Primal-dual subgradient methods for convex problems

Yu. Nesterov *

March 2002,
September 2005
(after revision)

Abstract

In this paper we present a new approach for constructing subgradient schemes for different types of nonsmooth problems with convex structure. Our methods are primal-dual since they are always able to generate a feasible approximation to the optimum of an appropriately formulated dual problem. Besides other advantages, this useful feature provides the methods with a reliable stopping criterion. The proposed schemes differ from the classical approaches (divergent series methods, mirror descent methods) by presence of two control sequences. The first sequence is responsible for aggregating the support functions in the dual space, and the second one establishes a dynamically updated scale between the primal and dual spaces. This additional flexibility allows to guarantee a boundedness of the sequence of primal test points even in the case of unbounded feasible set. We present the variants of subgradient schemes for nonsmooth convex minimization, minimax problems, saddle point problems, variational inequalities, and stochastic optimization. In all situations our methods are proved to be optimal from the view point of worst-case black-box lower complexity bounds.

Keywords: convex optimization, subgradient methods, non-smooth optimization, minimax problems, saddle points, variational inequalities, stochastic optimization, black-box methods, lower complexity bounds.

*Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium; e-mail: nesterov@core.ucl.ac.be.

This paper presents research results of the Belgian Program on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office.
The scientific responsibility rests with its author.

1 Introduction

Prehistory. The results presented in this paper are not very new. Most of them were obtained by the author in 2001 – 2002. However, a further purification of the developed framework led to rather surprising results related to the smoothing technique. Namely, in [10] it was shown that many nonsmooth convex minimization problems with an appropriate explicit structure can be solved up to absolute accuracy ϵ in $O(\frac{1}{\epsilon})$ iterations of special gradient-type methods. Recall that the exact *lower* complexity bound for *any* black-box subgradient scheme was established on the level of $O(\frac{1}{\epsilon^2})$ iterations (see [8], or [9], Section 3.1, for a more recent exposition). Thus, in [10] it was shown that the gradient-type methods of *Structural Optimization* outperform the black-box ones by an order of magnitude.

At that moment of time, the author got an illusion that the importance of black-box approach in Convex Optimization will be irreversibly vanishing, and, finally, this approach will be completely replaced by other ones based on a clever use of problem's structure (interior-point methods, smoothing, etc.). This explains why the results included in this paper were not published at time. 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 too complex for constructing a good self-concordant barrier or for applying a smoothing technique. Note also, that optimization schemes sometimes are employed for modelling certain *adjustment processes* in real-life systems. In this situation, we are not free in selecting the type of optimization scheme and in the choice of its parameters. However, the results on convergence and the rate of convergence of corresponding methods remain interesting.

These considerations encouraged the author to publish the above mentioned results on primal-dual subgradient methods for nonsmooth convex problems. Note that some elements of developed technique were used by the author later on in different papers related to smoothing approach (see [10, 11, 12]). Nevertheless, for a reader convenience, we include in this paper all necessary proofs.

Motivation. Historically, a subgradient method with *constant* step was the first numerical scheme suggested for approaching a solution to optimization problem

$$\min_x \{f(x) : \; x \in R^n\} \tag{1.1}$$

with nonsmooth convex objective function f (see [15] for historical remarks). In a convergent variant of this method [6, 14] we need to choose in advance a sequence of steps $\{\lambda_k\}_{k=0}^{\infty}$ satisfying the *divergent-series rule*:

$$\lambda_k > 0, \quad \lambda_k \to 0, \quad \sum_{k=0}^{\infty} \lambda_k = \infty. \tag{1.2}$$

Then, for $k \geq 0$ we can iterate:

$$x_{k+1} \;\; = \;\; x_k - \lambda_k g_k, \quad k \geq 0, \tag{1.3}$$

with some $g_k \in \partial f(x_k)$. In another variant of this scheme, the step direction is normalized:

$$x_{k+1} \;\; = \;\; x_k - \lambda_k g_k / \|g_k\|_2, \quad k \geq 0, \tag{1.4}$$

where $\|\cdot\|_2$ denotes the standard Euclidean norm in R^n, introduced by the standard inner product:

$$\langle x, y \rangle \;=\; \sum_{i=1}^{n} x^{(i)} y^{(i)}, \quad x, y \in R^n.$$

Since the objective function f is nonsmooth, we cannot expect its subgradients be vanishing in a neighborhood of optimal solution to (1.1). Hence, the condition $h_k \to 0$ is necessary for convergence of the processes (1.3) or (1.4). On the other hand, assuming that $\|g_k\|_2 \le L$, for any $x \in R^n$ we get

$$\begin{aligned} \|x - x_{k+1}\|_2^2 &\overset{(1.3)}{=} \|x - x_k\|_2^2 + 2\lambda_k \langle g_k, x - x_k \rangle + \lambda_k^2 \|g_k\|_2^2 \\ &\le \|x - x_k\|_2^2 + 2\lambda_k \langle g_k, x - x_k \rangle + \lambda_k^2 L^2. \end{aligned}$$

Hence, for any $x \in R^n$ with $\frac{1}{2}\|x - x_0\|_2^2 \le D$

$$\begin{aligned} f(x) \;&\ge\; l_N(x) \overset{\text{def}}{=} \sum_{k=0}^{N} \lambda_k [f(x_k) + \langle g_k, x - x_k \rangle] \Big/ \sum_{k=0}^{N} \lambda_k \\ &\ge\; \left\{ \sum_{i=0}^{N} \lambda_k f(x_k) - D - \tfrac{1}{2} L^2 \sum_{k=0}^{N} \lambda_k^2 \right\} \Big/ \sum_{k=0}^{N} \lambda_k. \end{aligned} \tag{1.5}$$

Thus, denoting $f_D^* \;=\; \min\limits_x \{ f(x) : \; \frac{1}{2}\|x - x_0\|_2^2 \le D \}$,

$$\bar{f}_N \;=\; \frac{\sum\limits_{k=0}^{N} \lambda_k f(x_k)}{\sum\limits_{k=0}^{N} \lambda_k}, \quad \kappa_N \;=\; \frac{2D + L^2 \sum\limits_{k=0}^{N} \lambda_k^2}{2 \sum\limits_{k=0}^{N} \lambda_k},$$

we conclude that

$$\bar{f}_N \;-\; f_D^* \;\le\; \kappa_N.$$

Note that the conditions (1.2) are necessary and sufficient for $\kappa_N \to 0$.

In the above analysis, convergence of the process (1.3) is based on the fact that the derivative of $l_N(x)$, the *lower linear model* of the objective function, is vanishing. This model is very important since it provides us also with a reliable stopping criterion. However, examining the structure of linear function $l_N(\cdot)$, we can observe a very strange feature:

New subgradients enter the model with decreasing weights. (1.6)

This feature contradicts common sense. It contradicts also to general principles of iterative schemes, in accordance to which the new information is more important than the old one. Thus, something is wrong. Unfortunately, in our situation a simple treatment is hardly possible: we have seen that decreasing weights ($\equiv$ steps) are *necessary* for convergence of the primal sequence $\{x_k\}_{k=0}^{\infty}$.

The above contradiction served as a point of departure for the developments presented in this paper. The proposed alternative looks quite natural. Indeed, we have seen that in primal space it is necessary to have a vanishing sequence of steps, but in the dual space (the space of linear functions), we would like to apply non-decreasing weights.

Consequently, we need *two different* sequences of parameters, each of which is responsible for some related processes in primal and dual spaces. The idea to relate the primal minimization sequence with a master process existing in the dual space is not new. It was implemented first in the *mirror descent* methods (see [8], and [5] with [4] for newer versions and historical remarks). However, the divergent series somehow penetrated in this approach too. Therefore, in Euclidean situation the mirror descent method coincides with subgradient one and, consequently, shares the drawback (1.6).

In this paper we consider the *primal-dual* subgradient schemes. It seems that this intrinsic feature of *all* subgradient methods was not recognized yet explicitly. Consider, for example, the scheme (1.3). From inequality (1.5), it is clear that

$$f_D^* \;\geq\; \hat{f}_N(D) \;\stackrel{\rm def}{=}\; \min_x \{l_N(x) : \tfrac{1}{2}\|x - x_0\|_2^2 \leq D\}.$$

Note that the value $\hat{f}_N(D)$ can be easily computed. On the other hand, in Convex Optimization there is only one way to get a lower bound for the *optimal* solution of a minimization problem. For that, we need to find a *feasible* solution to a certain dual problem.[1] Thus, computability of $\hat{f}_N(D)$ implies that we are able to point out a dual solution. And convergence of the primal-dual gap $\bar{f}_N(D) - \hat{f}_N$ to zero implies that the dual solution approaches the optimal one. Below, we will discuss in details the meaning of the dual solutions generated by the proposed schemes.

Finally, note that the subgradient schemes proposed in this paper are different from the standard "search methods". For example, for problem (1.1) we suggest to use the following process:

$$x_{k+1} = \arg\min_{x \in R^n} \left\{ \tfrac{1}{k+1} \sum_{i=0}^{k} [f(x_i) + \langle g_i, x - x_i \rangle] + \mu_k \|x - x_k\|_2^2 \right\}, \qquad (1.7)$$

where $\mu_k = O(\frac{1}{\sqrt{k}}) \to 0$. Note that in this scheme no artificial elements are needed to ensure the boundedness of the sequence of test points.

Contents. The paper is organized as follows. In Section 2 we consider a general scheme of Dual Averaging (DA) and prove the upper bounds for corresponding gap functions. These bounds will be specified later on for the particular problem classes. We give also two main variants of DA-methods, the method of Simple Dual Averages (SDA) and the method of Weighted Dual Averages.

In Section 3 we apply DA-methods to minimization over simple sets. In Section 3.1 we consider different forms of DA-methods as applied to a general minimization problem. The main goal of Section 3 is to show that all DA-methods are primal-dual. For that, we always consider a kind of dual problem and point out the sequence which converges to its solution. In Section 3.2 this is done for general unconstrained minimization problem. In Section 3.3 we do that for minimax problems. It is interesting that approximations of dual multipliers in this case are proportional to the number of times the corresponding functional components were active during the minimization process (compare with [2]). Finally, in Section 3.4 we consider primal-dual schemes for solving minimization

[1] Depending on available information on the *structure* of the problem, this dual problem can be posed in different forms. Therefore, sometimes we prefer to call such problems the *adjoint* ones.

problems with simple functional constraints. We manage to obtain for such problems an approximate primal-dual solution despite the fact that usually even the complexity of computation of the value of dual objective function is comparable with complexity of the initial problem.

In the next Section 4 we apply SDA-method to saddle point problems, and in Section 5 it is applied to variational inequalities. It is important that for solvable problems SDA-method generates a bounded sequence even for unbounded feasible set. In Section 6 we consider a stochastic version of SDA-method. We prove that this method generates a random sequence of points with expected value of the objective function convergent to the optimal value of stochastic optimization problem.

