Primal-dual subgradient methods for minimizing uniformly convex functions

Anatoli Iouditski, Yuri Nesterov

▶ To cite this version:

Anatoli Iouditski, Yuri Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. 2010. hal-00508933v1

HAL Id: hal-00508933
https://hal.archives-ouvertes.fr/hal-00508933v1

Preprint submitted on 8 Aug 2010 (v1), last revised 8 Jan 2014 (v3)

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

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

Primal-dual subgradient methods for minimizing uniformly convex functions

Anatoli Juditsky
LJK, Université J. Fourier
B.P. 53
38041 Grenoble Cedex 9, France.
E-mail: juditsky@imag.fr

Yuri Nesterov
CORE, Catholic University of Louvain
34 voie du Roman Pays
1348 Louvain-la-Neuve, Belgium.
E-mail: nesterov@core.ucl.ac.be

August 9, 2010

Abstract

We discuss non-Euclidean deterministic and stochastic algorithms for optimization problems with strongly and uniformly convex objectives. We provide accuracy bounds for the performance of these algorithms and design methods which are adaptive with respect to the parameters of strong or uniform convexity of the objective: in the case when the total number of iterations N is fixed, their accuracy coincides, up to a logarithmic in N factor with the accuracy of optimal algorithms.

1 Introduction

Let E be a (primal) finite-dimensional real vector space. In this paper we consider the optimization problem:

$$\min_x \{f(x) : x \in Q\}, \tag{1}$$

where Q is a closed convex set in E and function f is *uniformly convex* and Lipschitz-continuous on Q. Recall that a function f is called *uniformly convex* on $Q \subset E$ with convexity parameters $\rho = \rho(f) \geq 2$ and $\mu = \mu(f, \rho)$ if for all x and y from Q and any $\alpha \in [0, 1]$ we have

$$f(\alpha x + (a - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) - \tfrac{1}{2}\mu\alpha(1 - \alpha)\|x - y\|^\rho. \tag{2}$$

The function f which is uniformly convex with $\rho = 2$ is called *strongly convex*. Uniform convexity with $2 \leq \rho \leq \infty$ and $\mu \geq 0$ implies usual convexity.

In this paper we discuss deterministic and stochastic first order algorithms for (large scale) non-Euclidean uniformly convex objectives, thus extending non-Euclidean first order methods (see, e.g. [7, 11] and references therein) to uniformly convex optimization.

Uniformly convex functions have been introduced to optimization in [14] and extensively studied (cf. [1], [2], and [16]). The worst-case complexity bounds for the problem (1) with the exact oracle of the first order oracle are readily available (see [6]). Namely, for any method tuned to the relative accuracy ϵ the number of calls to the oracle is not less than $O(\epsilon^{-1})$ (which is much better than the corresponding bound $O(\epsilon^{-2})$ for much larger class of Lipschitz-continuous convex functions equipped with the first order oracle). The corresponding bound for uniformly convex problems with the convexity parameter ρ reads $O\left(\epsilon^{-\frac{2(\rho-1)}{\rho}}\right)$. Note that in the case of the stochastic oracle these bounds holds also for problems with smooth objective. Though optimal Euclidean algorithms for strongly convex optimization are readily available (see, e.g., [13]), they cannot be directly transposed to the non-Euclidean framework.

The results presented in this paper are not very new, as they were developed by the authors in 2004-2005. However, because of the immediate lack of application and, more importantly, due to new first order methods based on smoothing of structured problems with better complexity characteristics which were developed in [9, 10] at that time, the authors got an impression that the proposed algorithms of black-box (non-structured) uniformly convex optimization are of very limited interest. However, the developments of the last years clearly demonstrated that in some situations the black-box methods are irreplaceable. Indeed, the structure of a convex problem may be simply too complex for applying a smoothing technique. In particular, non-Euclidean first order methods of convex optimization have attracted much attention lately in relation, in particular, with very large scale applications arising in statistics and learning. For instance, some new applications involving large scale strongly convex optimization has been recently reported (see, e.g., [4, 5, 15]). These considerations encouraged the authors to publish the above mentioned results on subgradient methods for uniformly convex problems.

In this paper we develop minimax optimal primal-dual minimization schemes for uniformly convex problems as in (1) in the spirit of [11]. We also study the performance of multistage dual averaging procedures when applied to uniformly convex stochastic minimization problems. In particular, we show that such procedures attain the minimax rates of convergence on the considered problem class. We also provide confidence sets for approximate solutions of stochastic uniformly convex problems.

It is well known that performance of "classical" optimization routines for strongly (and uniformly) convex problems can become very poor when the parameters of strong (uniform) convexity are not known *a priori* (see, e.g. Section 2.1 in [7]). In the case of deterministic and stochastic optimization we develop *adaptive* minimization procedures in the case when the total number N of the method iterations is fixed. The accuracy of these procedures (which do not require a priori knowledge of parameters of uniform convexity) coincides, up to a logarithmic in N factor, with the accuracy of optimal algorithms (which "know" the exact parameters). It is worth to note that we do not know if it is possible to construct adaptive optimization procedures tuned to the fixed accuracy with analogous proprieties.

The paper is organized as follows: in Section 2 we define the basic ingredients of the minimization problem in question. Then we study the properties of the primal-dual subgradient algorithms in the problem with an exact deterministic oracle in Section 3 and show how the dual solutions can be produced in Section 4. In Section 5 we develop optimal algorithms for stochastic uniformly convex optimization and show how confidence sets for approximate solutions can be constructed. Finally, Section 6 contains some details of computation aspects of proposed routines.

2 Problem statement and basic assumptions

2.1 Notations and generalities

Let E^* be the dual of E. We denote the value of linear function $s \in E^*$ at $x \in E$ by $\langle s, x\rangle$. For measuring distances in E, let us fix some (primal) norm $\|\cdot\|$. This norm defines a primal unit ball

$$B = \{x \in E : \ \|x\| \le 1\}.$$

The dual norm $\|\cdot\|_*$ on E^* is introduced, as usual, by

$$\|s\|_* = \max_x \{\langle s, x\rangle : \ x \in B\}, \quad s \in E^*.$$

For other balls in E we adopt the following notation:

$$B_R(x) = \{y \in E : \ \|y - x\| \le R\}, \quad x \in E.$$

If a uniformly convex function f is *subdifferentiable* at x, then

$$f(y) \ge f(x) + \langle f'(x), y - x\rangle + \tfrac{1}{2}\mu\|y - x\|^\rho \quad \forall y \in Q, \tag{3}$$

where $f'(x) \in E^*$ denotes one of *subgradients* of f at $x \in Q$. If f is subdifferentiable at two points $x, y \in Q$, then

$$\langle f'(x) - f'(y), x - y\rangle \ge \mu\|x - y\|^\rho. \tag{4}$$

2.2 Problem statement

We consider the optimization problem (1) with the uniformly convex function f with convexity parameters $\rho(f)$ and $\mu(f)$. The basic assumption we make about the objective, and which is supposed to hold through the paper, is that f is Lipschitz-continuous on Q:

Assumption 1. We assume that all subgradients of the objective function are bounded:

$$\|f'(x)\|_* \le L, \quad \text{for any } \ x \in Q.$$

We are to study the performance of an iterative minimization schemes, and we consider two settings which differ with respect to the information available to the method at each iteration.

- *deterministic setting:* let x_k be the search points at iteration k, $k = 0, 1, ...$. We suppose that an exact subgradient observations $g_k = f'(x_k)$ and the exact objective values $f(x_k)$ are available;
- *stochastic setting:* the observation g_k of the subgradient $f'(x_k)$, requested by the method at the k-th iteration, is supplied by a *stochastic oracle*, i.e. g_k is a random vector.

To be more precise, suppose that we are given the probability space $(\Omega, \mathcal{F}, P)$ and a filtration $(\mathcal{F}_k)$, $k = -1, 0, 1, ...$ (non-decreasing family of σ-algebras which satisfies "usual" conditions).

Let

$$g_k \equiv g(x_k, \omega_k),$$

where

- $\{\omega_k\}_{k=0}^{\infty}$ is sequence of random parameters taking values in Ω, such that ω_k is $\mathcal{F}_k$-measurable;
- x_k is the k-th search point generated by the method. We suppose that x_k is $\mathcal{F}_{k-1}$-measurable (indeed, x_k is a measurable function of x_0 and observations $g_1, ..., g_{k-1}$ at iterations $1, ..., k-1$).

We also consider the following assumptions specific to the stochastic problem:

Assumption 2. The oracle is unbiased. Namely,

$$\mathbf{E}_{k-1}[g(x_k, \omega_k)] \in \partial f(x_k), \quad \text{a.s.} \quad x_k \in Q, \quad k = 0, 1, ...$$

Here $\mathbf{E}_k$ stands for the expectation conditioned by $\mathcal{F}_k$ (then $\mathbf{E} = \mathbf{E}_{-1}$ is the "full" expectation).

Let us denote

$$\xi_k = g_k - f'(x_k),$$

the stochastic perturbation. Note that $\mathbf{E}_{k-1}[\xi_k] = 0$ a.s. for $k = 0, 1, ...$ We suppose that the intensity of the sequence $\{g_k\}_{k=0}^{\infty}$ is bounded.

Assumption 3. We assume that

$$\sup_k \mathbf{E}\|\xi_k\|_*^2 \leq \sigma^2 < \infty \;\; \text{for} \quad k = 0, 1, ... \tag{5}$$

We will also use a stronger bound on the tails of the distribution of (ξ_k):

Assumption 4. There exists $\sigma < \infty$ such that

$$\mathbf{E}_{k-1}\left[\exp\left\{\|\xi_k\|_*^2 \sigma^{-2}\right\}\right] \leq \exp(1) \;\; \text{a.s.,} \quad k = 0, 1, ... \tag{6}$$

Note that by the Jensen inequality (6) implies (5).

2.3 Prox-function of the unit ball

Assume that we know a *prox-function* $d(x)$ of the ball B. This means that d is continuous and strongly convex on B in terms of (2) with some convexity parameter $\mu(d) > 0$. Moreover, we assume that

$$d(x) \geq d(0) = 0, \quad x \in B.$$

Hence, in view of (3) we have

$$d(x) \geq \tfrac{1}{2}\mu(d)\|x\|^2, \quad \forall x \in Q \cap B.$$

An important characteristic of the prox-function is its maximal value on the unit ball:

$$d(x) \;\leq\; A(d), \quad x \in B. \tag{7}$$

Therefore,

$$\mu(d) \;\;\leq\;\; 2A(d). \tag{8}$$

If the function d is growing quadratically, another important characteristics is its constant of quadratic growth $C(d)$ which we define as the smallest C such that

$$d(x) \leq C\|x\|^2. \tag{9}$$

We have

$$\mu(d) \leq 2C(d) \quad \text{and} \quad A(d) \leq C(d).$$

Example 1. Let $E = \mathbb{R}^n$ and let B be a unit Euclidean ball in $\mathbb{R}^n$. We choose the norm $\|\cdot\|$ to be the Euclidean norm on $\mathbb{R}^n$, so that the function $d(x) = \|x\|_2^2/2$ is strongly convex with $\mu(d) = 1$ and $C(d) = A(d) = 1/2$.

Example 2. Let again $E = \mathbb{R}^n$ and let B be the standard hyperoctahedron in $\mathbb{R}^n$, i.e. a unit l_1-ball: $B = \{x \in \mathbb{R}^n | \, \|x\|_1 \leq 1\}$, where

$$\|x\|_1 = \sum_{i=1}^n |x^{(i)}|.$$

We take $\|x\| = \|x\|_1$ and consider for $p > 1$ the function d,

$$d(x) = \tfrac{1}{2} \left(\sum_{i=1}^n |x_i|^p \right)^{2/p} = \tfrac{1}{2}\|x\|_p^2.$$

The function d is strongly convex with $\mu(d) = O(1) n^{\frac{p-1}{p}}$, and for $p = 1 + \frac{1}{\ln n}$ we have $\mu(d) = O(1)(\ln n)^{-1}$ (see, e.g. [6]). Further, we clearly have $A(d) = C(d) = 1/2$.

Note that norm-type prox-functions are not the only possible in the hyperoctahedron setting. Another example of prox-function of the l_1-unit ball B, which is very interesting from the computational point of view, is as follows:

$$\begin{aligned} d(x) &= \min\{ \sum_{i=1}^n [\, \psi(u^{(i)}) + \psi(v^{(i)}) \,] : \sum_{i=1}^n \left[u^{(i)} + v^{(i)} \right] = 1, \\ &\qquad x^{(i)} = u^{(i)} - v^{(i)}, \; u^{(i)} \geq 0, \; v^{(i)} \geq 0, \; i = 1, \dots, n \,\} \; + \; \ln(2n), \qquad (10) \\ \psi(t) &= \begin{cases} t \ln t, & t > 0, \\ 0, & t = 0. \end{cases} \end{aligned}$$

In order to show that this function is strongly convex on the standard hyperoctahedron $B = \{x \in \mathbb{R}^n | \, \|x\|_1 \leq 1\}$, we need the following general result.

Lemma 1. *Let Q be a bounded closed convex set in E containing the origin. If function $f(x)$ is strongly convex on Q with parameter $\mu \geq 0$, then its symmetrization*

$$f^0(x) = \min_{u,v,\alpha} \{ f(u) + f(v) : \; x = u - v, \; u \in \alpha Q, \; v \in (1-\alpha) Q, \; \alpha \in [0,1] \},$$

is strongly convex on the set $Q^0 = \mathrm{Conv}\,\{Q, -Q\}$ with convexity parameter $\frac{1}{2}\mu(f)$.

Proof: Consider two points $x_i \in Q^0$, $i = 1, 2$. Suppose that

$$x_i = u_i - v_i, \quad u_i \in \alpha_i Q, \quad v_i \in (1-\alpha_i) Q, \quad \alpha_i \in [0,1],$$

$$f^0(x_i) = f(u_i) + f(v_i), \quad i = 1, 2.$$

Let us choose an arbitrary $\alpha \in [0,1]$. Then,

$$\begin{aligned} x(\beta) &\stackrel{\mathrm{def}}{=} \beta x_1 + (1-\beta) x_2 \\ &= \beta(u_1 - v_1) + (1-\beta)(u_2 - v_2) \\ &= \beta u_1 + (1-\beta) u_2 - (\beta v_1 + (1-\beta) v_2). \end{aligned}$$

Denote $\gamma = \beta\alpha_1 + (1-\beta)\alpha_2$. Then

$$1 - \gamma = \beta(1-\alpha_1) + (1-\beta)(1-\alpha_2).$$

Note that $u_i = \alpha_i \bar{u}_i$, and $v_i = (1-\alpha_i)\bar{v}_i$ for some $\bar{u}_i$ and $\bar{v}_i$ from Q, $i = 1, 2$. Therefore, denoting

$$\tau = \beta\alpha_1/\gamma, \quad \xi = \beta(1-\alpha_1)/(1-\gamma),$$

we obtain

$$\begin{aligned}
x(\beta) &= \beta\alpha_1\bar{u}_1 + (1-\beta)\alpha_2\bar{u}_2 - (\beta(1-\alpha_1)\bar{v}_1 + (1-\beta)(1-\alpha_2)\bar{v}_2) \\
&= \gamma(\tau\bar{u}_1 + (1-\tau)\bar{u}_2) - (1-\gamma)(\xi\bar{v}_1 + (1-\xi)\bar{v}_2) \\
&\overset{\mathrm{def}}{=} \gamma\bar{u}_3 - (1-\gamma)\bar{v}_3
\end{aligned}$$

with some $\bar{u}_3$ and $\bar{v}_3$ from Q. Hence, $u_3 = \gamma\bar{u}_3 \in \gamma Q$, and $v_3 = (1-\gamma)\bar{v}_3 \in (1-\gamma)Q$. Consequently, by definition of function f^0 and using inclusions $u_i, v_i \in Q$, $i = 1, 2$, we obtain

$$\begin{aligned}
f^0(x(\beta)) &\leq f(u_3) + f(v_3) \\
&= f(\beta u_1 + (1-\beta)u_2) + f(\beta v_1 + (1-\beta)v_2) \\
&\leq \beta f(u_1) + (1-\beta)f(u_2) - \tfrac{1}{2}\mu\beta(1-\beta)\|u_1 - u_2\|^2 \\
&\quad +\beta f(v_1) + (1-\beta)f(v_2) - \tfrac{1}{2}\mu\beta(1-\beta)\|v_1 - v_2\|^2 \\
&= \beta f^0(x_1) + (1-\beta)f^0(x_2) - \tfrac{1}{2}\mu\beta(1-\beta)\left[\, \|u_1 - u_2\|^2 + \|v_1 - v_2\|^2 \,\right].
\end{aligned}$$

It remains to note that

$$2\|u_1 - u_2\|^2 + 2\|v_1 - v_2\|^2 \quad \geq \quad \|u_1 - u_2 - (v_1 - v_2)\|^2 \;=\; \|x_1 - x_2\|^2.$$

∎

Thus, for function $d(x)$ defined by (10) we can take

$$\mu(d) = \tfrac{1}{2}, \quad A(d) = \ln(2n).$$

Note that d does not satisfy the quadratic growth condition (9)

For $z \in Q$, consider the set

$$Q_R(z) \overset{\mathrm{def}}{=} Q \cap B_R(z).$$

This set can be equipped with a prox-function

$$d_{z,R}(x) = d\left(\frac{1}{R}(x - z)\right).$$

Thus, the prox-center of the set $Q_R(z)$ is z, and $\mu(d_{z,R}) = \frac{1}{R^2}\mu(d)$. Moreover, by (7),

$$d_{z,R}(x) \leq A(d), \quad \forall x \in Q_R(z).$$

In what follows we need the objects:

$$V_{z,R,\beta}(s) \;=\; \max_x \{\langle s, x - z\rangle - \beta d_{z,R}(x) : \; x \in Q_R(z)\}, \tag{11}$$

and

$$\pi_{z,R,\beta}(s) \;=\; \arg\max_x \{\langle s, x - z\rangle - \beta d_{z,R}(x) : \; x \in Q_R(z)\}.$$

Note that $\operatorname{dom} V_{z,R,\beta} = E^*$. Let us mention some properties of function $V_{z,R,\beta}$ (cf. Lemma 1 [11]):

- if $\beta_1 \leq \beta_2$ then $V_{z,R,\beta_1}(s) \geq V_{z,R,\beta_2}(s)$;
- the function $V_{z,R,\beta}$ is convex and differentiable on E^*. Moreover, its gradient is Lipschitz continuous with the constant $\frac{R^2}{\beta\mu(d)}$:

$$\|V'_{z,R,\beta}(s_1) - V'_{z,R,\beta}(s_2)\| \;\leq\; \frac{R^2}{\beta\mu(d)}\|s_1 - s_2\|_*, \quad \forall s_1, s_2 \in E^*.$$

- For any $s \in E^*$,

$$V'_{z,R,\beta}(s) + z = \pi_{z,R,\beta}(s) \in Q_R(z).$$

3 Deterministic methods for uniformly convex functions

We start with the description of the basic tool – the dual averaging procedure, which originates in [11].

3.1 Method of Dual Averaging

At each phase the dual averaging (DA) method will be applied to the following auxiliary problem:

$$\min_x \{f(x) : \; x \in Q_R(\bar{x})\}. \tag{12}$$

Its feasible set is endowed with the following prox-function:

$$d_{\bar{x},R}(x) = d\left(\tfrac{1}{R}(x - \bar{x})\right).$$

Consider now the generic scheme of Dual Averaging as applied to the problem (12).

Algorithm 1.

Initialization: Set $x_0 = \bar{x}$, $s_0 = 0 \in E^*$. Choose $\beta_0 > 0$.

Iteration $(k \geq 0)$:

1. Choose $\lambda_k > 0$. Set $s_{k+1} = s_k + \lambda_k f'(x_k)$, where $\{\lambda_i\}_{i=0}^{\infty}$ is a sequence of positive parameters.
2. Choose $\beta_{k+1} \geq \beta_k$. Set $x_{k+1} = \pi_{\bar{x},R,\beta_{k+1}}(-s_{k+1})$.

The process is terminated after N iterations. The resulting point is defined as follows:

$$x_N(\bar{x}, R) \;=\; \left(\sum_{i=0}^{N} \lambda_i\right)^{-1} \sum_{i=0}^{N} \lambda_i x_i. \tag{13}$$

The result below underlies the following developments (cf. Theorem 1 of [11].):

Proposition 1. *For any $x \in Q_R(\bar{x})$,*

$$\sum_{i=0}^{k} \lambda_i \langle f'(x_i), x_i - x\rangle \le \; d_{\bar{x},R}(x)\beta_{k+1} + \frac{R^2}{2\mu(d)} \sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_i} \|f'(x_i)\|_*^2. \tag{14}$$

Let $\lambda_i = 1$ and $\beta_i = \gamma\sqrt{N+1}$, $i = 0, ..., N$ with some $\gamma > 0$. We form the *gap value*

$$\delta_k(\bar{x}, R) = \max_x \left\{ \frac{1}{k} \sum_{i=0}^{k} \langle f'(x_i), x_i - x\rangle : \; x \in Q_R(\bar{x}) \right\}, \tag{15}$$

where $\{\lambda_i\}_{i=0}^{\infty}$ is a sequence of positive parameters. In view of (13) we have the following lemma:

Lemma 2. *Let us choose an arbitrary $\bar{x} \in Q$ and let x^* be the optimal solution of problem (12). Then the approximate solution supplied by Algorithm 1 with the constant gain $\beta_i = \gamma\sqrt{N+1}$ satisfies*

$$\begin{aligned} f(x_N(\bar{x}, R)) - f(x^*) \;&\le\; \frac{1}{\sqrt{N+1}} \left(\gamma A(d) + \frac{L^2 R^2}{2\gamma\mu(d)} \right), \\ \|x_N(\bar{x}, R) - x^*\|^\rho \;&\le\; \frac{\delta_N(\bar{x}, R)}{\mu(f)} \le \frac{1}{\mu(f)\sqrt{N+1}} \left(\gamma A(d) + \frac{L^2 R^2}{2\gamma\mu(d)} \right). \end{aligned}$$

Proof: In view of conditions of the lemma, $x^* \in Q_R(\bar{x})$. From the assumptions on function f, we conclude that

$$\langle f'(x_i), x_i - x^*\rangle \;\ge\; f(x_i) - f(x^*),$$

$$\langle f'(x_i), x_i - x^*\rangle \;\overset{(4)}{\ge}\; \mu(f) \cdot \|x_i - x^*\|^\rho, \quad i = 0, \ldots, N.$$

Hence,

$$\begin{aligned} (N+1)\delta_N(\bar{x}, R) \;&\ge\; \sum_{i=0}^{N} [f(x_i) - f(x^*)] \;\ge\; (N+1)[f(x_N(\bar{x}, R)) - f(x^*)], \\ (N+1)\delta_N(\bar{x}, R) \;&\ge\; \mu(f) \sum_{i=0}^{N} \|x_i - x^*\|^\rho \ge \mu(f)(N+1)\|x_N(\bar{x}, R) - x^*\|^\rho. \end{aligned}$$

It remains to note that $d_{\bar{x},R}(x) \le A(d)$ for any $x \in Q_R(\bar{x})$ use the inequality (14). ∎

Under the premises of the lemma we can establish the following immediate bounds:

Corollary 1. *Let x^* be an optimal solution of (12). Then for the choice*

$$\gamma = \frac{LR}{\sqrt{2\mu(d)A(d)}}$$

we have the estimates:

$$\begin{aligned} f(x_N(\bar{x}, R)) - f(x^*) &\leq LR\sqrt{\frac{2A(d)}{\mu(d)(N+1)}}, \\ \|x_N(\bar{x}, R) - x^*\|^\rho &\leq \frac{LR}{\mu(f)}\sqrt{\frac{2A(d)}{\mu(d)(N+1)}}. \end{aligned} \tag{16}$$

3.2 Multi-step algorithms

Now we are ready to analyze multistage procedures for uniformly convex functions. In this section we assume that the constants L, $\mu(f)$, ρ and $R_0 \geq \|x^* - x_0\|$ are known. Let us fix $\epsilon > 0$ and let x_0 be an arbitrary element of Q.