Finally, in Section 7 we consider two applications of DA-methods to modelling. In Section 7.1 it is shown that a natural adjustment strategy of a balanced development can be seen as an implementation of DA-method. Hence, it is possible to prove that in the limit this process converges to a solution of some optimization problem. In Section 7.2, for a multi-period optimization problem we compare the efficiency of static strategies based on preliminary planning and a certain dynamic adjustment process (based on SDA-method). It is clear that the dynamic adjustment computationally is much cheaper. On the other hand, we show that its results are comparable with the results of preliminary planning based on complete knowledge of the future.

We conclude the paper with Section 8, where we discuss the results. For completeness of presentation, we put in Appendix some nonstandard results on strongly convex functions.

Notations and generalities. Let E be a finite-dimensional real vector space and E^* be its dual. 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 system of primal balls:

$$B_r(x) = \{y \in E : \ \|y - x\| \leq r\}.$$

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

$$\|s\|_* = \max_x \{\langle s, x \rangle : \ x \in B_1(0)\}, \quad s \in E^*.$$

Let Q be a closed convex set in E. Assume that we know a *prox-function* $d(x)$ of the set Q. This means that $d(x)$ is a continuous function with domain belonging to Q, which is strongly convex on Q with respect to $\|\cdot\|$: $\quad \forall x, y \in Q, \quad \forall \alpha \in [0, 1]$,

$$d(\alpha x + (1-\alpha)y) \ \leq \ \alpha d(x) + (1-\alpha) d(y) - \tfrac{1}{2}\sigma\alpha(1-\alpha)\|x - y\|^2, \tag{1.8}$$

where $\sigma \geq 0$ is the *convexity parameter*.[2] Denote by x_0 the *prox-center* of the set Q:

$$x_0 = \arg\min_x \{d(x) : \ x \in Q\}. \tag{1.9}$$

Without loss of generality, we assume that $d(x_0) = 0$. In view of Lemma 6 in Appendix, the prox-center is well-defined and

$$d(x) \geq \tfrac{1}{2}\sigma\|x - x_0\|^2, \quad x \in Q.$$

[2]Note that in this definition we do not assume $d(x)$ to be differentiable on Q. Since this way of defining a strongly convex function is not standard, we present a justification for some properties of such functions in the Appendix.

Let us illustrate the above terminology by two examples.

1. Standard Euclidean norm. In this case

$$\|x\| \;=\; \|x\|_2 \;\equiv\; \langle x, x\rangle^{1/2}.$$

The natural choice of the distance function is then $d(x) = \frac{1}{2}\|x - x_0\|_2^2$ with some $x_0 \in Q$. Hence, $\sigma = 1$.

2. l_1-norm. Let us choose $E = R^n$ and

$$\|x\| \;=\; \|x\|_1 \;\equiv\; \sum_{i=1}^{n} |x^{(i)}|. \tag{1.10}$$

Define $Q = \Delta_n \stackrel{\rm def}{=} \{x \in R^n_+ :\; \sum\limits_{i=1}^{n} x^{(i)} = 1\}$ and consider the *entropy function*:

$$d(x) \;=\; \ln n + \sum_{i=1}^{n} x^{(i)} \ln x^{(i)}.$$

Then for any $x \in Q$ we have $d(x) \geq d(x_0) = 0$ where $x_0 = (\frac{1}{n}, \dots \frac{1}{n})^T \in Q$. Note that for any $x > 0$ function $d(x)$ is two times continuously differentiable and

$$\langle \nabla^2 d(x)h, h\rangle \;=\; \sum_{i=1}^{n} \frac{(h^{(i)})^2}{x^{(i)}}, \quad h \in R^n.$$

Hence, by Cauchy-Schwartz inequality,

$$\left(\sum_{i=1}^{n} |h^{(i)}|\right)^2 \;\leq\; \left(\sum_{i=1}^{n} |x^{(i)}|\right) \cdot \langle \nabla^2 d(x)h, h\rangle. \tag{1.11}$$

Consequently, in view of Lemma 7 in Appendix, $d(x)$ is strongly convex on Q with parameter $\sigma = 1$.

2 Main algorithmic schemes

Let Q be a closed convex set in E endowed with a prox-function $d(x)$. We allow Q to be unbounded (for example, $Q \equiv E$). For our analysis we need to define two support-type functions of the set Q:

$$\begin{aligned} \xi_D(s) &= \max_{x \in Q}\{\langle s, x - x_0\rangle :\; d(x) \leq D\}, \\ V_\beta(s) &= \max_{x \in Q}\{\langle s, x - x_0\rangle - \beta d(x)\}, \end{aligned} \tag{2.1}$$

where $D \geq 0$ and $\beta > 0$ are some parameters. The first function is a usual support function for the set

$$\mathcal{F}_D = \{x \in Q :\; d(x) \leq D\}.$$

The second one is a proximal-type approximation of the support function of set Q. Since $d(\cdot)$ is strongly convex, for any positive D and β we have $\mathrm{dom}\, \xi_D = \mathrm{dom}\, V_\beta = E^*$. Note that both of the functions are nonnegative.

Let us mention some properties of function $V(\cdot)$. If $\beta_2 \geq \beta_1 > 0$, then for any $s \in E^*$ we have

$$V_{\beta_2}(s) \leq V_{\beta_1}(s). \tag{2.2}$$

Note that the level of smoothness of function $V_\beta(\cdot)$ is controlled by parameter β.

Lemma 1 *Function $V_\beta(\cdot)$ is convex and differentiable on E^*. Moreover, its gradient is Lipschitz continuous with constant $\frac{1}{\beta\sigma}$:*

$$\|\nabla V_\beta(s_1) - \nabla V_\beta(s_2)\| \leq \tfrac{1}{\beta\sigma}\|s_1 - s_2\|_*, \quad \forall s_1, s_2 \in E^*. \tag{2.3}$$

For any $s \in E_*$, vector $\nabla V_\beta(s)$ belongs to Q:

$$\nabla V_\beta(s) \;=\; \pi_\beta(s) - x_0, \quad \pi_\beta(s) \stackrel{\rm def}{=} \arg\min_{x \in Q}\{-\langle s, x\rangle + \beta d(x)\}. \tag{2.4}$$

Proof:
In view of definition (2.1), function $V_\beta(s)$ is a maximum of linear functions. Therefore it is convex. Since $d(x)$ is strongly convex, the point $\pi_\beta(s) \in Q$ is uniquely defined (see Lemma 6 in Appendix).

Finally, let us fix two points $s_1, s_2 \in E^*$ and arbitrary $\alpha \in (0, 1)$. Denote

$$x_1 = \pi_\beta(s_1), \quad x_2 = \pi_\beta(s_2), \quad x(\alpha) = \alpha x_1 + (1-\alpha)x_2.$$

Since for any fixed $s \in E^*$, function $-\langle s, x\rangle + \beta d(x)$ is strongly convex on Q with convexity parameter $\beta\sigma$, using inequality (9.5) in Appendix we get

$$\begin{aligned}
\langle s_1, x(\alpha)\rangle - \beta d(x(\alpha)) &\leq \langle s_1, x_1\rangle - \beta d(x_1) - \tfrac{1}{2}\beta\sigma\|x(\alpha) - x_1\|^2, \\
\langle s_2, x(\alpha)\rangle - \beta d(x(\alpha)) &\leq \langle s_2, x_2\rangle - \beta d(x_2) - \tfrac{1}{2}\beta\sigma\|x(\alpha) - x_2\|^2.
\end{aligned}$$

Summing up these inequalities with coefficients α and $1-\alpha$ respectively and using (1.8), we get

$$\begin{aligned}
&\alpha\langle s_1, x(\alpha) - x_1\rangle + (1-\alpha)\langle s_2, x(\alpha) - x_2\rangle \\
&\leq \beta[d(x(\alpha)) - \alpha d(x_1) - (1-\alpha)d(x_2)] - \tfrac{1}{2}\beta\sigma\alpha(1-\alpha)\|x_1 - x_2\|^2 \\
&\leq -\beta\sigma\alpha(1-\alpha)\|x_1 - x_2\|^2.
\end{aligned}$$

Thus, $\beta\sigma\|x_1 - x_2\|^2 \leq \langle s_1 - s_2, x_1 - x_2\rangle \leq \|s_1 - s_2\|_* \cdot \|x_1 - x_2\|$ and this is (2.3). □

As a trivial corollary of (2.3) we get the following inequality:

$$V_\beta(s + \delta) \leq V_\beta(s) + \langle \delta, \nabla V_\beta(s)\rangle + \tfrac{1}{2\sigma\beta}\|\delta\|_*^2 \quad \forall s, \delta \in E^*. \tag{2.5}$$

Note that in view of definition (1.9) we have $\pi_\beta(0) = x_0$. This implies $V_\beta(0) = 0$ and $\nabla V_\beta(0) = 0$. Thus, in this case inequality (2.5) with $s = 0$ yields

$$V_\beta(\delta) \leq \tfrac{1}{2\sigma\beta}\|\delta\|_*^2 \quad \forall \delta \in E^*. \tag{2.6}$$

In the sequel, we assume that the set Q is simple enough for computing vector $\pi_\beta(s)$ exactly. For our analysis we need the following relation between the functions (2.1).

Lemma 2 *For any $s \in E^*$ and $\beta \geq 0$ we have*

$$\xi_D(s) \;\leq\; \beta D + V_\beta(s). \tag{2.7}$$

Proof:
Indeed,

$$\begin{aligned}
\xi_D(s) &= \max_{x \in Q} \{ \langle s, x - x_0 \rangle : \; d(x) \leq D \} \\
&= \max_{x \in Q} \min_{\beta \geq 0} \; \{ \; \langle s, x - x_0 \rangle + \beta \, [D - d(x)] \; \} \\
&\leq \min_{\beta \geq 0} \max_{x \in Q} \; \{ \; \langle s, x - x_0 \rangle + \beta \, [D - d(x)] \; \} \\
&\leq \beta D + V_\beta(s).
\end{aligned}$$

□

Consider now the sequences

$$X_k = \{x_i\}_{i=0}^k \subset Q, \quad G_k = \{g_i\}_{i=0}^k \subset E^*, \quad \Lambda_k = \{\lambda_i\}_{i=0}^k \subset R_+.$$

Typically, the test points x_i and the weights λ_i are generated by some algorithmic scheme and the points g_i are computed by a black-box oracle $\mathcal{G}(\cdot)$, related to a specific convex problem:

$$g_i = \mathcal{G}(x_i), \quad i \geq 0.$$

In this paper we consider only the problem instances, for which there exists a solution $x^* \in Q$ satisfying the condition

$$\langle g_i, x_i - x^* \rangle \geq 0, \quad i \geq 0. \tag{2.8}$$

We are going to approximate the primal and dual solutions of our problem using the following aggregate objects:

$$\begin{aligned}
S_k &= \sum_{i=0}^k \lambda_i, & \hat{x}_{k+1} &= \tfrac{1}{S_k} \sum_{i=0}^k \lambda_i x_i, \\
s_{k+1} &= \sum_{i=0}^k \lambda_i g_i, & \hat{s}_{k+1} &= \tfrac{1}{S_k} s_{k+1},
\end{aligned} \tag{2.9}$$

with $\hat{x}_0 = x_0$ and $s_0 = 0$.