Algorithm 2.

Initialization: Set $y_0 = x_0$ and $m = \lfloor \log_2 \frac{\mu(f)}{\epsilon} R_0^\rho \rfloor + 1$.[1)] Let $\tau = \frac{2(\rho-1)}{\rho}$.

Stage $k = 1, \ldots, m$:

1. Define $N_k = \lfloor 2^{\tau k} \frac{4L^2 A(d)}{\mu^2(f)\mu(d)R_0^{2(\rho-1)}} \rfloor$ and $R_k^\rho = 2^{-k} R_0^\rho$.
2. Compute $y_k = x_{N_k}(y_{k-1}, R_{k-1})$ with $\gamma_k = \frac{LR_{k-1}}{\sqrt{2\mu(d)A(d)}}$.

Output: $\widehat{x}_\epsilon(y_0, R_0) := y_m$.

Note that the parameters of the algorithm satisfy the following relations:

$$N_k + 1 \geq 2^{\tau k} \frac{4L^2 A(d)}{\mu^2(f)\mu(d)R_0^{2(\rho-1)}} \geq N_k, \qquad 2^m \geq \frac{\mu(f)}{\epsilon} R_0^\rho \geq 2^{m-1}. \tag{17}$$

Theorem 1. *The points $\{y_k\}_{k=1}^m$ generated by Algorithm 2 satisfy the following conditions:*

$$\|y_k - x^*\|^\rho \leq R_k^\rho = 2^{-k} R_0^\rho, \quad k = 0, \ldots, m, \tag{18}$$

$$\delta_{N_k}(y_{k-1}, R_{k-1}) \leq \mu(f) R_k^\rho = \mu(f) 2^{-k} R_0^\rho, \quad k = 1, \ldots, m. \tag{19}$$

Moreover, $f(\widehat{x}_\epsilon(y_0, R_0)) - f^* \leq \epsilon$ and the total number $N(\epsilon)$ of iterations in the scheme does not exceed

$$\left(\frac{2^{m+1}}{R_0^\rho}\right)^\tau \frac{4L^2 A(d)}{\mu^2(f)\mu(d)} \overset{(17)}{\leq} \frac{4^{\tau+1} L^2 A(d)}{\mu(f)^{\frac{2}{\rho}} \mu(d)} \epsilon^{-\tau}. \tag{20}$$

---

1)Here $\lfloor a \rfloor$ stands for the largest integer strictly smaller than a.

Proof: Indeed, for $k=0$, (18) is valid. Assume it is valid for some $k \geq 0$. Note that

$$\sqrt{N_{k+1}+1} \quad \overset{(17)}{\geq} \quad \left(\left(\frac{2^k}{R_0^\rho} \right)^\tau \frac{8L^2 A(d)}{\mu^2(f)\mu(d)} \right)^{1/2} \quad = \quad 2 \frac{L\sqrt{2A(d)}}{\mu(f) R_k^{\rho-1}\sqrt{\mu(d)}}.$$

Therefore, in view of Proposition 1 and Corollary 1, we have

$$\delta_{N_{k+1}}(y_k, R_k) \quad \leq \quad \frac{LR_k\sqrt{2A(d)}}{\sqrt{\mu(d)(N_{k+1}+1)}} \leq \frac{\mu(f)}{2} R_k^\rho = \mu(f) R_{k+1}^\rho,$$

and this is (19) for the next value of the iteration counter. Further,

$$\|y_{k+1} - x^*\|^\rho \leq \mu(f)^{-1} \delta_{N_{k+1}}(y_k, R_k) \leq \; R_{k+1}^\rho,$$

and this is (18) for $k+1$.

Finally, at the end of the m-th stage, in view of Lemma 2 and (19) we have

$$f(\widehat{x}_\epsilon(y_0, R_0)) - f^* \quad \leq \quad \delta_{N_m}(y_{m-1}, R_{m-1}) \overset{(19)}{\leq} \quad \mu(f) R_m^\rho \quad = \quad 2^{-m}\mu(f) R_0^\rho \quad \overset{(17)}{\leq} \quad \epsilon.$$

The complexity of the method can be estimated as follows:

$$N(\epsilon) \quad \overset{(17)}{\leq} \quad \sum_{k=1}^m 2^k \tau \frac{4L^2 A(d)}{\mu^2(f)\mu(d) R_0^{2(\rho-1)}} \quad < \quad \left(\frac{2^{m+1}}{R_0^\rho} \right)^\tau \frac{4L^2 A(d)}{(2^\tau - 1)\mu^2(f)\mu(d)}.$$

To conclude (20) it suffices to notice that by (17), $2^{m+1} \leq 4\frac{\mu(f)R_0^\rho}{\epsilon}$. ∎

An important particular case of Theorem 1 is the case of strongly convex objective f. In the latter case $\tau = 1$ and the analytical complexity of Algorithm 2 does not exceed

$$\frac{16L^2 A(d)}{\mu(f)\mu(d)} \, \epsilon^{-1}.$$

The method can be easily rewritten for the case when the total number N of calls to the oracle is fixed a priori. Suppose that $N \geq [2^\tau(2^\tau+1)] \frac{4L^2 A(d)}{\mu^2(f)\mu(d) R_0^{2(\rho-1)}}$ (in the opposite case the bound of Corollary 1 for one-stage method provides better accuracy). Consider the following procedure:

Algorithm 3.

Initialization: Set $y_0 = x_0$, $\tau = \frac{2(\rho-1)}{\rho}$, compute $N_j = \lfloor 2^{\tau j} \frac{4L^2 A(d)}{\mu(f)^2 \mu(d) R_0^{2(\rho-1)}} \rfloor$ while $\sum_j N_j \leq N$. Set

$$m(N) \quad = \quad \max\{k : \; \textstyle\sum_{j=1}^k N_j \leq N\}.$$

Stage $k = 1, \ldots, m(N)$: Set $R_k^\rho = 2^{-k} R_0^\rho$. Compute $y_k = x_{N_k}(y_{k-1}, R_{k-1})$ with

$$\gamma_k \quad = \quad \frac{LR_{k-1}}{\sqrt{2\mu(d)A(d)}}.$$

Output: $\widehat{x}_N = y_{m(N)}$.

Corollary 1. *We have*

$$f(\widehat{x}_N) - f^* \leq 2\left(\frac{8L^2A(d)}{\mu(f)^{\frac{2}{\rho}}\mu(d)N}\right)^{1/\tau}. \tag{21}$$

Proof. Indeed, as in the proof of Theorem 1 we conclude that

$$f(\widehat{x}_N) - f^* \leq 2^{-m(N)}\mu(f)R_0^\rho.$$

Now it suffices to notice that the number $m(N)$ of the stages of the algorithm can be easily bounded:

$$\frac{N}{2} \leq \sum_{k=1}^{m(N)} N_k \leq \left(\frac{2^{m(N)+1}}{R_0^\rho}\right)^\tau \frac{4L^2A(d)}{(2^\tau-1)\mu^2(f)\mu(d)}.$$

Thus,

$$2^{-m(N)} \leq 2\left(\frac{8L^2A(d)}{\mu^2(f)\mu(d)N}\right)^{1/\tau} R_0^{-\rho},$$

and the bound (21) follows. ∎

3.3 Methods with quadratically growing prox-function

We propose here a slightly different version of multi-stage procedures for the case when the prox-function satisfies the condition (9) of quadratic growth.

The result below is an immediate consequence of Proposition 1 (cf. Lemma 2 and Corollary 1):

Corollary 2. *Let x^* be an optimal solution of (12). Suppose that the prox-function d satisfies (9) and that $\|\bar{x} - x^*\| \leq r \leq R$. Then the approximate solution $x_N(\bar{x}, R)$, provided by Algorithm 1 with*

$$\gamma = \frac{R^2L}{r\sqrt{2C(d)\mu(d)}},$$

satisfies

$$f(x_N(\bar{x}, R)) - f(x^*) \leq rL\sqrt{\frac{2C(d)}{\mu(d)(N+1)}}, \tag{22}$$

$$\|x_N(\bar{x}, R) - x^*\|^\rho \leq \frac{rL}{\mu(f)}\sqrt{\frac{2C(d)}{\mu(d)(N+1)}}. \tag{23}$$

Indeed, to show (22) and (23) it suffices to use (14) and to observe that due to (9) $d_{\bar{x},R}(x^*) \leq C(d)\frac{r^2}{R^2}$.

The following multi-stage scheme exploits the "scalability property" (9) of the prox-function d. It starts from arbitrary $x_0 \in Q$. As in the previous section, we assume that the constants L, $\mu(f)$ and the diameter R_0 of Q are known.

Algorithm 4.

Initialization: Set $y_0 = x_0$, $\tau = \frac{2(\rho-1)}{\rho}$ and $m = \lfloor \log_2 \frac{\mu(f)}{\epsilon} R_0^\rho \rfloor + 1$.

Stage $k = 1, \ldots, m$:

1. Define $N_k = \lfloor 2^{\tau k} \frac{4L^2 C(d)}{\mu^2(f)\mu(d) R_0^{2(\rho-1)}} \rfloor$ and $r_k^\rho = 2^{-k} R_0^\rho$.
2. Compute $y_k = x_{N_k}(y_{k-1}, R_0)$ with $\gamma_k = \frac{LR_0^2}{r_{k-1}\sqrt{2C(d)\mu(d)}}$.

Output: Set the approximate solution $\widehat{x}_\epsilon = y_m$.

We would like to stress the difference between Algorithms 2 and 4: in Algorithm 4 the delation parameter $R = R_0$ of the prox-function d remains the same through all the stages of the method. Only the gain γ_k and the duration N_k of the stage depend on the stage index k.

We have the following analogue of Theorem 1 in this case:

Theorem 2. *Suppose that*

$$N \geq N(\epsilon) = \frac{4^{\tau+1} L^2 C(d)}{\mu(f)^{\frac{2}{\rho}} \mu(d)} \epsilon^{-\tau}.$$

Then the approximate solution $\widehat{x}_N$, provided by Algorithm 4 satisfies:

$$Ef(\widehat{x}_\epsilon) - f^* \leq \epsilon.$$

Proof: As in the proof of Theorem 1, the result of the theorem follows immediately from the relations:

$$E\|y_k - x^*\|^\rho \leq r_k^\rho = 2^{-k} R_0^\rho \tag{24}$$

and

$$Ef(y_k) - f^* \leq \mu(f) r_k^\rho \leq \mu(f) 2^{-k} R_\rho^2. \tag{25}$$

Indeed, using the relations above we write:

$$Ef(\widehat{x}) - f^* \quad \leq \quad \mu(f) r_m^\rho = 2^{-m} \mu(f) R_0^\rho \leq \epsilon.$$

Let us verify the bounds (24) and (25). Assume that (24) valid for some $k \geq 0$. Note that

$$\sqrt{N_{k+1} + 1} > \frac{2^{\tau k/2}}{R_0^{\rho-1}} \left(\frac{8L^2 C(d)}{\mu^2(f)\mu(d)} \right)^{1/2} = 2 \frac{L}{\mu(f) r_k} \sqrt{\frac{2C(d)}{\mu(d)}}.$$

Therefore, in view of Corollary 4, we have

$$E\|y_{k+1} - x^*\|^\rho \quad \leq \quad \frac{L r_k}{\mu(f)} \sqrt{\frac{2C(d)}{\mu(d)(N_{k+1} + 1)}} \leq \frac{r_k^\rho}{2} = r_{k+1}^\rho,$$

and

$$Ef(y_{k+1}) - f^* \leq (L + \sigma) r_k \sqrt{\frac{2C(d)}{\mu(d)(N_{k+1} + 1)}} \leq \frac{\mu(f)}{2} r_k^\rho = \mu(f) r_{k+1}^\rho.$$

∎

The method can be rewritten in the when the total number N of calls to the oracle is fixed. Suppose that $N \geq 2^\tau (2^\tau + 1) \frac{4L^2 C(d)}{\mu^2(f)\mu(d) R_0^{2(\rho-1)}}$. Consider the following procedure:

Algorithm 5.

Initialization: Set $y_0 = x_0$, $\tau = \frac{2(\rho-1)}{\rho}$, compute $N_j = \left\lfloor 2^{\tau j} \frac{4L^2 C(d)}{\mu(f)^2 \mu(d) R_0^{2(\rho-1)}} \right\rfloor$, while $\sum_j N_j \leq N$. Set $m(N) = \max\{k : \sum_{j=1}^k N_j \leq N\}$.

Stage $k = 1, \dots, m(N)$:

Set $r_k^\rho = 2^{-k} R_0^\rho$. Compute $y_k = x_{N_k}(y_{k-1}, R_0)$ with $\gamma_k = \frac{LR_0^2}{r_{k-1}\sqrt{2C(d)\mu(d)}}$.

Termination: Set the approximate solution $\widehat{x}_N = y_{m(N)}$.

Corollary 3. *We have*

$$Ef(\widehat{x}_N) - f^* \leq 2\left(\frac{8L^2 C(d)}{\mu(f)^{\frac{2}{\rho}} \mu(d) N}\right)^{1/\tau}.$$

The proof of the corollary is completely analogous to that of Corollary 1.

3.4 Adaptive algorithm

Consider the setting in which the total number N of calls to the oracle is fixed and suppose that the convexity parameters ρ $\mu(f)$ are unknown. We propose a multi-stage procedure which does not require the knowledge of these parameters and attains the accuracy of the method which "knows" the convexity parameters up to a logarithmic in N factor. Following the terminology used in statistical and control literature, we call such procedures adaptive (with respect to unknown parameters). In what follows we suppose that the bounds L and R_0 are known *a priori*.

We analyze here the following adaptive version of Algorithm 3 (we leave the construction and analysis of adaptive version of Algorithm 5 as an exercise to the reader):

Algorithm 6.

Initialization: Set $y_0 = x_0$, $m = \left[\frac{1}{2}\log_2 \frac{\mu(d)N}{A(d)\log_2 N}\right] - 1$ [2], $N_0 = [N/m]$, and

$$R_k = 2^{-k} R_0, \quad k = 1, ..., m.$$

Stage $k = 1, ..., m$: Compute $y_k = \widehat{x}_{N_0}(y_{k-1}, R_{k-1})$ with $\gamma_k = \frac{LR_{k-1}}{\sqrt{2\mu(d)A(d)}}$.

Output: $\widehat{x}_N = \mathrm{argmin}_{k=1,\dots,m} f(y_k)$.

Theorem 3. *The approximate solution $\widehat{x}_N$ satisfies for $N \geq 4$*

$$f(\widehat{x}_N) - f^* \quad \leq \quad 2\left(\frac{16L^2 A(d)\log_2 N}{\mu(f)^{\frac{2}{\rho}} \mu(d) N}\right)^{\frac{\rho}{2(\rho-1)}}.$$

[2] here $[a]$ stands here for the largest integer less or equal to a

Proof: Note that m satisfies

$$2^m \quad \leq \quad \tfrac{1}{2}\sqrt{\tfrac{\mu(d)N}{A(d)\log_2 N}}. \tag{26}$$

Thus by (8), $m\leq\frac{1}{2}\log_2 N$. Assume now that $\mu(f) \leq \frac{4L}{R_0^{\rho-1}}\sqrt{\frac{A(d)\log_2 N}{\mu(d)N}}$. We have

$$\begin{aligned} f(y_1) - f^* \quad &\leq \quad \delta_{N_0}(y_0, R_0) \leq LR_0\sqrt{\frac{2A(d)}{\mu(d)(N_0+1)}} \;\leq\; LR_0\sqrt{\frac{2mA(d)}{\mu(d)N}} \\ &\leq \quad LR_0\sqrt{\frac{A(d)\log_2 N}{\mu(d)N}} \leq \left(\frac{16L^2A(d)\log_2 N}{\mu(f)^{\frac{2}{\rho}}\mu(d)N}\right)^{\frac{\rho}{2(\rho-1)}}, \end{aligned}$$

what implies the statement of the theorem in this case. Next, let us denote $\mu_0 = 2^{-m}LR_0^{1-\rho}$ so that

$$2LR_0^{1-\rho}\sqrt{\frac{A(d)\log_2 N}{\mu(d)N}} \leq \mu_0 < 4LR_0^{1-\rho}\sqrt{\frac{A(d)\log_2 N}{\mu(d)N}}, \tag{27}$$

and $\mu_k = 2^{(\rho-1)k}\mu_0$, $k = 1, ..., m$. Observe that from the available information we can derive an upper bound on the unknown parameter $\mu(f)$, namely,

$$\mu(f) \quad \leq \quad \tfrac{L}{R_0^{\rho-1}} \leq \mu_m.$$

Suppose now that the true $\mu(f)$ satisfies $\mu_0 \leq \mu(f) \leq \mu_m$. We need the following auxiliary result.

Lemma 1. *Let k^* satisfy $\mu_{k^*} \leq \mu(f) \leq 2^{\rho-1}\mu_{k^*}$. For $1 \leq k \leq k^*$, the points $\{y_k\}_{k=1}^m$ generated by Algorithm 6 satisfy the following relations:*

$$\|y_{k-1} - x^*\| \quad \leq \quad R_{k-1} = 2^{-k+1}R_0, \tag{28}$$

$$\delta_{N_0}(y_{k-1}, R_{k-1}) \quad \leq \quad \mu_k R_k^{\rho} = 2^{-k}\mu_0 R_0^{\rho}. \tag{29}$$

For $k^* < k \leq m$, we have

$$f(y_k) \leq f(y_{k^*}) + \mu_{k^*}R_{k^*}^{\rho}. \tag{30}$$

Proof. Let us prove first (28) and (29). Indeed, for $k = 1$ (28) is valid. Assume it is valid for some $k \geq 1$. We write

$$\begin{aligned} \mu(f) \quad &\geq \quad \mu_k \;=\; 2^{(\rho-1)k}\mu_0 \;=\; \left(\frac{2^k}{R_0}\right)^{\rho-1} \cdot L2^{-m} \\ &\overset{(26)}{\geq} \quad \left(\frac{2^k}{R_0}\right)^{\rho-1} 2L\sqrt{\frac{A(d)\log_2 N}{\mu(d)N}} \;\geq\; \frac{2L}{R_k^{\rho-1}}\sqrt{\frac{2A(d)}{\mu(d)(N_0+1)}}. \end{aligned}$$

Therefore,

$$\delta_{N_0}(y_{k-1}, R_{k-1}) \quad \overset{(16)}{\leq} \quad \frac{LR_{k-1}\sqrt{2A(d)}}{\sqrt{\mu(d)(N_0+1)}} \;\leq\; \tfrac{1}{2}\mu_k R_k^{\rho-1}R_{k-1} \;=\; \mu_k R_k^{\rho}. \tag{31}$$

That is (29). Moreover,

$$\|y_k - x^*\|^\rho \;\leq\; \mu(f)^{-1}\delta_{N_0}(y_{k-1}, R_{k-1}) \leq \tfrac{\mu_k}{\mu(f)} R_k^\rho \leq R_k^\rho,$$

and this is (28) for the next index value. Further, as in (31), for $k > k^*$ we have

$$\begin{aligned} f(y_k) - f(y_{k-1}) \;&\leq\; \delta_{N_0}(y_{k-1}, R_{k-1}) \;\leq\; LR_{k-1}\sqrt{\tfrac{2A(d)}{\mu(d)(N_0+1)}} \\ &=\; 2^{k^*-k} LR_{k^*-1}\sqrt{\tfrac{2A(d)}{\mu(d)(N_0+1)}} \overset{(31)}{\leq} 2^{k^*-k}\mu_{k^*} R_{k^*}^\rho. \end{aligned}$$

Then

$$f(y_k) - f(y_{k^*}) \;=\; \sum_{j=k^*+1}^{k} f(y_j) - f(y_{j-1}) \leq \sum_{j=k^*+1}^{k} 2^{k^*-j}\mu_{k^*} R_{k^*}^\rho \leq \mu_{k^*} R_{k^*}^\rho.$$

This proves the lemma. ∎

Now we can finish the proof of the theorem. Recall that $\mu_0 \leq \mu(f) \leq \mu_m$. At the end of the k^*-th stage we have

$$\begin{aligned} f(y_{k^*}) - f^* \;&\leq\; \delta_{N_0}(y_{k^*-1}, R_{k^*-1}) \leq \mu_{k^*} R_{k^*}^\rho \leq 2\frac{\mu_{k^*}^{\frac{1}{\rho-1}}}{\mu(f)^{\frac{1}{\rho-1}}}\mu_{k^*} R_{k^*}^\rho \\ &=\; 2\frac{\mu_0^{\frac{\rho}{\rho-1}} R_0^\rho}{\mu(f)^{\frac{1}{\rho-1}}} \overset{(27)}{\leq} 2\left(\frac{16L^2 A(d)\log_2 N}{\mu(f)^{\frac{2}{\rho}}\mu(d) N}\right)^{\frac{\rho}{2(\rho-1)}}. \end{aligned}$$

∎

4 Generating dual solutions

In order to speak about primal-dual solutions, we need to fix somehow the structure of objective function in problem (1). Let us assume that

$$f(x) \;=\; \max_{w\in S}\; \Psi(x,w), \quad x \in Q,$$

where S is a closed convex set, and function Ψ is convex in the first argument $x \in Q$ and concave in the second argument $u \in S$. Let us assume that Ψ is subdifferentiable in x at any $(x,w) \in Q \times S$. Then we can take

$$\begin{aligned} f'(x) \;&=\; \Psi'_x(x, w(x)), \\ w(x) \;&\in\; \operatorname{Arg}\max_{w\in S} \Psi(x,w). \end{aligned} \tag{32}$$

Thus, we can define the dual function $\eta(w) \;=\; \min\limits_{x\in Q}\; \Psi(x,w)$, and the dual maximization problem

$$\text{Find } f^* = \max_w\{\eta(w) :\; w \in S\}.$$

For any $w \in S$, we assume that $\Psi(\cdot,w)$ is uniformly convex on Q with convexity parameters $\rho = \rho(\Psi)$ and $\mu(\Psi)$).

The following result is quite standard (cf. Lemma 3 [12]).