As we will see later, the quality of the test sequence X_k can be naturally described by the following *gap function*:

$$\delta_k(D) = \max_x \left\{ \sum_{i=0}^k \lambda_i \langle g_i, x_i - x \rangle : \; x \in \mathcal{F}_D, \; \right\}, \quad D \geq 0. \tag{2.10}$$

Using notation (2.9), we get an explicit representation of the gap:

$$\delta_k(D) = \sum_{i=0}^k \lambda_i \langle g_i, x_i - x_0 \rangle + \xi_D(-s_{k+1}). \tag{2.11}$$

Sometimes we will use an *upper gap function*

$$\begin{aligned} \Delta_k(\beta, D) &= \beta D + \sum_{i=0}^{k} \lambda_i \langle g_i, x_i - x_0 \rangle + V_\beta(-s_{k+1}) \\ &= \sum_{i=0}^{k} \lambda_i \langle g_i, x_i - \pi_\beta(-s_{k+1}) \rangle + \beta \cdot (D - d(\pi_\beta(-s_{k+1}))) \,. \end{aligned} \tag{2.12}$$

In view of (2.7) and (2.11), for any non-negative D and β we have

$$\delta_k(D) \le \Delta_k(\beta, D). \tag{2.13}$$

Since Q is a simple set, the values of the gap functions can be easily computed. Note that for some D these values can be negative. However, if the solution x^* of our problem exists (in the sense of (2.8)), then for

$$D \ge d(x^*) \quad (\Rightarrow \; x^* \in \mathcal{F}_D) \,,$$

the value $\delta_k(D)$ is non-negative independently on the sequences X_k, Λ_k and G_k, involved in its definition.

Consider now the generic scheme of *Dual Averaging* (DA-scheme).

Initialization: Set $s_0 = 0 \in E^*$. Choose $\beta_0 > 0$.
Iteration ($k \ge 0$)**:**   **1.** Compute $g_k = \mathcal{G}(x_k)$.   **2.** Choose $\lambda_k > 0$. Set $s_{k+1} = s_k + \lambda_k g_k$.   **3.** Choose $\beta_{k+1} \ge \beta_k$. Set $x_{k+1} = \pi_{\beta_{k+1}}(-s_{k+1})$.

(2.14)

Theorem 1 *Let the sequences X_k, G_k and Λ_k be generated by (2.14). Then:*

1. *For any $k \ge 0$ and $D \ge 0$ we have:*

$$\delta_k(D) \;\le\; \Delta_k(\beta_{k+1}, D) \;\le\; \beta_{k+1} D + \tfrac{1}{2\sigma} \sum_{i=0}^{k} \tfrac{\lambda_i^2}{\beta_i} \|g_i\|_*^2. \tag{2.15}$$

2. *Assume that the solution x^* in the sense (2.8) exists. Then*

$$\tfrac{1}{2}\sigma \|x_{k+1} - x^*\|^2 \le d(x^*) + \tfrac{1}{2\sigma\beta_{k+1}} \sum_{i=0}^{k} \tfrac{\lambda_i^2}{\beta_i} \|g_i\|_*^2. \tag{2.16}$$

3. *Assume that x^* is an interior solution: $B_r(x^*) \subseteq \mathcal{F}_D$ for some positive r and D. Then*

$$\|\hat{s}_{k+1}\|_* \le \tfrac{1}{rS_k} \left[\beta_{k+1} D + \tfrac{1}{2\sigma} \sum_{i=0}^{k} \tfrac{\lambda_i^2}{\beta_i} \|g_i\|_*^2 \right]. \tag{2.17}$$

Proof:
1. In view of (2.13), we need to prove only the second inequality in (2.15). By the rules of scheme (2.14), for $i \geq 1$ we obtain

$$V_{\beta_{i+1}}(-s_{i+1}) \overset{(2.2)}{\leq} V_{\beta_i}(-s_{i+1}) \overset{(2.5)}{\leq} V_{\beta_i}(-s_i) - \lambda_i \langle g_i, \nabla V_{\beta_i}(-s_i) \rangle + \tfrac{\lambda_i^2}{2\sigma\beta_i} \|g_i\|_*^2$$

$$\overset{(2.4)}{=} V_{\beta_i}(-s_i) + \lambda_i \langle g_i, x_0 - x_i \rangle + \tfrac{\lambda_i^2}{2\sigma\beta_i} \|g_i\|_*^2.$$

Thus,

$$\lambda_i \langle g_i, x_i - x_0 \rangle \leq V_{\beta_i}(-s_i) - V_{\beta_{i+1}}(-s_{i+1}) + \tfrac{\lambda_i^2}{2\sigma\beta_i} \|g_i\|_*^2, \quad i = 1, \ldots, k.$$

The summation of all these inequalities results in

$$\sum_{i=0}^{k} \lambda_i \langle g_i, x_i - x_0 \rangle \leq V_{\beta_1}(-s_1) - V_{\beta_{k+1}}(-s_{k+1}) + \tfrac{1}{2\sigma} \sum_{i=1}^{k} \tfrac{\lambda_i^2}{\beta_i} \|g_i\|_*^2. \tag{2.18}$$

But in view of (2.6) $V_{\beta_1}(-s_1) \leq \frac{\lambda_0^2}{2\sigma\beta_1} \|g_0\|_*^2 \leq \frac{\lambda_0^2}{2\sigma\beta_0} \|g_0\|_*^2$. Thus, (2.18) results in (2.15).

2. Let us assume that x^* exists. Then,

$$\begin{aligned} 0 &\overset{(2.8)}{\leq} \sum_{i=0}^{k} \lambda_i \langle g_i, x_i - x^* \rangle \\ &\overset{(2.18)}{\leq} \langle s_{k+1}, x_0 - x^* \rangle - V_{\beta_{k+1}}(-s_{k+1}) + \tfrac{1}{2\sigma} \sum_{i=0}^{k} \tfrac{\lambda_i^2}{\beta_i} \|g_i\|_*^2 \\ &\overset{(2.1)}{=} \langle s_{k+1}, x_{k+1} - x^* \rangle + \beta_{k+1} d(x_{k+1}) + \tfrac{1}{2\sigma} \sum_{i=0}^{k} \tfrac{\lambda_i^2}{\beta_i} \|g_i\|_*^2 \\ &\overset{(9.5)}{\leq} \beta_{k+1} d(x^*) - \tfrac{1}{2}\beta_{k+1}\sigma \|x_{k+1} - x^*\|^2 + \tfrac{1}{2\sigma} \sum_{i=0}^{k} \tfrac{\lambda_i^2}{\beta_i} \|g_i\|_*^2, \end{aligned}$$

and that is (2.16).

3. Let us assume now that x^* is an interior solution. Then

$$\begin{aligned} \delta_k(D) &= \max_x \left\{ \sum_{i=0}^{k} \lambda_i \langle g_i, x_i - x \rangle : \; x \in \mathcal{F}_D \right\} \\ &\geq \max_x \left\{ \sum_{i=0}^{k} \lambda_i \langle g_i, x_i - x \rangle : \; x \in B_r(x^*) \right\} \geq r \|s_{k+1}\|_*. \end{aligned}$$

Thus, (2.17) follows from (2.15). □

The form of inequalities (2.15) and (2.16) suggests some natural strategies for choosing the parameters β_i in the scheme (2.14). Let us define the following sequence:

$$\hat{\beta}_0 = \hat{\beta}_1 = 1, \quad \hat{\beta}_{i+1} = \hat{\beta}_i + \tfrac{1}{\hat{\beta}_i}, \; i \geq 1. \tag{2.19}$$

The advantage of this sequence is justified by the following relation:

$$\hat{\beta}_{k+1} = \sum_{i=0}^{k} \tfrac{1}{\hat{\beta}_i}, \quad k \geq 0.$$

Thus, this sequence can be used for balancing the terms appearing in the right-hand side of inequality (2.15). Note that the growth of the sequence can be estimated as follows.

Lemma 3

$$\sqrt{2k-1} \leq \hat{\beta}_k \leq \frac{1}{1+\sqrt{3}} + \sqrt{2k-1}, \quad k \geq 1. \tag{2.20}$$

Proof:
From (2.19) we have $\hat{\beta}_1 = 1$ and $\hat{\beta}_{k+1}^2 = \hat{\beta}_k^2 + \hat{\beta}_k^{-2} + 2$ for $k \geq 1$. This gives the first inequality in (2.20). On the other hand, since the function $\beta + \frac{1}{\beta}$ is increasing for $\beta \geq 1$ and $\hat{\beta}_k \geq 1$, the second inequality in (2.20) can be justified by induction. □

We will consider two main strategies for choosing λ_i in (2.14):

- *Simple averages:* $\lambda_k = 1$.
- *Weighted averages:* $\lambda_k = \frac{1}{\|g_k\|_*}$.

For the convenience of referencing, let us write down the corresponding algorithmic schemes in an explicit form. Both theorems below are straightforward consequences of Theorem 1.

Method of Simple Dual Averages

Initialization: Set $s_0 = 0 \in E^*$. Choose $\gamma > 0$.

Iteration ($k \geq 0$):

1. Compute $g_k = \mathcal{G}(x_k)$. Set $s_{k+1} = s_k + g_k$.

2. Choose $\beta_{k+1} = \gamma \hat{\beta}_{k+1}$. Set $x_{k+1} = \pi_{\beta_{k+1}}(-s_{k+1})$. (2.21)

Theorem 2 *Denote* $L_k = \max\limits_{0 \leq i \leq k} \|g_i\|_*$. *For method (2.21) we have* $S_k = k+1$ *and*

$$\delta_k(D) \leq \hat{\beta}_{k+1} \left(\gamma D + \frac{L_k^2}{2\sigma\gamma} \right).$$

Moreover, if the solution x^* *in the sense (2.8) exists, then the scheme (2.21) generates a bounded sequence:*

$$\|x_k - x^*\|^2 \leq \frac{2}{\sigma} d(x^*) + \frac{L_k^2}{\sigma^2 \gamma^2}, \quad k \geq 0.$$

Method of Weighted Dual Averages

Initialization: Set $s_0 = 0 \in E^*$. Choose $\rho > 0$.