Lemma 2. *Define* $\bar{x} = \frac{1}{N+1}\sum_{i=0}^{N} x_i$, *and* $\bar{w}_N = \frac{1}{N+1}\sum_{i=0}^{N} w(x_i)$. *Then*

$$f(\bar{x}_N) - \eta(\bar{w}_N) \quad \leq \quad l_N^* \stackrel{\rm def}{=} \max_x \left\{ \tfrac{1}{N+1}\sum_{i=0}^{N} \langle f'(x_i), x_i - x\rangle - \tfrac{1}{2}\mu(\Psi)\|x - \bar{x}_N\|^\rho :\ x \in Q \right\}. \qquad (33)$$

Proof. Since Ψ is convex in the first argument, for any $x \in Q$ we have

$$\begin{aligned} \langle f'(x_i), x_i - x\rangle &\stackrel{(32)}{=} \langle \Psi'_x(x_i, w(x_i)), x_i - x\rangle \\ &\geq \Psi(x_i, w(x_i)) - \Psi(x, w(x_i)) + \tfrac{1}{2}\mu(\Psi)\|x - x_i\|^\rho \\ &= f(x_i) - \Psi(x, w(x_i)) + \tfrac{1}{2}\mu(\Psi)\|x - x_i\|^\rho. \end{aligned}$$

Hence,

$$\begin{aligned} l_N^* &= \tfrac{1}{N+1}\max_x \left\{ \sum_{i=0}^{N} \langle f'(x_i), x_i - x\rangle - \mu(\Psi)\tfrac{N+1}{2}\|x - \bar{x}_N\|^\rho :\ x \in Q \right\} \\ &\geq \tfrac{1}{N+1}\max_x \left\{ \sum_{i=0}^{N} [\langle f'(x_i), x_i - x\rangle - \tfrac{1}{2}\mu(\Psi)\|x - x_i\|^\rho] :\ x \in Q \right\} \\ &\geq \tfrac{1}{N+1}\max_x \left\{ \sum_{i=0}^{N} [f(x_i) - \Psi(x, w(x_i))] :\ x \in Q \right\} \\ &\geq f(\bar{x}_N) - \min_{x\in Q} \Psi(x, \bar{w}_N) \ = \ f(\bar{x}_N) - \eta(\bar{w}_N). \end{aligned}$$

∎

Let us prove now several auxiliary results. Let $l(x)$ be an affine function on E. Let us fix a point $\bar{y} \in Q$. Consider the function

$$\psi(r) = \max_x \{ l(x) :\ x \in Q_r(\bar{y})\}, \quad r \geq 0.$$

Note that $\psi(r)$ is an increasing concave function of r and

$$\psi(r) \geq \psi(0) = l(\bar{y}).$$

Let us fix some $\bar{r} > 0$ and choose an arbitrary $\bar{x} \in Q_{\bar{r}}(\bar{y})$. For some $\mu > 0$ define

$$\lambda_\mu^*(x) = \max_y \{ l(y) - \tfrac{1}{2}\mu\|y - x\|^\rho :\ y \in Q\}. \qquad (34)$$

We need to bound from above the value $\lambda_\mu^*(\bar{x})$.

Lemma 3. *For any* $b > 0$ *we have*

$$\lambda_\mu^*(\bar{x}) \leq \lambda^*_{(1+b)^{1-\rho}\mu}(\bar{y}) + \frac{\mu}{2b^{\rho-1}}\bar{r}^\rho. \qquad (35)$$

Proof. Consider $y_\mu(\bar{x})$, the optimal solution of optimization problem in (34) with $x = \bar{x}$. Then

$$\lambda_\mu^*(\bar{x}) = l(y_\mu(\bar{x})) - \tfrac{1}{2}\mu\|y_\mu(\bar{x}) - \bar{x}\|^\rho.$$

On the other hand, for any $b > 0$,

$$\begin{array}{rcl} \|y_\mu(\bar{x}) - \bar{y}\|^\rho & \leq & (\|y_\mu(\bar{x}) - \bar{x}\| + \|\bar{x} - \bar{y}\|)^\rho \\ & \leq & (1+b)^{\rho-1}\|y_\mu(\bar{x}) - \bar{x}\|^\rho + (1+b^{-1})^{\rho-1}\|\bar{x} - \bar{y}\|^\rho \\ & \leq & (1+b)^{\rho-1}\|y_\mu(\bar{x}) - \bar{x}\|^\rho + (1+b^{-1})^{\rho-1}\bar{r}^\rho. \end{array}$$

Hence,

$$\lambda_\mu^*(\bar{x}) \leq l(y_\mu(\bar{x})) - \tfrac{\mu}{2}\tfrac{\|y_\mu(\bar{x}) - \bar{y}\|^\rho}{(1+b)^{\rho-1}} + \tfrac{1}{2b^{\rho-1}}\bar{r}^\rho \leq \lambda^*_{(1+b)^{1-\rho}\mu}(\bar{y}) + \tfrac{\mu}{2b^{\rho-1}}\bar{r}^\rho.$$

∎

Lemma 4.

$$\lambda_\mu^*(\bar{y}) \leq \psi(\bar{r}) + \frac{\rho - 1}{\rho}\left(\frac{2}{\mu\rho}\right)^{\frac{1}{\rho-1}}\left(\frac{\psi(\bar{r}) - \psi(0)}{\bar{r}}\right)^{\frac{\rho}{\rho-1}}.$$

Proof. Indeed, denote $\hat{t} = \|y_\mu(\bar{y}) - \bar{y}\|$. Then

$$\lambda_\mu^*(\bar{y}) \;=\; l(y_\mu(\bar{y})) - \tfrac{1}{2}\mu\hat{t}^\rho \;\leq\; \psi(\hat{t}) - \tfrac{1}{2}\mu\hat{t}^\rho \;\leq\; \max_{t\geq 0}\{\psi(t) - \tfrac{1}{2}\mu t^\rho\}.$$

Since $\psi(t)$ is concave,

$$\psi(t) \;\leq\; \psi(\bar{r}) + \psi'(\bar{r})(t - \bar{r}) \;\leq\; \psi(\bar{r}) + \psi'(\bar{r})t.$$

Note that

$$\psi'(\bar{r})t - \tfrac{1}{2}\mu t^\rho \leq \frac{\rho - 1}{\rho}\left(\frac{2\psi'(\bar{r})^\rho}{\mu\rho}\right)^{\frac{1}{\rho-1}},$$

thus

$$\lambda_\mu^*(\bar{y}) \leq \psi(\bar{r}) + \frac{\rho - 1}{\rho}\left(\frac{2\psi'(\bar{r})^\rho}{\mu\rho}\right)^{\frac{1}{\rho-1}}.$$

On the other hand,

$$\psi(0) \leq \psi(\bar{r}) + \psi'(\bar{r})(0 - \bar{r}).$$

Thus, $\psi'(\bar{r}) \leq \frac{1}{\bar{r}}(\psi(\bar{r}) - \psi(0))$. ∎

When substituting the result into (35) we obtain

Corollary 2.

$$\lambda_\mu^*(\bar{x}) \;\leq\; \psi(\bar{r}) + (1+b)\tfrac{\rho-1}{\rho}\left(\tfrac{2}{\mu\rho}\right)^{\frac{1}{\rho-1}}\left(\tfrac{\psi(\bar{r}) - \psi(0)}{\bar{r}}\right)^{\frac{\rho}{\rho-1}} + \tfrac{\mu}{2b^{\rho-1}}\bar{r}^\rho. \tag{36}$$

Let us apply now the above results to Algorithm 2. Let us choose $\mu = \mu(\Psi)$,

$$\bar{y} = y_{m-1}, \quad \bar{x} = y_m, \quad \bar{r} = R_{m-1}, \quad l(x) = \tfrac{1}{1+N_m}\sum_{i=0}^{N_m}\langle f'(x_i), x_i - x\rangle,$$

where the points $\{x_i\}_{i=0}^{N_m}$ were generated during the last mth stage of the algorithm. Note that

$$2^m \;\; \geq \;\; \tfrac{\mu(\Psi)}{\epsilon} R_0^{\rho} \; \geq \; 2^{m-1}. \tag{37}$$

Therefore

$$\tfrac{2\epsilon}{\mu(\Psi)} \; \geq \; \bar{r}^{\rho} \;\; = \;\; 2^{1-m} R_0^{\rho} \; \geq \; \tfrac{\epsilon}{\mu(\Psi)}. \tag{38}$$

Further,

$$\begin{aligned}
\psi(\bar{r}) \;\; &= \;\; \delta_{N_m}(y_{m-1}, R_{m-1}) \; \stackrel{(19)}{\leq} \; \mu(\Psi) 2^{-m} R_0^{\rho} \; \stackrel{(37)}{\leq} \; \epsilon, \\
\psi(0) \;\; &= \;\; \tfrac{1}{1+N_m} \sum_{i=0}^{N_m} \langle f'(x_i), x_i - y_{m-1} \rangle \; \geq \; \tfrac{1}{1+N_m} \sum_{i=0}^{N_m} [f(x_i) - f(y_{m-1})] \\
&\geq \;\; f^* - f(y_{m-1}) \; \stackrel{(19)}{\geq} \; -\mu(\Psi) 2^{1-m} R_0^{\rho} \; \stackrel{(38)}{\geq} \; -2\epsilon.
\end{aligned}$$

Hence, using the above inequalities in (36), we obtain

$$\begin{aligned}
\lambda^*_{\mu(\Psi)}(y_m) \;\; &\leq \;\; \epsilon + (1+b) \tfrac{\rho-1}{\rho} \left(\tfrac{2}{\mu(\Psi)\rho} \right)^{\frac{1}{\rho-1}} \left(\tfrac{3\epsilon}{\bar{r}} \right)^{\frac{\rho}{\rho-1}} + \tfrac{\mu(\Psi)}{2b^{\rho-1}} \bar{r}^{\rho} \\
&\leq \;\; \epsilon + \tfrac{(1+b)(\rho-1)}{2} \left(\tfrac{6}{\rho} \right)^{\frac{\rho}{\rho-1}} \kappa^{-\frac{1}{\rho-1}} \epsilon + \tfrac{\kappa \epsilon}{2b^{\rho-1}} \\
&\leq \;\; \epsilon \left(1 + \tfrac{(1+b)(\rho-1)}{2} \left(\tfrac{6}{\rho} \right)^{\frac{\rho}{\rho-1}} + b^{1-\rho} \right),
\end{aligned} \tag{39}$$

where we set $\kappa = \frac{\mu(\Psi)\bar{r}^{\rho}}{\epsilon}$ and used the fact that $1 \leq \kappa \leq 2$ due to (38).

When setting $b = (\frac{\rho}{6})^{\frac{1}{\rho-1}} 2^{\frac{1}{\rho}}$ we obtain

$$\lambda^*_{\mu(\Psi)}(y_m) \leq \epsilon \left(1 + 3 \, \frac{6^{\frac{1}{\rho-1}} + 2^{\frac{1}{\rho}} \rho^{\frac{1}{\rho-1}}}{\rho^{\frac{\rho}{\rho-1}}} + \frac{6}{2^{\frac{\rho-1}{\rho}} \rho} \right).$$

Note that a finer estimate can be obtained for $\rho = 2$. To this end it suffices to verify that for the choice $b = 1/3$ the right-hand side of (39) is decreasing in κ for $0 \leq \kappa \leq 2$. Therefore,

$$\lambda^*_{\mu(\Psi)}(y_m) \;\; \leq \;\; 8.5 \, \epsilon.$$

It remains to note that

$$\lambda^*_{\mu(\Psi)}(y_m) \;\; = \;\; \max_y \left\{ \tfrac{1}{1+N_m} \sum_{i=0}^{N_m} \langle f'(x_i), x_i - y \rangle - \tfrac{1}{2} \mu(\Psi) \| y - y_m \|^{\rho} : \; y \in Q \right\},$$

and $y_m = x_{N_m}(y_{m-1}, R_{m-1}) \stackrel{(13)}{=} \frac{1}{1+N_m} \sum\limits_{i=0}^{N_m} x_i$. Thus, applying Lemma 2, we come to the following statement:

Theorem 4. *Let assumptions of Theorem 1 hold and let $\widehat{x}_{\epsilon}(y_0, R_0)$ be the approximate solution, supplied by Algorithm 2. Define $\bar{w}_{N_m} = \frac{1}{1+N_m} \sum\limits_{i=0}^{N_m} w(x_i)$. Then*

$$f(\widehat{x}_{\epsilon}(y_0, R_0)) - \eta(\bar{w}_{N_m}) \leq C(\rho) \, \epsilon,$$

where

$$C(\rho) \leq \left(1 + 3\,\frac{6^{\frac{1}{\rho-1}} + 2^{\frac{1}{\rho}}\rho^{\frac{1}{\rho-1}}}{\rho^{\frac{\rho}{\rho-1}}} + \frac{6}{2^{\frac{\rho-1}{\rho}}\rho}\right).$$

Furthermore, when the objective f is strongly convex ($\rho = 2$), $f(\widehat{x}_\epsilon(y_0, R_0)) - \eta(\bar{w}_{N_m}) \leq 8.5\,\epsilon$.