Iteration ($k \geq 0$):

1. Compute $g_k = \mathcal{G}(x_k)$. Set $s_{k+1} = s_k + g_k / \|g_k\|_*$.

2. Choose $\beta_{k+1} = \frac{\hat{\beta}_{k+1}}{\rho\sqrt{\sigma}}$. Set $x_{k+1} = \pi_{\beta_{k+1}}(-s_{k+1})$.

(2.22)

Theorem 3 *For method (2.22) we have $S_k \geq \frac{k+1}{L_k}$ and*

$$\delta_k(D) \leq \frac{\hat{\beta}_{k+1}}{\sqrt{\sigma}} \left(\frac{D}{\rho} + \frac{1}{2}\rho \right).$$

Moreover, if the solution x^* in the sense (2.8) exists, then the scheme (2.22) generates a bounded sequence:

$$\|x_k - x^*\|^2 \leq \frac{1}{\sigma}(2d(x^*) + \rho^2).$$

In the next sections we show how to apply the above results to different classes of problems with convex structure.

3 Minimization over simple sets

3.1 General minimization problem

Consider the following minimization problem:

$$\min_x \{ f(x) : \; x \in Q \}, \tag{3.1}$$

where f is a convex function defined on E and Q is a closed convex set. Recall that we assume Q to be simple, which means computability of exact optimal solutions to both minimization problems in (2.1).

In order to solve problem (3.1), we need a black-box oracle, which is able to compute a subgradient of objective function at any test point:

$$\mathcal{G}(x) \in \partial f(x), \quad x \in E.$$

Then, we have the following interpretation of the gap function $\delta_k(D)$. Denote

$$\begin{aligned} l_k(x) &= \tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i [f(x_i) + \langle g_i, x - x_i \rangle], \quad (g_i \in \partial f(x_i)) \\ \hat{f}_k(D) &= \min_x \{ l_k(x) : \; x \in \mathcal{F}_D \}, \\ f_D^* &= \min_x \{ f(x) : \; x \in \mathcal{F}_D \}. \end{aligned}$$

Since $f(\cdot)$ is convex, in view of definition (2.10) of gap function, we have

$$\tfrac{1}{S_k} \delta_k(D) \;=\; \tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i f(x_i) - \hat{f}_N(D) \;\geq\; f(\hat{x}_{k+1}) - f_D^*. \tag{3.2}$$

Thus, we can justify the rate of convergence of methods (2.21) and (2.22) as applied to problem (3.1). In the estimates below we assume that

$$\|g\|_* \leq L \quad \forall g \in \partial f(x), \; \forall x \in Q.$$

1. Simple averages. In view of Theorem 2 and inequalities (2.20), (3.2) we have

$$f(\hat{x}_{k+1}) - f_D^* \leq \tfrac{0.5 + \sqrt{2k+1}}{k+1} \left(\gamma D + \tfrac{L^2}{2\sigma\gamma} \right). \tag{3.3}$$

Note that parameters D and L are not used explicitly in this scheme. However, their estimates are needed for choosing a reasonable value of γ. The optimal choice of γ is as follows:

$$\gamma^* = \tfrac{L}{\sqrt{2\sigma D}}.$$

In the method of Simple Dual Averages (SDA), the accumulated lower linear model is remarkably simple:

$$l_k(x) = \tfrac{1}{k+1} \sum_{i=0}^{k} [f(x_i) + \langle g_i, x - x_i \rangle].$$

Thus, its algorithmic scheme looks as follows:

$$\boxed{ x_{k+1} \;=\; \arg\min_{x \in Q} \left\{ \tfrac{1}{k+1} \sum_{i=0}^{k} [f(x_i) + \langle g_i, x - x_i \rangle] + \mu_k d(x) \right\}, \quad k \geq 0, } \tag{3.4}$$

where $\{\mu_k\}_{k=0}^{\infty}$ is a sequence of positive scaling parameters. In accordance to the rules of (2.21), we choose $\mu_k = \frac{\beta_{k+1}}{k+1}$. However, the rate of convergence of the method (3.4) remains on the same level for any $\mu_k = O\left(\frac{1}{\sqrt{k}}\right)$.

Note that for method (3.4) we do not need Q to be bounded. For example, we can have $Q = E$. Nevertheless, if the solution x^* of problem (3.1) do exist, then, in accordance to Theorem 2, the generated sequence $\{x_k\}_{k=0}^{\infty}$ is bounded. This feature may look surprising since $\mu_k \to 0$ and no special caution is taken in (3.4) to ensure the boundedness of the sequence.

2. Weighted averages. In view of Theorem 3 and inequalities (2.20), (3.2) we have

$$f(\hat{x}_{k+1}) - f_D^* \leq \frac{0.5+\sqrt{2k+1}}{(k+1)\sqrt{\sigma}} L \left(\frac{1}{\rho} D + \frac{\rho}{2} \right). \tag{3.5}$$

As in (2.21), parameters D and L are not used explicitly in method (2.22). But now, in order to choose a reasonable value of ρ, we need a reasonable estimate only for D. The optimal choice of ρ is as follows:

$$\rho^* = \sqrt{2D}.$$

3.2 Primal-dual problem

For the sake of notation, in this section we assume that the prox-center of the set Q is at the origin:

$$x_0 = \arg\min_{x \in Q} d(x) \; = \; 0 \in E. \tag{3.6}$$

Let us assume that the optimal solution x^* of problem (3.1) exists. Then, choosing $D \geq d(x^*)$, we can rewrite this problem in an equivalent form:

$$f^* \; = \; \min_x \{ f(x) : \; x \in \mathcal{F}_D \}. \tag{3.7}$$

Consider the conjugate function

$$f_*(s) = \sup_{x \in E} [\, \langle s, x \rangle - f(x) \,]. \tag{3.8}$$

Note that for any $x \in E$ we have

$$\partial f(x) \subseteq \operatorname{dom} f_*. \tag{3.9}$$

Since $\operatorname{dom} f = E$, a converse representation to (3.8) is always valid:

$$f(x) = \max_s [\, \langle s, x \rangle - f_*(s) : \; s \in \operatorname{dom} f_* \,], \quad x \in E.$$

Hence, we can rewrite the problem (3.7) in a dual form:

$$\begin{aligned} f^* \; &= \; \min_{x \in \mathcal{F}_D} \max_{s \in \operatorname{dom} f_*} [\, \langle s, x \rangle - f_*(s) \,] \\ &= \; \max_{s \in \operatorname{dom} f_*} \min_{x \in \mathcal{F}_D} [\, \langle s, x \rangle - f_*(s) \,] \\ &= \; \max_{s \in \operatorname{dom} f_*} [\, -\xi_D(-s) - f_*(s) \,]. \end{aligned}$$

Thus, we come to the following dual problem:

$$-f^* \; = \; \min_s [\, f_*(s) + \xi_D(-s) : \; s \in \operatorname{dom} f_* \,]. \tag{3.10}$$

As usual, it is worth to unify (3.1) and (3.10) in a single *primal-dual* problem:

$$0 \; = \; \min_{x,s} [\, \psi_D(x,s) \stackrel{\text{def}}{=} f(x) + f_*(s) + \xi_D(-s) : \; x \in Q, \; s \in \operatorname{dom} f_* \,]. \tag{3.11}$$

Let us show that DA-methods converge to a solution of this problem.

Theorem 4 *Let pair $(\hat{x}_{k+1}, \hat{s}_{k+1})$ be defined by (2.9) with sequences X_k, G_k and Λ_k being generated by method (2.14) for problem (3.7). Then $\hat{x}_{k+1} \in Q$, $\hat{s}_{k+1} \in \operatorname{dom} f_*$, and*

$$\psi_D(\hat{x}_{k+1}, \hat{s}_{k+1}) \;\leq\; \tfrac{1}{S_k}\delta_k(D). \tag{3.12}$$

Proof:
Indeed, point $\hat{x}_{k+1}$ is feasible since it is a convex combination of feasible points. In view of inclusion (3.9), same arguments also work for $\hat{s}_{k+1}$. Finally,

$$\begin{aligned}
\psi_D(\hat{x}_{k+1}, \hat{s}_{k+1}) &= \xi_D(-\hat{s}_{k+1}) + f(\hat{x}_{k+1}) + f_*(\hat{s}_{k+1}) \\
&\overset{(2.9)}{\leq} \xi_D(-\hat{s}_{k+1}) + \tfrac{1}{S_k}\sum_{i=0}^{k} \lambda_i [f(x_i) + f_*(g_i)] \\
&\overset{(2.1),\,(3.8)}{=} \tfrac{1}{S_k}\xi_D(-s_{k+1}) + \tfrac{1}{S_k}\sum_{i=0}^{k} \lambda_i \langle g_i, x_i \rangle \\
&\overset{(2.11),\,(3.6)}{=} \tfrac{1}{S_k}\delta_k(D).
\end{aligned}$$

□

Thus, for corresponding methods, the right-hand sides of inequalities (3.3), (3.5) establish the rate of convergence of primal-dual function in (3.12) to zero. In the case of interior solution x^*, the optimal solution of the dual problem (3.10) is attained at $0 \in E^*$. In this case the rate of convergence of $\hat{s}_{k+1}$ to the origin is given by (2.17).

Example of the dual problem considered in this section demonstrates the general primal-dual abilities of DA-methods (2.14). However, as we will see in the next sections, these schemes can provide us with much more interesting dual information. For that we need to employ an available information on the structure of minimization problem.

3.3 Minimax problems

Consider the following variant of problem (3.1):

$$\min_x \{f(x) = \max_{1 \leq j \leq p} f_j(x) :\; x \in Q\}, \tag{3.13}$$

where $f_j(\cdot)$, $j = 1, \ldots, p$, are convex functions defined on E. Of course, this problem admits a dual representation. Let us fix $D \geq d(x^*)$. Recall that we denote by Δ_p a standard simplex in R^p:

$$\Delta_p \;=\; \left\{ y \in R^p_+ :\; \sum_{j=1}^{p} y^{(j)} = 1 \right\}.$$

Then

$$\begin{aligned}
f^* &= \min_{x \in Q} \max_{1 \leq j \leq p} f_j(x) \\
&= \min_{x \in \mathcal{F}_D} \max_{y \in \Delta_p} \sum_{j=1}^{p} y^{(j)} f_j(x) \\
&= \max_{y \in \Delta_p} \min_{x \in \mathcal{F}_D} \sum_{j=1}^{p} y^{(j)} f_j(x).
\end{aligned}$$

Thus, defining $\phi_D(y) = \min\limits_{x \in \mathcal{F}_D} \sum\limits_{j=1}^{p} y^{(j)} f_j(x)$, we obtain the dual problem

$$f^* \;=\; \max_{y \in \Delta_p} \phi_D(y). \tag{3.14}$$

Let us show that the DA-schemes generate also an approximation to the optimal solution of the dual problem. For that, we need to employ the structure of the oracle $\mathcal{G}$ for the objective function in (3.13). Note that in our case

$$\begin{aligned} \partial f(x) &= \mathrm{Conv}\,\{\partial f_j(x) : \; j \in I(x)\}, \\ I(x) &= \{j : \; f_j(x) = f(x)\}. \end{aligned}$$

Thus, for any g_k in the method (2.14) as applied to the problem (3.13) we can define a vector $y_k \in \Delta_p$ such that

$$\begin{aligned} y_k^{(j)} &= 0, \quad j \notin I(x_k), \\ g_k &= \sum_{j \in I(x_k)} y_k^{(j)} g_{k,j}, \end{aligned} \tag{3.15}$$

where $g_{k,j} \in \partial f_j(x_k)$ for $j \in I(x_k)$. Denote

$$\hat{y}_{k+1} \;=\; \tfrac{1}{S_k} \sum_{i=0}^{k} y_i.$$

Theorem 5 *Let pair $(\hat{x}_{k+1}, \hat{y}_{k+1})$ be defined by sequences X_k, G_k and Λ_k generated by method (2.14) for problem (3.13). Then this pair is primal-dual feasible and*

$$0 \;\le\; f(\hat{x}_{k+1}) - \phi_D(\hat{y}_{k+1}) \;\le\; \tfrac{1}{S_k} \delta_k(D). \tag{3.16}$$

Proof:
Indeed, the pair $(\hat{x}_{k+1}, \hat{y}_{k+1})$ is feasible in view of convexity of primal-dual set $Q \times \Delta_p$. Further, denote $F(x) = (f_1(x), \ldots, f_p(x))^T \in R^p$. Then in view of (3.15), for any $k \ge 0$ we have

$$\begin{aligned} \langle g_k, x_k - x \rangle &= \sum_{j \in I(x_k)} y_k^{(j)} \langle g_{k,j}, x_k - x \rangle \\ &\ge \sum_{j \in I(x_k)} y_k^{(j)} [f_j(x_k) - f_j(x)] \\ &= \langle y_k, F(x_k) - F(x) \rangle \\ &= f(x_k) - \langle y_k, F(x) \rangle. \end{aligned}$$

Therefore

$$\begin{aligned}
\tfrac{1}{S_k}\delta_k(D) &= \tfrac{1}{S_k}\max_{x\in\mathcal{F}_D}\left\{\sum_{i=0}^{k}\lambda_i\langle g_i, x_i - x\rangle\right\} \\
&\geq \tfrac{1}{S_k}\max_{x\in\mathcal{F}_D}\left\{\sum_{i=0}^{k}\lambda_i[f(x_i) - \langle y_i, F(x)\rangle]\right\} \\
&= \tfrac{1}{S_k}\sum_{i=0}^{k}\lambda_i f(x_i) - \phi_D(\hat{y}_{k+1}) \;\geq\; f(\hat{x}_{k+1}) - \phi_D(\hat{y}_{k+1}).
\end{aligned}$$

□

Let us write down an explicit form of SDA-method as appied to problem (3.13). Denote by e_j the jth coordinate vector in R^p.

Initialization: Set $l_0(x) \equiv 0$, $m_0 = 0 \in Z^p$.
Iteration ($k \geq 0$)**:** **1.** Choose any $j_k^* :\ f_{j_k^*}(x_k) = f(x_k)$. **2.** Set $l_{k+1}(x) = \frac{k}{k+1}l_k(x) + \frac{1}{k+1}[f(x_k) + \langle g_{k,j_k^*}, x - x_k\rangle]$. **3.** Compute $x_{k+1} = \arg\min_{x\in Q}\left\{l_{k+1}(x) + \frac{\gamma\hat{\beta}_{k+1}}{k+1}d(x)\right\}$. **4.** Update $m_{k+1} = m_k + e_{j_k^*}$.
Output: $\hat{x}_{k+1} = \frac{1}{k+1}\sum_{i=0}^{k} x_i, \quad \hat{y}_{k+1} = \frac{1}{k+1}m_{k+1}$.

(3.17)

In (3.17), the entry of the optimal dual vector is approximated by the *frequency* of detecting the corresponding functional component as the maximal one.

Note that SDA-method can be applied also to a *continuous* minimax problem. In this case the objective function has the following form:

$$f(x) = \max_y[\langle y, F(x)\rangle :\ y \in Q_d\},$$

where Q_d is a bounded closed convex set in R^p. For this problem an approximation to the optimal dual solution is obtained as an average of the vectors $y(x_k)$ defined by

$$y(x) \in \mathrm{Arg}\max_y[\langle y, F(x)\rangle :\ y \in Q_d\}.$$

Corresponding modifications of the method (3.17) are straightforward.

3.4 Problems with simple functional constraints

Let us assume that the set Q in problem (3.1) has the following structure:

$$Q \;=\; \{\; x \in \bar{Q} : \quad Ax = b, \quad F(x) \le 0 \;\}, \tag{3.18}$$

where $\bar{Q}$ is a closed convex set in E, $\dim E = n$, $b \in R^m$, A is an $m \times n$-matrix, and $F(x) : \bar{Q} \to R^p$ is a component-wise convex function. Usually, depending on the importance of corresponding functional components, we can freely decide whether they should be hidden in the set $\bar{Q}$ or not. In any case, representation of the set Q in the form (3.18) is not unique; therefore we call the corresponding dual problem the *adjoint* one.

Let us introduce dual variables $u \in R^m$ and $v \in R^p_+$. Assume that the optimal solution x^* of (3.1), (3.18) exists and $D \ge d(x^*)$. Denote

$$\phi_D(u,v) \;=\; \min_{x \in \bar{Q}} \{ f(x) + \langle u, b - Ax \rangle + \langle v, F(x) \rangle : \; d(x) \le D \}. \tag{3.19}$$

In this definition $d(x)$ is still a prox-function of the set Q with prox-center $x_0 \in Q$. Then the adjoint problem to (3.1), (3.18) consists in

$$\max_{u,v} \{ \phi_D(u,v) : \; u \in R^m, \; v \in R^p_+ \}. \tag{3.20}$$

Note that the complexity of computation of the objective function in this problem can be comparable with the complexity of the initial problem (3.1). Nevertheless, we will see that DA-methods (2.14) are able to approach its optimal solutions. Let us consider two possibilities.

1. Let us fix $D \ge d(x^*)$. Denote by $\hat{u}_k$, $\hat{v}_k$ the optimal multipliers for essential constraints in the following optimization problem:

$$\begin{aligned} \tfrac{1}{S_k}\delta_k \;&=\; \max_{x \in \bar{Q}} \left\{ \tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i \langle g_i, x_i - x \rangle : \; Ax = b, \; F(x) \le 0, \; d(x) \le D \right\} \\ &=\; \max_{x \in \bar{Q},\, d(x) \le D} \; \min_{u \in R^m, v \in R^p_+} \; \hat{\mathcal{L}}(x,u,v), \end{aligned} \tag{3.21}$$

$$\hat{\mathcal{L}}(x,u,v) \;=\; \tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i \langle g_i, x_i - x \rangle + \langle u, Ax - b \rangle - \langle v, F(x) \rangle.$$

And let $\hat{x}_k$ be its optimal solution. Then for any $x \in \bar{Q}$ with $d(x) \le D$ we have

$$\langle -\tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i g_i + A^T \hat{u}_k - \sum_{j=1}^{p} \hat{v}_k^{(j)} \psi_j, \; x - \hat{x}_k \rangle \;\le\; 0,$$

with some $\psi_j \in \partial F^{(j)}(\hat{x}_k)$, $j = 1, \ldots, p$, and

$$\sum_{j=1}^{p} \hat{v}_k^{(j)} F^{(j)}(\hat{x}_k) = 0.$$

Therefore, for any such x we get

$$
\begin{aligned}
& f(x) + \langle \hat{u}_k, b - Ax \rangle + \langle \hat{v}_k, F(x) \rangle \\
\geq \; & \tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i [f(x_i) + \langle g_i, x - x_i \rangle] + \langle \hat{u}_k, b - Ax \rangle + \sum_{j=1}^{p} \hat{v}_k^{(j)} [F^{(j)}(\hat{x}_k) + \langle \psi_j, x - \hat{x}_k \rangle] \\
\geq \; & \tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i [f(x_i) + \langle g_i, \hat{x}_k - x_i \rangle] \; = \; \tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i f(x_i) \; - \; \tfrac{1}{S_k} \delta_k(D).
\end{aligned}
$$

Thus, we have proved that

$$
0 \; \leq \; f(\hat{x}_{k+1}) - \phi_D(\hat{u}_k, \hat{v}_k) \; \leq \; \tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i f(x_i) - \phi_D(\hat{u}_k, \hat{v}_k) \; \leq \; \tfrac{1}{S_k} \delta_k(D). \tag{3.22}
$$

Hence, the upper estimate for quality of this primal-dual pair can be derived from Item 1 of Theorem 1.

2. The above suggestion to form an approximate solution to the adjoint problem (3.20) by the dual multipliers of problem (3.21) has two drawbacks. First of all, it is necessary to guarantee that the auxiliary parameter D is sufficiently big. Secondly, the problem (3.21) needs additional computational efforts. Let us show that the approximate solution to problem (3.20) can be obtained for free, using the objects necessary for finding the point $\pi_{\beta_{k+1}}(-s_{k+1})$.

Denote by $\bar{u}_k$, $\bar{v}_k$ the optimal multipliers for functional constraints in the following optimization problem:

$$
\begin{aligned}
x_{k+1} = \pi_{\beta_{k+1}}(-s_{k+1}) \; &= \; \arg\max_{x \in \bar{Q}} \left\{ -\tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i \langle g_i, x \rangle - \tfrac{\beta_{k+1}}{S_k} d(x) : \; Ax = b, \; F(x) \leq 0 \right\} \\
&= \; \arg\max_{x \in \bar{Q}} \; \min_{u \in R^m, v \in R^p_+} \; \bar{\mathcal{L}}(x, u, v), \\
\bar{\mathcal{L}}(x, u, v) \; &= \; -\tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i \langle g_i, x \rangle - \tfrac{\beta_{k+1}}{S_k} d(x) + \langle u, Ax - b \rangle - \langle v, F(x) \rangle.
\end{aligned} \tag{3.23}
$$

Then for any $x \in \bar{Q}$ we have

$$
\langle -\tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i g_i - \tfrac{\beta_{k+1}}{S_k} d' + A^T \bar{u}_k - \sum_{j=1}^{p} \bar{v}_k^{(j)} \psi_j, \; x - x_{k+1} \rangle \; \leq \; 0,
$$

with some $d' \in \partial d(x_{k+1})$ and $\psi_j \in \partial F^{(j)}(x_{k+1})$, $j = 1, \ldots, p$, and

$$
\sum_{j=1}^{p} \bar{v}_k^{(j)} F^{(j)}(x_{k+1}) = 0.
$$

Therefore, for all such x we obtain

$$\begin{aligned} & f(x) + \langle \bar{u}_k, b - Ax \rangle + \langle \bar{v}_k, F(x) \rangle \\ \geq & \frac{1}{S_k} \sum_{i=0}^{k} \lambda_i [f(x_i) + \langle g_i, x - x_i \rangle] + \langle \bar{u}_k, b - Ax \rangle + \sum_{j=1}^{p} \bar{v}_k^{(j)} [F^{(j)}(x_{k+1}) + \langle \psi_j, x - x_{k+1} \rangle] \\ \geq & \frac{1}{S_k} \sum_{i=0}^{k} \lambda_i [f(x_i) + \langle g_i, x_{k+1} - x_i \rangle] + \frac{\beta_{k+1}}{S_k} \langle d', \ x_{k+1} - x \ \rangle \\ \geq & \frac{1}{S_k} \sum_{i=0}^{k} \lambda_i f(x_i) - \frac{\beta_{k+1}}{S_k} d(x) - \frac{1}{S_k} \left[\sum_{i=0}^{k} \lambda_i \langle g_i, x_i - x_{k+1} \rangle - \beta_{k+1} d(x_{k+1}) \right]. \end{aligned}$$

Since in definition (3.19) we need $d(x) \leq D$, in view of (2.12) the above estimates results in

$$\phi_D(\bar{u}_k, \bar{v}_k) \ \geq \frac{1}{S_k} \sum_{i=0}^{k} \lambda_i f(x_i) - \frac{1}{S_k} \Delta(\beta_{k+1}, D). \tag{3.24}$$

Therefore, as above, the upper estimate for quality of this primal-dual pair can be derived from Item 1 of Theorem 1.