5 Stochastic programming with uniformly convex objective

In order to rewrite the results of Sections 3 in the stochastic framework we substitute for $f'(x_k)$ its observation $g_k = f'(x_k) + \xi_k$ into the iteration of Algorithm 1. The following statement is a stochastic counterpart of Proposition 1:

Proposition 2. *Let (x_k), $k = 0, 1, ...$ be the search points of Algorithm 1 with g_k substituted for $f'(x_k)$. Then for any $x \in Q \cap B_R(\bar{x})$,*

$$\sum_{i=1}^{k} \lambda_i \langle f'(x_i), x_i - x\rangle \leq d_{\bar{x},R}(x)\beta_{k+1} + \frac{R^2}{2\mu(d)} \sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_i} \|f'(x_i)\|_*^2 + \sum_{i=0}^{k} \zeta_i, \tag{40}$$

where

$$\|\zeta_i\|_* \leq 2\lambda_i \|\xi_i\|_* R, \quad \zeta_i \leq -\lambda_i\langle \xi_i, \tilde{x}_i - x\rangle + \frac{R^2\lambda_i^2\|\xi_i\|_*^2}{2\mu(d)\beta_i}, \tag{41}$$

and $(\tilde{x}_i), i = 1, ..., k$ are $\mathcal{F}_{i-1}$-measurable random vectors, $\tilde{x}_i \in Q \cap B_R(\bar{x})$

In this section we propose two families of multi-stage methods for uniformly convex stochastic programming problem described in Section 2.2. The first one is based on the dual averaging scheme with the prox-function which satisfies the condition (9) of quadratic growth. As we shall see immediately, one can easily obtain the bounds for the average value of the objective at the approximate solution and is a stochastic counterpart of Algorithm 4 and 5. On the other hand, the methods derived from those, presented in Section 3.2 better suit the case when the confidence bounds on the error of the approximate solutions are required.

5.1 Expectation bounds for methods with prox-function of quadratic growth

When taking the expectation with respect to the distribution of ξ_i we obtain the following simple counterpart of Lemma 2:

Lemma 3. *Let $\bar{x} \in Q$ satisfy $\mathbf{E}\|\bar{x} - x^*\|^2 \leq R^2$, where x^* is the optimal solution of problem (12), and let $\lambda_k = 1$ and $\beta_k = \gamma\sqrt{N+1}$, $k = 0, .., N$. Suppose that Assumptions 2 and 3 hold. Then the approximate solution supplied by Algorithm 1 satisfies*

$$\begin{aligned}
\mathbf{E}f(x_N(\bar{x}, R)) - f^* &\leq \frac{1}{N+1}\sum_{i=0}^{N} \mathbf{E}\langle f'(x_i), x_i - x^*\rangle \\
&\leq \frac{1}{\sqrt{N+1}}\left(\gamma \mathbf{E} d_{\bar{x},R}(x^*) + \frac{R^2(L^2+\sigma^2)}{2\mu(d)\gamma}\right), \\
\mathbf{E}\|x_N(\bar{x}, R) - x^*\|^\rho &\leq \frac{1}{\mu(\Psi)\sqrt{N+1}}\left(\gamma \mathbf{E} d_{\bar{x},R}(x^*) + \frac{R^2(L^2+\sigma^2)}{2\mu(d)\gamma}\right).
\end{aligned}$$

Suppose now that $\mathbf{E}\|\bar{x}-x^*\|^2 \le r^2$. Using the relation $d_{\bar{x},R}(x^*) \le C(d)\frac{r^2}{R^2}$ we get the following (cf Corollary 2)

Corollary 4. *Suppose that $\bar{x} \in Q$ satisfy*

$$\mathbf{E}\|\bar{x}-x^*\|^2 \le r^2,$$

and let

$$\gamma = \frac{R^2}{r}\sqrt{\frac{L^2+\sigma^2}{2C(d)\mu(d)}},$$

Then

$$\mathbf{E}f(x_N(\bar{x},R)) - f^* \le r\sqrt{\frac{2C(d)(L^2+\sigma^2)}{\mu(d)(N+1)}}, \tag{42}$$

$$\mathbf{E}\|x_N(\bar{x},R)-x^*\|^\rho \le \frac{r}{\mu(f)}\sqrt{\frac{2C(d)(L^2+\sigma^2)}{\mu(d)(N+1)}}, \tag{43}$$

When comparing the above statement to the result of Corollary 2 we observe that the only difference between the two is that in Corollary 4 the quantity L^2 is substituted with $L^2+\sigma^2$. When modifying in the same way the parameters of Algorithm 5 we obtain the multistage procedure for the stochastic problem.

Assume that the parameters L, ρ, $\mu(f)$ and the diameter R_0 of Q are known. The method starts from an arbitrary $x_0 \in Q$.

Algorithm 7.

Initialization: Set $y_0 = x_0$, $\tau = \frac{2(\rho-1)}{\rho}$ and $m = \lfloor \log_2 \frac{\mu(f)}{\epsilon} R_0^\rho \rfloor + 1$.

Stage $k = 1, \ldots, m$:

1. Define $N_k = \lfloor 2^{\tau k} \frac{4(L^2+\sigma^2)C(d)}{\mu^2(f)\mu(d)R_0^{2(\rho-1)}} \rfloor$ and $r_k^\rho = 2^{-k}R_0^\rho$.
2. Compute $y_k = x_{N_k}(y_{k-1}, R_0)$ with $\gamma_k = \frac{R_0^2}{r_{k-1}}\sqrt{\frac{L^2+\sigma^2}{2C(d)\mu(d)}}$.

Output: Set the approximate solution $\widehat{x}_\epsilon = y_m$.

We have the following stochastic analogue of Theorem 2:

Theorem 5. *Suppose that*

$$N \ge N(\epsilon) = \frac{4^{\tau+1}(L^2+\sigma^2)C(d)}{\mu(f)^{\frac{2}{\rho}}\mu(d)}\epsilon^{-\tau}.$$

Then the approximate solution $\widehat{x}_N$, provided by Algorithm 7 satisfies:

$$\mathbf{E}f(\widehat{x}_\epsilon) - f^* \le \epsilon.$$

The proof of the theorem follows the lines of that of Theorem 2. It suffices to substitute the bounds (42) and (43) for those of (22) and (23). We leave this simple exercise to the reader.

The method can be rewritten for the case when the total number N of calls to the oracle is fixed.

Suppose that

$$N \geq 2^{\tau}(2^{\tau}+1)\frac{4(L^2+\sigma^2)C(d)}{\mu^2(f)\mu(d)R_0^{2(\rho-1)}}.$$

Consider the following procedure:

Algorithm 8.

Initialization: Set $y_0 = x_0$, $\tau = \frac{2(\rho-1)}{\rho}$, compute $N_j = \left\lfloor 2^{\tau j}\frac{4(L^2+\sigma^2)C(d)}{\mu(f)^2\mu(d)R_0^{2(\rho-1)}}\right\rfloor$, while $\sum_j N_j \leq N$. Set $m(N) = \max\{k:\ \sum_{j=1}^{k} N_j \leq N\}$.

Stage $k = 1, \ldots, m(N)$:

Set $r_k^{\rho} = 2^{-k}R_0^{\rho}$. Compute $y_k = x_{N_k}(y_{k-1}, R_0)$ with $\gamma_k = \frac{R_0^2}{r_{k-1}}\sqrt{\frac{L^2+\sigma^2}{2C(d)\mu(d)}}$.

Termination: Set the approximate solution $\widehat{x}_N = y_{m(N)}$.

Corollary 5. *We have*

$$\mathbf{E}f(\widehat{x}_N) - f^* \leq 2\left(\frac{8(L^2+\sigma^2)C(d)}{\mu(f)^{\frac{2}{\rho}}\mu(d)N}\right)^{1/\tau}.$$

Exactly in the same way it was done in the deterministic settings, we can provide an adaptive version of the method. To this end the adaptive method of Algorithm 6 for deterministic problem should be slightly modified: we have to change the way the approximate solution $\widehat{x}_N$ is formed, as the exact observations of the objective function are not available anymore. Fortunately, we can take as the output of the algorithm the approximate solution y_m, generated at the last stage.

Consider the following procedure:

Algorithm 9.

Initialization: Set $y_0 = x_0$, $m = \left[\frac{1}{2}\log_2\frac{\mu(d)N}{C(d)\log_2 N}\right] - 1$, $N_0 = [N/m]$, $r_k = 2^{-k}R_0$, $k = 1, ..., m$.

Stage $k = 1, ..., m$: Compute $y_k = x_{N_0}(y_{k-1}, R_0)$ with with $\gamma_k = \frac{R_0^2}{r_{k-1}}\sqrt{\frac{L^2+\sigma^2}{2C(d)\mu(d)}}$.

Termination: Set the approximate solution $\widehat{x}_N = y_m$.

Theorem 6. *The approximate solution $\widehat{x}_N$, supplied by Algorithm 9, satisfies for $N > 4$:*

$$\mathbf{E}f(\widehat{x}_N) - f^* \leq 4\left(\frac{16(L^2+\sigma^2)C(d)\log_2 N}{\mu(f)^{\frac{2}{\rho}}\mu(d)N}\right)^{\frac{\rho}{2(\rho-1)}}.$$

Proof: The proof of the theorem follows exactly the lines of that of Theorem 3. Using the notation k^*, introduced in Lemma 1, we get (cf. (30))

$$\mathbf{E}f(y_m) \leq \mathbf{E}f(y_{k^*}) + \mu_{k^*} r_{k^*}^{\rho}.$$

Thus

$$\mathbf{E}f(y_m) - f^* \leq 2\mu_{k^*} r_{k^*}^{\rho} \leq 4\left(\frac{16(L^2+\sigma^2)C(d)\log_2 N}{\mu(f)^{\frac{2}{\rho}}\mu(d)N}\right)^{\frac{\rho}{2(\rho-1)}}.$$

∎

5.2 Confidence sets for uniformly convex stochastic programs

In this section we establish confidence bounds for the approximate solutions, delivered by multistage stochastic algorithms. Consider dual averaging Algorithm 1 in which we substitute the exact subgradient with the observation $g_k = f'(x_k) + \xi_k$. Let $\delta_N(\bar{x}, R)$ be the gap value, defined in (15).