4 Saddle point problems

Consider the general saddle point problem:

$$\min_{u \in U} \max_{v \in V} \ f(u, v), \tag{4.1}$$

where U is a closed convex set in E_u, V is a closed convex set in E_v, function $f(\cdot, v)$ is convex in the first argument on E_u for any $v \in V$, and function $f(u, \cdot)$ is concave in the second argument on E_v for any $u \in U$.

For set U we assume existence of prox-function $d_u(\cdot)$ with prox-center u_0, which is strongly convex on U with respect to the norm $\| \cdot \|_u$ with convexity parameter σ_u. For the set V we introduce the similar assumptions.

Note that the point $x^* = (u^*, v^*) \in Q \stackrel{\text{def}}{=} U \times V$ is a solution to the problem (4.1) if and only if

$$f(u^*, v) \leq f(u^*, v^*) \leq f(u, v^*) \quad \forall u \in U, \ v \in V.$$

Since for any $x \stackrel{\text{def}}{=} (u, v) \in Q$ we have

$$\begin{aligned} f(u, v^*) & \leq f(u, v) + \langle g_v, v^* - v \rangle \quad \forall g_v \in \partial f_v(u, v), \\ f(u^*, v) & \geq f(u, v) + \langle g_u, u^* - u \rangle \quad \forall g_u \in \partial f_u(u, v), \end{aligned}$$

we conclude that

$$\langle g_u, u - u^* \rangle + \langle -g_v, v - v^* \rangle \geq 0 \tag{4.2}$$

for any $g = (g_u, g_v)$ from $\partial f_u(u, v) \times \partial f_v(u, v)$. Thus, the oracle

$$\mathcal{G}: \quad x = (u, v) \in Q \quad \Rightarrow \quad g(x) = (g_u(x), -g_v(x)) \tag{4.3}$$

satisfies condition (2.8).

Further, let us fix some $\alpha \in (0,1)$. Then we can introduce the following prox-function of the set Q:

$$d(x) = \alpha d_u(u) + (1-\alpha) d_v(v).$$

In this case, the prox-center of Q is $x_0 = (u_0, v_0)$. Defining the norm for $E = E_u \times E_v$ by

$$\|x\| = \left[\alpha\sigma_u \|u\|_u^2 + (1-\alpha)\sigma_v \|v\|_v^2 \right]^{1/2}.$$

we get for function $d(\cdot)$ the convexity parameter $\sigma = 1$. Note that the norm for the dual space $E^* = E_u^* \times E_v^*$ is defined now as

$$\|g\|_* = \left[\tfrac{1}{\alpha\sigma_u} \|g_u\|_{u,*}^2 + \tfrac{1}{(1-\alpha)\sigma_v} \|g_v\|_{v,*}^2 \right]^{1/2}.$$

Assuming now that partial subdifferentials of function f are uniformly bounded,

$$\|g_u\|_{u,*} \le L_u, \quad \|g_v\|_{v,*} \le L_v, \quad \forall g \in \partial f_u(u,v) \times \partial f_v(u,v), \ \forall (u,v) \in Q,$$

we obtain the following bound for the answers of oracle $\mathcal{G}$:

$$\|g\|_*^2 \le L^2 \stackrel{\text{def}}{=} \tfrac{L_u^2}{\alpha\sigma_u} + \tfrac{L_v^2}{(1-\alpha)\sigma_v}. \tag{4.4}$$

Finally, if $d_u(u^*) \le D_u$ and $d_v(v^*) \le D_v$, then $x^* = (u^*, v^*) \in \mathcal{F}_D$ with

$$D = \alpha D_u + (1-\alpha) D_v. \tag{4.5}$$

Let us apply now to (4.1) SDA-method (2.21). In accordance to Theorem 2, we get the following bound:

$$\tfrac{1}{S_k}\delta_k(D) \le \tfrac{\hat\beta_{k+1}}{k+1}\,\Omega, \quad \Omega \stackrel{\text{def}}{=} \gamma D + \tfrac{L^2}{2\gamma},$$

with L and D defined by (4.4) and (4.5) respectively. With optimal $\gamma = \frac{L}{\sqrt{2D}}$ we obtain

$$\begin{aligned} \Omega &= [2L^2 D]^{1/2} = \left[2\left(\tfrac{L_u^2}{\alpha\sigma_u} + \tfrac{L_v^2}{(1-\alpha)\sigma_v}\right) \cdot (\alpha D_u + (1-\alpha) D_v)\right]^{1/2} \\ &= \left[2\left(\tfrac{L_u^2 D_u}{\sigma_u} + \tfrac{L_v^2 D_v}{\sigma_v} + \tfrac{\alpha L_v^2 D_u}{(1-\alpha)\sigma_v} + \tfrac{(1-\alpha) L_u^2 D_v}{\alpha\sigma_u}\right)\right]^{1/2} \end{aligned}$$

Minimizing the latter expression in α we obtain

$$\Omega = \sqrt{2}\left(L_u\sqrt{\tfrac{D_u}{\sigma_u}} + L_v\sqrt{\tfrac{D_v}{\sigma_v}}\right).$$

Thus, we have shown that under an appropriate choice of parameters we can ensure for SDA-method the following rate of convergence for the gap function:

$$\tfrac{1}{S_k}\delta_k(D) \le \tfrac{\hat\beta_{k+1}}{k+1}\,\sqrt{2}\left(L_u\sqrt{\tfrac{D_u}{\sigma_u}} + L_v\sqrt{\tfrac{D_v}{\sigma_v}}\right). \tag{4.6}$$

It remains to show that the vanishing gap results in approaching the solution of the saddle point problem (4.1).

Let us introduce two auxiliary functions:

$$\begin{aligned} \phi(u) &= \max_{v \in V} \{ f(u,v) : \; d_v(v) \leq D_v \}, \\ \psi(v) &= \min_{u \in U} \{ f(u,v) : \; d_u(u) \leq D_u \}. \end{aligned}$$

In view of our assumptions, $\phi(\cdot)$ is convex on U and $\psi(\cdot)$ is concave on V. Moreover, for any $u \in U$ and $v \in V$ we have

$$\psi(v) \leq f^* \leq \phi(u),$$

where $f^* = f(u^*, v^*)$.

Theorem 6 *Let point $\hat{x}_{k+1} = (\hat{u}_{k+1}, \hat{v}_{k+1})$ be defined by (2.9) with sequences X_k, G_k and Λ_k being generated by method (2.14) for problem (4.1) with oracle $\mathcal{G}$ defined by (4.3). Then $\hat{x}_{k+1} \in Q$ and*

$$0 \leq \phi(\hat{u}_{k+1}) - \psi(\hat{v}_{k+1}) \; \leq \; \tfrac{1}{S_k} \delta_k(D). \tag{4.7}$$

Proof:
Indeed,

$$\begin{aligned} \tau_k &\stackrel{\text{def}}{=} \tfrac{1}{S_k} \max_{u \in U} \left\{ \sum_{i=0}^{k} \lambda_i \langle g_u(u_i, v_i), u_i - u \rangle : \; d_u(u) \leq D_u \right\} \\ &\geq \tfrac{1}{S_k} \max_{u \in U} \left\{ \sum_{i=0}^{k} \lambda_i [f(u_i, v_i) - f(u, v_i)] : \; d_u(u) \leq D_u \right\} \\ &\geq \tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i f(u_i, v_i) - \min_{u \in U} \{ f(u, \hat{v}_{k+1}) : \; d_u(u) \leq D_u \} \\ &= \tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i f(u_i, v_i) - \psi(\hat{v}_{k+1}). \end{aligned}$$

Similarly,

$$\begin{aligned} \sigma_k &\stackrel{\text{def}}{=} \tfrac{1}{S_k} \max_{v \in V} \left\{ \sum_{i=0}^{k} \lambda_i \langle g_v(u_i, v_i), v - v_i \rangle : \; d_v(v) \leq D_v \right\} \\ &\geq \tfrac{1}{S_k} \max_{v \in V} \left\{ \sum_{i=0}^{k} \lambda_i [f(u_i, v) - f(u_i, v_i)] : \; d_v(v) \leq D_v \right\} \\ &\geq \max_{v \in V} \{ f(\hat{u}_{k+1}, v) : \; d_v(v) \leq D_v \} - \tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i f(u_i, v_i) \\ &= \phi(\hat{u}_{k+1}) - \tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i f(u_i, v_i). \end{aligned}$$

Thus,

$$\begin{aligned}\phi(\hat{u}_{k+1}) - \psi(\hat{v}_{k+1}) &\leq \tau_k + \sigma_k \\ &\leq \tfrac{1}{S_k} \max_{x \in Q} \left\{ \sum_{i=0}^{k} \lambda_i \langle g(x_i), x_i - x \rangle : \; d(x) \leq \alpha D_u + (1-\alpha) D_v \right\} \\ &= \tfrac{1}{S_k} \delta_k(D).\end{aligned}$$

□

Note that we did not assume boundedness of the sets U and V. Parameter D from inequality (4.7) does not appear in DA-method (2.14) in explicit form.

5 Variational inequalities

Let Q be a closed convex set endowed, as usual, with a prox-function $d(x)$. Consider a continuous operator $V(\cdot) : \; Q \to E^*$, which is *monotone* on Q:

$$\langle V(x) - V(y), x - y \rangle \geq 0 \quad \forall x, y \in Q.$$

In this section we are interested in *variational inequality problem*:

$$\text{Find } x^* \in Q: \quad \langle V(x), x - x^* \rangle \geq 0 \quad \forall x \in Q. \tag{5.1}$$

Sometimes this problem is called a *weak variational inequality*. Since $V(\cdot)$ is continuous and monotone, the problem (5.1) is equivalent to its *strong* variant:

$$\text{Find } x^* \in Q: \quad \langle V(x^*), x - x^* \rangle \geq 0 \quad \forall x \in Q. \tag{5.2}$$

We assume that for this problem there exists at least one solution x^*.

Let us fix $D > d(x^*)$. A standard measure for quality of approximate solution to (5.1) is the *restricted merit function*

$$f_D(x) = \max_{y \in Q} \{ \langle V(y), x - y \rangle : \; d(y) \leq D \}. \tag{5.3}$$

Let us present two important facts related to this function (our exposition is partly taken from Section 2 in [12]).

Lemma 4 *Function $f_D(x)$ is well defined and convex on E. For any $x \in Q$ we have $f_D(x) \geq 0$, and $f_D(x^*) = 0$. Conversely, if $f_D(\hat{x}) = 0$ for some $\hat{x} \in Q$ with $d(\hat{x}) < D$, then $\hat{x}$ is a solution to problem (5.1).*

Proof:
Indeed, function $f_D(x)$ is well defined since the set $\mathcal{F}_D$ is bounded. It is convex in x as a maximum of a parametric family of linear functions. If $x \in \mathcal{F}_D$, then taking in (5.3) $y = x$ we guarantee $f_D(x) \geq 0$. If $x \in Q$ and $d(x) > D$, consider $y \in Q$ satisfying condition

$$d(y) = D, \quad y = \alpha x + (1-\alpha) x^*,$$

for certain $\alpha \in (0,1)$. Then $y \in \mathcal{F}_D$ and

$$\langle V(y), x - y \rangle = \frac{1-\alpha}{\alpha} \langle V(y), y - x^* \rangle \geq 0.$$

Thus, $f_D(x) \geq 0$ for all $x \in Q$. On the other hand, since x^* is a solution to (5.1), $\langle V(y), x^* - y \rangle \leq 0$ for all y from $\mathcal{F}_D$. Hence, since $x^* \in \mathcal{F}_D$, we get $f_D(x^*) = 0$.

Finally, let $f_D(\hat{x}) = 0$ for some $\hat{x} \in Q$ and $d(\hat{x}) < D$. This means that $\hat{x}$ is a solution of the weak variational inequality

$$\langle V(y), y - \hat{x} \rangle \geq 0 \quad \forall y \in \mathcal{F}_D.$$

Since $V(y)$ is continuous, we conclude that $\hat{x}$ is a solution to the strong variational inequality

$$\langle V(\hat{x}), y - \hat{x} \rangle \geq 0 \quad \forall y \in \mathcal{F}_D.$$

Hence, the minimum of linear function $l(y) = \langle V(\hat{x}), y \rangle$, $y \in \mathcal{F}_D$, is attained at $y = \hat{x}$. Moreover, at this point the constraint $d(y) \leq D$ is not active. Thus, this constraint can be removed and we conclude that $\hat{x}$ is a solution to (5.2) and, consequently, to (5.1). $\square$

Let us use for problem (5.1) the oracle $\mathcal{G} : \; x \to V(x)$.

Lemma 5 *For any $k \geq 0$ we have*

$$f_D(\hat{x}_{k+1}) \leq \frac{1}{S_k} \delta_k(D). \tag{5.4}$$

Proof:
Indeed, since $V(x)$ is a monotone operator,

$$\begin{aligned} f_D(\hat{x}_{k+1}) &= \max_x \{ \langle V(x), \hat{x}_{k+1} - x \rangle : \; x \in \mathcal{F}_D \} \\ &= \frac{1}{S_k} \max_x \left\{ \sum_{i=0}^{k} \lambda_i \langle V(x), x_i - x \rangle : \; x \in \mathcal{F}_D \right\} \\ &\leq \frac{1}{S_k} \max_x \left\{ \sum_{i=0}^{k} \lambda_i \langle V(x_i), x_i - x \rangle : \; x \in \mathcal{F}_D \right\} \equiv \frac{1}{S_k} \delta_k(D). \end{aligned}$$

$\square$

Let us assume that

$$\|V(x)\|_* \leq L \quad \forall x \in Q.$$

Then, for example, by Theorem 2 we conclude that SDA-method (2.21) as applied to the problem (5.1) converges with the following rate:

$$f_D(\hat{x}_{k+1}) \leq \frac{\hat{\beta}_{k+1}}{k+1} \left(\gamma d(x^*) + \frac{L^2}{2\sigma\gamma} \right). \tag{5.5}$$

Note that the sequence $\{x_k\}_{k=0}^{\infty}$ remains bounded even for unbounded feasible set Q:

$$\|x_k - x^*\|^2 \leq \frac{2}{\sigma} d(x^*) + \frac{L^2}{\sigma^2 \gamma^2}. \tag{5.6}$$

6 Stochastic optimization

Let Ξ be a probability space endowed with a probability measure, and let Q be a closed convex set in E endowed with a prox-function $d(x)$. We are given a cost function

$$f: \; Q \times \Xi \mapsto R.$$

This mapping defines a family of random variables $f(x, \xi)$ on Ξ. We assume that the expectation in ξ, $\mathbf{E}_\xi(f(x, \xi))$ is well-defined for all $x \in Q$. Our problem of interest is the stochastic optimization problem:

$$\min_x \{\phi(x) \equiv \mathbf{E}_\xi(f(x, \xi)) : \; x \in Q\}. \tag{6.1}$$

We assume that problem (6.1) is solvable and denote by x^* one of its solutions with $\phi^* \stackrel{\text{def}}{=} \phi(x^*)$.

In this section we make the standard convexity assumption.

Assumption 1 *For any $\xi \in \Xi$, function $f(\cdot, \xi)$ is convex and subdifferentiable on Q.*

Then, the function $\phi(x)$ is convex on Q.

Note that computation of the exact value of $\phi(x)$ may not be numerically possible, even when the distribution of ξ is known. Indeed, if $\Xi \subset R^m$, then computing $\hat{\phi}$, such that $\mid \hat{\phi} - \phi(x) \mid \leq \hat{\epsilon}$, may involve up to $O\left(\frac{1}{\hat{\epsilon}^m}\right)$ computations of $f(x, \xi)$ for different $\xi \in \Xi$. Hence, the standard deterministic notion of approximate solution to problem (6.1) becomes useless.

However, an alternative definition of the solution to (6.1) is possible. Indeed, it is clear that no solution concept can exclude failures in actual implementation. This is the very nature of decision under uncertainty not to have full control over the consequences of a given decision. In this context, there is no reason to require that the computed solution be the result of some deterministic process. A random computational scheme would be equally acceptable, if the solution meets some probabilistic criterion. Namely, let x be a random output of some random algorithm. It appears, that it is enough to assume that this output is good only *in average*:

$$\mathbf{E}(\phi(x)) - \phi^* \leq \epsilon. \tag{6.2}$$

Then, as it was shown in [13], any random optimization algorithm, which is able to generate such random output for any $\epsilon > 0$, can be transformed in an algorithm generating random solutions with an appropriate level of confidence. In [13] this transformation was justified for a *stochastic version* of subgradient method (1.3) (which is known to possess the feature (6.2)).

The goal of this section is to prove that a stochastic version of SDA-method (2.21) as applied to problem (6.1) is also able to produce a random variable $\tilde{x} \in Q$ satisfying condition (6.2).

In order to justify convergence of the method we need more assumptions.

Assumption 2 *At any point $x \in Q$ and any implementation of random vector $\xi \in \Xi$ the stochastic oracle $\mathcal{G}(x, \xi)$ always returns a predefined answer $f'(x, \xi) \in \partial_x f(x, \xi)$. Moreover, the answers of the oracle are uniformly bounded:*

$$\|f'(x, \xi)\|_* \; \leq \; L, \quad \forall x \in Q, \; \forall \xi \in \Xi.$$

Let us write down a stochastic modification of SDA-method in the same vein as [13]. In this section, we adopt notation $\tilde{\tau}$ for a particular result of drawing of a random variable τ. Thus, all $\tilde{\xi}_k \in R^m$ in the scheme below denote some observations of corresponding i.i.d. random vectors ξ_k, which are the identical copies of the random vector ξ defined in (6.1).

Method of Stochastic Simple Averages
Initialization: Set $\tilde{s}_0 = 0 \in E^*$. Choose $\gamma > 0$.
Iteration ($k \geq 0$)**:** **1.** Generate $\tilde{\xi}_k$ and compute $\tilde{g}_k = f'(\tilde{x}_k, \tilde{\xi}_k)$. **2.** Set $\tilde{s}_{k+1} = \tilde{s}_k + \tilde{g}_k$, $\tilde{x}_{k+1} = \pi_{\gamma\hat{\beta}_{k+1}}(-\tilde{s}_{k+1})$.

(6.3)

Denote by $\boldsymbol{\xi}_k$, $k \geq 0$, the collection of random variables $(\xi_0, \ldots, \xi_k)$. SSA-method (6.3) defines two sequences of random vectors

$$s_{k+1} = s_{k+1}(\boldsymbol{\xi}_k) \in E^*, \quad x_{k+1} = x_{k+1}(\boldsymbol{\xi}_k) \in Q, \quad k \geq 0,$$

with deterministic s_0 and x_0. Note that their implementations $\tilde{s}_{k+1}$ and $\tilde{x}_{k+1}$ are different for different runs of SSA-method.