Proposition 3. *Let $\bar{x}$ be a point of Q, $\lambda_k = 1$ and $\beta_k = \gamma\sqrt{N+1}$, $k = 0,..,N$. Suppose that Assumptions 2–4 hold. Then*

$$\mathrm{Prob}_{\bar{x}}\left[\delta_N(\bar{x}, R) \geq \frac{1}{\sqrt{N+1}}\left(\gamma A(d) + \frac{R^2(L^2+\sigma^2)}{2\gamma\mu(d)}\right) + 2R\sigma\sqrt{\frac{3\ln\alpha^{-1}}{N+1}}\right] \leq \alpha. \tag{44}$$

Proof: We need the following result which is essentially known (cf [3]):

Lemma 4. *Let ψ_i, $i = 0, ..., N$, be Borel functions on Ω such that ψ_i is $\mathcal{F}_i$-measurable, and let $\mu_i \geq 0$, $\nu_i > 0$ be deterministic reals. Assume that for all $i = 0, 1, 2, ...$ one has a.s.*

$$\mathbf{E}_{i-1}[\psi_i] \leq \mu_i, \quad \mathbf{E}_{i-1}[\exp\{\psi_i^2/\nu_i^2\}] \leq \exp\{1\},$$

Then for every $\Lambda \geq 0$

$$\mathrm{Prob}\left[\sum_{i=0}^{N}\psi_i > \sum_{i=0}^{N}\mu_i + \Lambda\sqrt{\sum_{i=0}^{N}\nu_i^2}\right] \leq \exp\{-\Lambda^2/3\} \tag{45}$$

Proof: It is immediately seen that $\exp\{s\} \leq s + \exp\{9s^2/16\}$ for all s. We conclude that if $0 \leq t \leq \frac{4}{3\nu_i}$, then

$$\begin{aligned}\mathbf{E}_{i-1}[\exp\{t\psi_i\}] &\leq t\mu_i + \mathbf{E}_{i-1}[\exp\{9t^2\psi_i^2/16\}]\\ &\leq t\mu_i + \exp\{9t^2\nu_i^2/16\} \leq \exp\{t\mu_i + 9t^2\nu_i^2/16\}.\end{aligned} \tag{46}$$

Besides this, we have $tx \leq \frac{3t^2\nu_i^2}{8} + \frac{2x^2}{3\nu_i^2}$, so that

$$\mathbf{E}_{i-1}[\exp\{t\psi_i\}] \leq \exp\left\{\frac{3t^2\nu_i^2}{8} + \frac{2}{3}\right\},$$

and the latter quantity is $\leq \exp\{3t^2\nu_i^2/4\}$ when $t \geq \frac{4}{3\nu_i}$. Invoking (46), we arrive at

$$\mathbf{E}_{n-1}[\exp\{t\phi_n\}] \leq \exp\{t\mu_i + 3t^2\nu_n^2/4\}$$

for any $t \geq 0$. It follows that

$$\begin{aligned}\mathbf{E}\left[\exp\left\{t\sum_{i=0}^{n}\psi_i\right\}\right] &= \mathbf{E}\left[\mathbf{E}_{n-1}\left[\exp\left\{t\sum_{i=0}^{n}\psi_i\right\}\right]\right]\\ &\leq \mathbf{E}\left[\exp\left\{t\sum_{i=0}^{n-1}\psi_i\right\}\right]\exp(t\mu_n + 3t^2\nu_n^2/4),\end{aligned}$$

whence for $t \geq 0$,

$$\mathbf{E}\left[\exp\left\{t\sum_{i=0}^{N}\psi_i\right\}\right] \leq \exp\left\{t\sum_{i=0}^{N}\mu_i + \frac{3t^2}{4}\sum_{i=0}^{N}\nu_i^2\right\}.$$

Therefore for $\Lambda \geq 0$ we get

$$\begin{aligned}&\mathrm{Prob}\left[\sum_{i=0}^{N}\psi_i > \sum_{i=0}^{N}\mu_i + \Lambda\sqrt{\sum_{i=0}^{N}\nu_i^2}\right]\\ &\leq \min_{t>0}\left[\mathbf{E}\left[\exp\left\{t\sum_{i=0}^{N}\psi_i\right\}\right]\exp\left\{-t\sum_{i=0}^{N}\mu_i - t\Lambda\sqrt{\sum_{i=0}^{N}\nu_i^2}\right\}\right]\\ &\leq \min_{t>0}\exp\left\{t\sum_{i=0}^{N}\mu_i + \frac{3t^2}{4}\sum_{i=0}^{N}\nu_i^2 - t\sum_{i=0}^{N}\mu_i - t\Lambda\sqrt{\sum_{i=0}^{N}\nu_i^2}\right\} = \exp\{-\Lambda^2/3\}\end{aligned}$$

as required in (45). ∎

Let us return to the proof of the proposition. From (41) and Assumption 4 we conclude that

$$\mathbf{E}_{i-1}\zeta_i \leq \frac{R^2\lambda_i^2\sigma^2}{2\mu(d)\beta_i} = \frac{R^2\sigma^2}{2\mu(d)\beta_i}$$

(recall that $E_{i-1}\xi_i = 0$ and $\tilde{x}_i$ is $\mathcal{F}_{i-1}$-measurable). Along with Assumption 4 this implies that random variables $\psi_i = \zeta_i$ satisfy the premises of Lemma 4 with $\mu_i = \frac{R^2\sigma^2}{2\mu(d)\beta_i}$ and $\nu_i = 2R\sigma$. Thus by (45),

$$\mathrm{Prob}\left[\sum_{i=0}^{N}\zeta_i \geq \frac{R^2\sigma^2}{2\mu(d)}\sum_{i=0}^{N}\beta_i^{-1} + 2\Lambda R\sigma\sqrt{N+1}\right] \leq \exp\{-\Lambda^2/3\}\ \ (=\alpha \text{ for } \Lambda = \sqrt{3\ln\alpha^{-1}}).$$

When substituting $\beta_i = \gamma\sqrt{N+1}$ we conclude (44) from (40). ∎

From (44) we obtain immediately:

Corollary 6. *Let $\bar{x}$ be a point of Q. Let*

$$\gamma = R\sqrt{\frac{L^2+\sigma^2}{2\mu(d)A(d)}}.$$

Then for all $\alpha \geq 0$, the approximate solution $x_N(\bar{x}, R)$ of Algorithm 1 satisfies

$$\mathrm{Prob}_{\bar{x}}\left[\delta_N(\bar{x}, R) \leq 2R\left[\sqrt{\frac{A(d)(L^2+\sigma^2)}{2\mu(d)(N+1)}} + \sigma\sqrt{\frac{\ln 3\alpha^{-1}}{N+1}}\right]\right] \geq 1-\alpha. \tag{47}$$

Corollary 6 allows us to compute the confidence sets for approximate solutions, provided by stochastic analogues of Algorithms 2 and 3 exactly in the same way as it was done in Section 3.2. For the sake of conciseness we present here only the result for the setting when the total number N of subgradient observations is fixed and the convexity parameters of the objective are unknown.

Algorithm 10.

Initialization: Set $y_0 = x_0$, $m = \left[\frac{1}{2}\log_2 \frac{\mu(d)N}{A(d)\log_2 N}\right] - 1$, $N_0 = [N/m]$, and

$$R_k = 2^{-k}R_0, \quad k = 1, ..., m.$$

Stage $k = 1, ..., m$: Compute $y_k = \widehat{x}_{N_0}(y_{k-1}, R_{k-1})$ with $\gamma_k = R_{k-1}\sqrt{\frac{N_0(L^2+\sigma^2)}{2\mu(d)A(d)}}$.

Output: $\widehat{x}_N = y_m$.

Theorem 7. *Let $\alpha \geq 0$. Then the approximate solution $\widehat{x}_N$ satisfies for $N \geq 4$*

$$\mathrm{Prob}\left[f(\widehat{x}_N) - f^* \leq \epsilon(N, \alpha)\right] \geq 1-\alpha,$$

where

$$\epsilon(N,\alpha) = 4\left(\frac{16}{(N_0+1)\mu(f)^{\frac{2}{\rho}}}\right)^{\frac{\rho}{2(\rho-1)}}\left(\sqrt{\frac{(L^2+\sigma^2)A(d)}{2\mu(d)}} + \sigma\sqrt{3\ln\left(\frac{\log_2 N}{2\alpha}\right)}\right)^{\frac{\rho}{\rho-1}}.$$

Proof: Let us denote $\bar{\alpha} = \frac{2\alpha}{\log_2 N}$ and

$$a(N_0, \bar{\alpha}) = \frac{2}{\sqrt{N_0+1}}\left(\sqrt{\frac{(L^2+\sigma^2)A(d)}{2\mu(d)}} + \sigma\sqrt{3\ln\bar{\alpha}^{-1}}\right).$$

We set

$$\mu_0 = 2R_0^{1-\rho}a(N_0, \bar{\alpha}) \quad \text{and} \quad \mu_k = 2^{(\rho-1)k}\mu_0, \quad k = 1, ..., m. \tag{48}$$

Note also that

$$\mu(f) \quad \leq \quad \frac{L}{R_0^{\rho-1}},$$

and by the definition of μ_0 and m we have $\mu(f) \leq \mu(m)$. Suppose first that the true $\mu(f)$ satisfies $\mu_0 \leq \mu(f) \leq \mu_m$. We start with the following auxiliary result.

Lemma 5. *Let k^* satisfy $\mu_{k^*} \le \mu(f) \le 2^{\rho-1}\mu_{k^*}$. Then for any $1 \le k \le k^*$, there exists a set $\mathcal{A}_k \subset \Omega$ of probability at least $1-k\bar\alpha$ such that for $\omega \in \mathcal{A}_k$ the points $\{y_k\}_{k=1}^m$ generated by Algorithm 10 satisfy*

$$\begin{aligned} \|y_{k-1}-x^*\| &\le R_{k-1} = 2^{-k+1}R_0, && (49)\\ f(y_k)-f^* &\le \mu_k R_k^\rho = 2^{-k}\mu_0 R_0^\rho. && (50) \end{aligned}$$

Further, for $k > k^*$ there is a set $\mathcal{C}_k \subset \Omega$ of probability at least $1-(k-k^*)\bar\alpha$ such that on $\mathcal{C}_k$

$$f(y_k) \le f(y_{k^*}) + \mu_{k^*}R_{k^*}^\rho. \tag{51}$$

Proof. Note that for $k=1$ (49) is valid. Assume it is valid for some $k \ge 1$. Note that by (47) of Corollary 6 there exists a random set, let us call it $\mathcal{B}_k$, such that $\text{Prob}[\mathcal{B}_k] \ge 1-\bar\alpha$ and on $\mathcal{B}_k$,

$$\begin{aligned} \delta_N(y_{k-1},R_{k-1}) &\le 2R_{k-1}\left(\sqrt{\frac{(L^2+\sigma^2)A(d)}{2\mu(d)(N_0+1)}} + \sigma\sqrt{\frac{3\ln\bar\alpha^{-1}}{N_0+1}}\right)\\ &= R_{k-1}a(N_0,\bar\alpha) \overset{(48)}{=} \frac{1}{2}\mu_k 2^{-(\rho-1)k}R_0^{\rho-1}R_{k-1} = \mu_k R_k^\rho. \end{aligned} \tag{52}$$

On the other hand, by our inductive hypothesis, $\|y_{k-1}-x^*\| \le R_{k-1}$ on $\mathcal{A}_{k-1}$. Let $\mathcal{A}_k = \mathcal{A}_{k-1}\cap\mathcal{B}_k$. Note that

$$\text{Prob}[\mathcal{A}_k] \ge \text{Prob}[\mathcal{A}_{k-1}] + \text{Prob}[\mathcal{B}_k] - 1 \ge 1-k\bar\alpha,$$

and we have on A_k:

$$\begin{aligned} f(y_k)-f^* &\le \delta_N(y_{k-1},R) \le \mu_k R_k^\rho,\\ \|y_k-x^*\|^\rho &\le \frac{\delta_N(y_{k-1},R_{k-1})}{\mu(f)} \le R_k^\rho, \end{aligned}$$

what is (50) and (49) for $k+1$.

To show (51) notice that, we have for $k > k^*$ (cf. (52))

$$f(y_k) - f(y_{k-1}) \le \delta_{N_0}(y_{k-1},R_{k-1}) \le \mu_k R_k^\rho$$

on some $\mathcal{B}_k \subset \Omega$ such that $\text{Prob}[\mathcal{B}_k] \ge 1-\bar\alpha$. Then we have on $\mathcal{C}_k = \cap_{j=k^*+1}^k \mathcal{B}_j$:

$$f(y_k) - f(y_{k^*}) = \sum_{j=k^*+1}^k f(y_j) - f(y_{j-1}) \le \sum_{j=k^*+1}^k 2^{k^*-j}\mu_{k^*}R_{k^*}^\rho \le \mu_{k^*}R_{k^*}^\rho.$$

Note that $\text{Prob}[\mathcal{C}_k] \ge 1-(k-k^*)\bar\alpha$. This proves the lemma. ∎

Now we can finish the proof of the theorem. Let $\mu_0 \le \mu(f) \le \mu_m$. At the end of the k^*-th stage we have on the set $\mathcal{A}_{k^*}$ of probability at least $1-k^*\bar\alpha$:

$$f(y_{k^*}) - f^* \le \delta_{N_0}(y_{k^*-1},R_{k^*-1}) \le \mu_{k^*}R_{k^*}^\rho.$$

Then on the set $\mathcal{A}_{k^*} \cap \mathcal{C}_m$ such that $\mathrm{Prob}[\mathcal{A}_{k^*} \cap \mathcal{C}_m] \geq 1 - m\bar{\alpha}$ (cf (51)) we have

$$\begin{aligned} f(y_m) - f^* &\leq 2\mu_{k^*} R_{k^*}^\rho \leq 4 \frac{\mu_{k^*}^{\frac{1}{\rho-1}}}{\mu(f)^{\frac{1}{\rho-1}}} \mu_{k^*} R_{k^*}^\rho \\ &= 4 \frac{\mu_0^{\frac{\rho}{\rho-1}} R_0^\rho}{\mu(f)^{\frac{1}{\rho-1}}} \overset{(48)}{\leq} 4 \left(\frac{2a(N_0, \bar{\alpha})}{\mu(f)^{\frac{1}{\rho}}} \right)^{\frac{\rho}{\rho-1}}. \end{aligned}$$

It suffices to recall now that by the definition of m, $m \leq \frac{1}{2} \log_2 N$, thus $m\bar{\alpha} \leq \alpha$.

If $\mu(f) < \mu(0)$, we have on $\mathcal{A}_1 = \mathcal{B}_1$ (cf. (52)):

$$\begin{aligned} f(y_1) - f^* &\leq R_0 a(N_0, \bar{\alpha}) = \frac{R_0}{a(N_0, \bar{\alpha})^{\frac{1}{\rho-1}}} a(N_0, \bar{\alpha})^{\frac{\rho}{\rho-1}} \\ &= 2^{\frac{1}{\rho-1}} \frac{a(N_0, \bar{\alpha})^{\frac{\rho}{\rho-1}}}{\mu_0^{\frac{1}{\rho-1}}} \leq 2^{\frac{1}{\rho-1}} \left(\frac{a(N_0, \bar{\alpha})}{\mu(f)^{\frac{1}{\rho}}} \right)^{\frac{\rho}{\rho-1}}. \end{aligned}$$

Finally, we conclude using (51): on $\mathcal{A}_1 \cap \mathcal{C}_m$ we have

$$f(y_m) - f^* \leq 2R_0 a(N_0, \bar{\alpha}) \leq \left(\frac{2a(N_0, \bar{\alpha})}{\mu(f)^{\frac{1}{\rho}}} \right)^{\frac{\rho}{\rho-1}}.$$

∎

6 Computational issues

The interest of the proposed algorithmic schemes is conditioned by our ability to compute efficiently the optimal solution $\pi_{z,R,\beta}(s)$ of the optimization problem (11). We present here two important examples in which the problem (11) can be solved quite efficiently. These are the standard simplex and the hyperoctahedron settings.

Let us measure the distances in $E = \mathbb{R}^n$ in l_1-norm:

$$\|x\| = \|x\|_1 = \sum_{i=1}^n |x^{(i)}|.$$

6.1 Simplex setup

Let $n \geq 2$ and let

$$Q = \{x \in \mathbb{R}^n | \, x \geq 0, \|x\|_1 = 1\}$$

be the standard simplex. We are to show how the problem (11) can be solved in this case. The problem (11) on $Q_R(z)$ for the function d as in (10) writes

$$\begin{aligned} &\min_{x,u,v} \{ \textstyle\sum_{i=1}^n [s_i x_i + u_i \ln u_i + v_i \ln v_i] : \quad \sum_{i=1}^n [u_i + v_i] = R, \quad \sum_{i=1}^n x_i = 1, \\ &x_i = z_i + u_i - v_i, \quad u_i \geq 0, \quad v_i \geq 0, \quad x_i \geq 0, \quad i = 1, ..., n. \} \end{aligned}$$

When eliminating the "x" variable and dualizing the coupling constraints we obtain the equivalent problem

$$\max_{\lambda,\mu} \left\{ \underline{L}(\lambda,\mu) \equiv \min_{u,v} L(u,v,\lambda,\mu) \;:\; z_i + u_i - v_i \geq 0, \;\; i = 1,...,n \right\}, \tag{53}$$

where

$L(u,v,\lambda,\mu) = \sum_{i=1}^n [r_i v_i + t_i u_i + u_i \ln u_i + v_i \ln v_i] - \lambda R - \mu \;:$

$r_i = s_i + \lambda - \mu, \;\; t_i = -s_i + \lambda + \mu.$

The dual problem (54) can be solved using a conventional method of convex optimization (ellipsoid or level), given the solution of the problem

$$\min_{u,v} \{ L(u,v,\lambda,\mu) \;:\; z_i + u_i - v_i \geq 0, \;\; i = 1,...,n \}.$$

Note that the latter problem an be decomposed into n 2-dimensional problems

$$\min_{u,v} \; su + tv + u \ln u + v \ln v, \;\; u \geq v - z. \tag{54}$$

One way to compute the minimizer is to compute the solution $(\bar{u},\bar{v})$ to the problem

$$\min_{u,v} \; [\psi(u,v) = su + tv + u \ln u + v \ln v], \;\; u = v - z,$$

namely,

$$\bar{u} = \frac{1}{2}\left(\sqrt{z^2 + 4e^{-2-s-t}} - z\right), \;\; \bar{v} = \frac{1}{2}\left(\sqrt{z^2 + 4e^{-2-s-t}} + z\right)$$

and to see if the subgradient

$$\psi'(u,v) = \begin{pmatrix} s + \ln u + 1 \\ t + \ln v + 1 \end{pmatrix}.$$

satisfies

$$\psi'_u(\bar{u},\bar{v}) + \psi'_v(\bar{u},\bar{v}) = 0 \;\;\text{ and }\;\; \psi'_u(\bar{u},\bar{v}) - \psi'_v(\bar{u},\bar{v}) > 0.$$

If this is the case, we take $\bar{u},\bar{v}$ as the minimizers, if not, the inequality constraint is not active at the optimal solution of (54) and we take

$$\bar{u} = e^{-1-s}, \;\; \bar{v} = e^{-1-t}.$$

6.2 Hyperoctahedron setup

Let now Q be a standard hyperoctahedron: $Q = \{x \in \mathbb{R}^n | \, \|x\|_1 \leq 1\}$. Let us see how the solution to (11) can be computed in this case.

When writing

$$x_i = w_i - y_i, \;\; w_i, \; y_i \geq 0, \;\; \sum_{i=1}^n [w_i + y_i] = 1,$$

the problem (11) on $Q_R(z)$ can be rewritten as

$$\min_{w,y,u,v} \{ \textstyle\sum_{i=1}^n [s_i(w_i - y_i) + u_i \ln u_i + v_i \ln v_i] : \;\; \sum_{i=1}^n [u_i + v_i] = R, \;\; \sum_{i=1}^n [w_i + y_i] = 1,$$
$$w_i - y_i = z_i + u_i - v_i, \;\; u_i \geq 0, \;\; v_i \geq 0, \;\; w_i \geq 0, \;\; y_i \geq 0, \;\; i = 1,...,n.\}$$

When dualizing the coupling constraints we come to

$$\max_{\lambda,\mu}\Big\{\underline{L}(\lambda,\mu)\equiv\min_{u,v,w,y}L(u,v,w,y,\lambda,\mu)\ :$$
$$z_i+u_i-v_i-w_i+y_i=0,\quad w_i\geq 0,\quad y_i\geq 0,\quad i=1,...,n\}$$

where

$$L(u,v,w,y,\lambda,\mu)=\sum_{i=1}^n[r_iv_i+t_iu_i+\mu(w_i+y_i)+u_i\ln u_i+v_i\ln v_i]-\lambda R-\mu\ :$$
$$r_i=s_i+\lambda,\quad t_i=-s_i+\lambda.$$

The computation of the dual function $\underline{L}(\lambda,\mu)$ boils down to evaluating solutions to n subproblems

$$\begin{array}{l}\min_{u,v}\ su+tv+\lambda(w+y)+u\ln u+v\ln v,\\ z+u-v-w+y=0,\quad w\geq 0,\quad y\geq 0.\end{array}\tag{55}$$

It is obvious that either w or y vanishes, and to find the solution to (55) it suffices to compare the optimal values of the problems

$$\min_{u,v}\ \psi_w(u,v)=su+tv+\lambda(z+u-v)+u\ln u+v\ln v,\quad z+u-v\geq 0,\quad (\text{case } y=0),$$
$$\min_{u,v}\ \psi_y(u,v)=su+tv-\lambda(z+u-v)+u\ln u+v\ln v,\quad z+u-v\leq 0,\quad (\text{case } w=0),$$

which are the same problems as (54) in the previous section.

References

[1] D. Azé, J.-P. Penot, Uniformly convex and uniformly smooth convex functions. *Ann. Fac. Sci. Toulouse*, VI. Sér., Math. 4, 705-730 (1995).

[2] Yu. Chekanov, Yu. Nesterov, A. Vladimirov. On uniformly convex functionals, *Vest. Mosk. Univ.*, **3**, Ser. XV, 12-23 (1978).

[3] I.A. Ibragimov, Yu.V. Linnik. *Independent and stationary sequences of random variables*, Wolters-Noordhoff Ser. Pure and Appl. Math. (1971).

[4] V. Lemaire, G. Pags Unconstrained recursive importance sampling, *Ann. Appl. Probab.* **20** 3, 1029-1067 (2010).

[5] B. Nadler, N. Srebro, X. Zhou Statistical Analysis of Semi-Supervised Learning: The Limit of Infinite Unlabelled Data, *NIPS 2009 Online papers*, http://books.nips.cc/nips22.html, to appear in *Advances in Neural Information Processing Systems* **22** edited by Y. Bengio et al., (2009).

[6] A.S. Nemirovski, D.B. Yudin, *Problem complexity and method efficiency in optimization*, Wiley-Interscience Series in Discrete Mathematics, John Wiley, **XV**, (1983).

[7] A. Nemirovski , A. Juditsky , G. Lan , A. Shapiro, Robust Stochastic Approximation Approach to Stochastic Programming, *SIAM J. Optim.* **19**, 4, 1574-1609 (2009).

[8] Yu. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Springer (2003).

[9] Yu. Nesterov. Smooth minimization of nonsmooth functions, *Math. Prog. Ser A*, **103**, 1, 127-152 (2005).

[10] Yu. Nesterov. Excessive Gap Technique in Nonsmooth Convex Minimization, *SIAM J. Optim.* **16**, 1, 235 - 249, (2005)

[11] Yu. Nesterov. Primal-dual subgradient methods for convex problems, *Math. Program., Ser. B* (2007) (Online).

[12] Yu. Nesterov. Barrier subgradient method. *Ciaco*, 2008, CORE DP2008/60, (2008).

[13] Yu. Nesterov, J. -Ph. Vial. Confidence level solutions for stochastic programming *Automatica* **44**, 6, 1559-1568 (2008).

[14] B. Polyak. Existence theorems and convergence of minimizing sequences in extremum problems with restrictions, *Sov. Math. Dokl.*, **7**, 72-75, (1967).

[15] L. Xiao, Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization, ISMP 2009, Chicago, August 23-28 (2009).

[16] C. Zalinescu. On uniformly convex functions, *J. Math. Anal. Appl*, **95**, 344-374 (1983).