Thus, for each particular run of this method, we can define a gap function

$$\tilde{\delta}_k(D) = \max_x \left\{ \sum_{i=0}^{k} \langle f'(\tilde{x}_i, \tilde{\xi}_i), \tilde{x}_i - x \rangle : \; x \in \mathcal{F}_D, \; \right\}, \quad D \geq 0, \tag{6.4}$$

which can be seen as a drawing of random variable $\delta_k(D)$. Since $\|\tilde{g}_k\|_* \leq L$, Theorem 2 results in the following uniform bound:

$$\tfrac{1}{S_k}\tilde{\delta}_k(D) \;\leq\; \tfrac{\hat{\beta}_{k+1}}{k+1}\left(\gamma D + \tfrac{L^2}{2\sigma\gamma}\right). \tag{6.5}$$

Let us choose now D big enough: $D \geq d(x^*)$. Then

$$\begin{aligned} \tfrac{1}{S_k}\tilde{\delta}_k(D) &\geq \tfrac{1}{k+1}\sum_{i=0}^{k} \langle f'(\tilde{x}_i, \tilde{\xi}_i), \tilde{x}_i - x^* \rangle \\ &\geq \tfrac{1}{k+1}\sum_{i=0}^{k} [f(\tilde{x}_i, \tilde{\xi}_i) - f(x^*, \tilde{\xi}_i)] \, . \end{aligned}$$

Since all ξ_i are i.i.d. random vectors, in average we have

$$\mathbf{E}_{\boldsymbol{\xi}_k}\left(\tfrac{1}{S_k}\delta_k(D)\right) \;\geq\; \tfrac{1}{k+1}\sum_{i=0}^{k} \mathbf{E}_{\boldsymbol{\xi}_k}\left(f(x_i, \xi_i)\right) - \phi^*. \tag{6.6}$$

Note that the random vector x_i is independent on ξ_j, $i \le j \le k$. Therefore,

$$\mathbf{E}_{\boldsymbol{\xi}_k}\left(f(x_i,\xi_i)\right) = \mathbf{E}_{\boldsymbol{\xi}_k}\left(\phi(x_i)\right),$$

and we conclude that

$$\begin{aligned} \tfrac{1}{k+1}\sum_{i=0}^{k}\mathbf{E}_{\boldsymbol{\xi}_k}\left(f(x_i,\xi_i)\right) &= \tfrac{1}{k+1}\sum_{i=0}^{k}\mathbf{E}_{\boldsymbol{\xi}_k}\left(\phi(x_i)\right) \\ &\ge \mathbf{E}_{\boldsymbol{\xi}_k}\left(\phi\left(\tfrac{1}{k+1}\sum_{i=0}^{k}x_i\right)\right). \end{aligned}$$

Thus, in view of inequalities (6.5), (6.6), we have proved the following theorem.

Theorem 7 *Let the random sequence $\{x_k\}_{k=0}^{\infty}$ be defined by SSA-method (6.3). Then for any $k \ge 0$ we have*

$$\mathbf{E}_{\boldsymbol{\xi}_k}\left(\phi\left(\tfrac{1}{k+1}\sum_{i=0}^{k}x_i\right)\right) - \phi^* \le \tfrac{\hat{\beta}_{k+1}}{k+1}\left(\gamma d(x^*) + \tfrac{L^2}{2\sigma\gamma}\right).$$

□

7 Applications in modelling

7.1 Algorithm of balanced development

Let us discuss a decision-making model, where a variant of DA-algorithm (2.14) has natural interpretation of certain rational dynamic investment strategy.

Assume we are going to develop our property by a given priori sequence $\Lambda = \{\lambda_k\}_{k=0}^{\infty}$ of consecutive investments of the capital. This money can be spent for buying certain products with unitary prices

$$p^{(j)} \ge 0, \quad j = 1,\dots,N.$$

All products are perfectly divisible; they are available on the market in unlimited quantities. Thus, if for investment k we buy quantity $X_k^{(j)} \ge 0$ of each product j, $j = 1,\dots,N$, then the balance equation can be written as

$$X_k = \lambda_k x_k, \quad \langle p, x_k\rangle = 1, \quad x_k \ge 0, \tag{7.1}$$

where the vectors $p = (p^{(1)},\dots,p^{(N)})^T$ and $x_k = (x_k^{(1)},\dots x^{(N)})^T$ are both from R^N.

In order to measure the progress of our property (or, the efficiency of our investments) we introduce a set of m characteristics (features). Let us assume that after $k \ge 0$ investments we are able to measure *exactly* an accumulated level $s_k^{(i)}$, $i = 1,\dots,m$, of each feature i in our property. On the other hand, we have a set of standards $b^{(i)} > 0$, $i = 1,\dots,m$, which are treated as lower bounds on "concentration" of each feature in a *perfectly balanced* unit of such a property.

Finally, let us assume that each product j is described by a vector

$$a_j = (a_j^{(1)}, \ldots, a_j^{(m)})^T, \quad j = 1, \ldots, N,$$

with entries being the unitary concentrations of corresponding characteristics. Introducing the matrix

$$A = (a_1, \ldots, a_N) \in R^{m \times N}$$

we can describe the dynamics of accumulated features as follows:

$$s_{k+1} = s_k + \lambda_k g_k, \quad g_k = Ax_k, \quad k \geq 0. \tag{7.2}$$

In accordance to our system of standards, the strategy of *optimal balanced development* can be found from the following optimization problem:

$$\tau^* \stackrel{\text{def}}{=} \max_{x,\tau} \{\tau : \; Ax \geq \tau b, \; \langle p, x \rangle = 1, \; x \geq 0\}. \tag{7.3}$$

Indeed, denote by x^* its optimal solution. Then, choosing in (7.2) $x_k = x^*$ we get

$$s_k = S_k \cdot Ax^* \; \geq \; S_k \cdot \tau^* b.$$

Thus, the efficiency of such an investment strategy is maximal possible and its quality even does not depend on the schedule of the payments.

However, note that in order to apply the above strategy, it is necessary to solve in advance a linear programming problem (7.3), which can be very complicated because of a high dimension. What can be done if this computation appears to be too difficult? Do we have some simple tools for approaching the optimal strategy in real life? Apparently, the latter question has a positive answer.

First of all, let us represent the problem (7.3) in a dual form. From the Lagrangian

$$\mathcal{L}(x, \tau, y, \lambda) \; = \; \tau + \langle Dy, Ax - \tau b \rangle + \lambda \cdot (1 - \langle p, x \rangle) \quad \rightarrow \quad \max_{\tau;\, x \geq 0} \min_{\lambda;\, y \geq 0},$$

where D is a diagonal matrix, $D^{(i,i)} = \frac{1}{b^{(i)}}$, $i = 1, \ldots, m$, we come to the following dual representation

$$\tau^* \; = \; \min_{y,\lambda} \{\lambda : \; A^T Dy \leq \lambda p, \; y \in \Delta_m\}.$$

Or, introducing the function $\phi(y) = \max\limits_{1 \leq j \leq N} \frac{1}{p^{(j)}} \langle a_j, Dy \rangle$, we obtain

$$\tau^* \; = \; \min_{y} \{\phi(y) : \; y \in \Delta_m\}. \tag{7.4}$$

Let us apply to problem (7.4) a variant of DA-scheme (2.14). For feasible set of this problem Δ_m, we choose the entropy prox-function

$$d(y) \; = \; \ln m + \sum_{i=1}^{m} y^{(i)} \ln y^{(i)}$$

with prox-center $y_0 = (\frac{1}{m}, \ldots, \frac{1}{m})^T$. From (1.11) we see that $d(y)$ is strongly convex on Δ_m in l_1-norm (1.10) with convexity parameter $\sigma = 1$. Note that for any $y \in \Delta_m$ we have

$$d(y) \; \leq \; \ln m. \tag{7.5}$$

Moreover, with this prox-function the computation of projection $\pi_\beta(s)$ is very cheap:

$$\pi_\beta(s)^{(i)} \;=\; e^{s^{(i)}/\beta} \cdot \left[\sum_{j=1}^{m} e^{s^{(j)}/\beta} \right]^{-1}, \quad i = 1, \dots, m, \tag{7.6}$$

(see, for example, Lemma 4 in [10]).

In the algorithm below, the sequence Λ is the same as in (7.2).

Algorithm of Balanced Development

Initialization: Set $s_0 = 0 \in E^*$. Choose $\beta_0 > 0$.

Iteration ($k \geq 0$):

1. Form the set $J_k = \left\{ j : \; \frac{1}{p^{(j)}} \langle a_j, Dy_k \rangle = \phi(y_k) \right\}$.

2. Choose any $x_k \geq 0$, $\langle p, x_k \rangle = 1$, with $x_k^{(j)} = 0$ for $j \notin J_k$.

3. Update $s_{k+1} = s_k + \lambda_k g_k$ with $g_k = Ax_k$.

4. Choose $\beta_{k+1} \geq \beta_k$ and for $i = 1, \dots, m$ define

$$y_{k+1}^{(i)} \;=\; e^{-s_{k+1}^{(i)}/(\beta_{k+1} b^{(i)})} \cdot \left[\sum_{j=1}^{m} e^{-s_{k+1}^{(j)}/(\beta_{k+1} b^{(j)})} \right]^{-1}.$$

(7.7)

Note that in this scheme $Dg_k \in \partial\phi(y_k)$ and

$$y_{k+1} = \pi_{\beta_{k+1}} \left(-D \sum_{j=0}^{k} \lambda_j g_j \right).$$

Therefore it is indeed a variant of (2.14), and in view of Theorem 1 and (7.5) we have

$$\begin{aligned} \tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i \phi(y_i) - \tau^* \;&\leq\; \tfrac{1}{S_k} \left(\beta_{k+1} \ln m + \tfrac{1}{2} L^2 \sum_{i=0}^{k} \tfrac{\lambda_i^2}{\beta_i} \right), \\ L \;=\; \max_{1 \leq j \leq N} \tfrac{1}{p^{(j)}} \|Da_j\|_* \;&=\; \max_{\substack{1 \leq j \leq N, \\ 1 \leq i \leq m}} \tfrac{1}{p^{(j)} b^{(i)}} \, |a_j^{(i)}|. \end{aligned} \tag{7.8}$$

BD-algorithm (7.7) admits the following interpretation. Its first three steps describe a natural investment strategy. Indeed, the algorithm updates a system of personal prices for characteristics

$$u_k \stackrel{\text{def}}{=} Dy_k, \quad k \geq 0.$$

By these prices, for any product j we can compute an estimate of its utility $\langle a_j, u_k \rangle$. Therefore, in Step 1 we form a subset of available products with the best *quality-price ratio* (Step 1):

$$J_k = \left\{ j : \frac{1}{p^{(j)}} \langle a_j, u_k \rangle = \max_{1 \le i \le N} \frac{1}{p^{(i)}} \langle a_i, u_k \rangle \right\}.$$

In Step 2 we share a unit of budget in an arbitrary way among the best products. And in Step 3 we buy the products within the budget λ_k and update the vector of accumulated features. A non-trivial part of interpretation is related to Step 4, which describes a dependence of the personal prices, developed *to the end* of period $k+1$, in the vector of accumulated characteristics s_{k+1} observed *during* this period. For this interpretation, we need to explain first the meaning of (7.6).

Relations (7.6) appear in so-called *logit* variant of discrete choice model (see, for example, [1]). In this model we need to choose one of m variants with utilities $s^{(i)}$, $i = 1, \ldots, m$. The utilities are observed with additive random errors $\epsilon^{(i)}$, which are i.i.d. in accordance to double-exponential distribution:

$$F(\tau) = \mathbf{Prob}\left[\epsilon^{(i)} \le \tau\right] = \exp\left\{-e^{-\gamma-\tau/\beta}\right\}, \quad i = 1, \ldots, m,$$

where $\gamma \approx 0.5772$ is the Euler's constant and $\beta > 0$ is a tolerance parameter. Then, the expected observation of the utility function is given by

$$U_\beta(s) = E\left(\max_{1 \le i \le m} [s^{(i)} + \epsilon^{(i)}]\right),$$

It can be proved that $U_\beta(s) = \beta d_*(s/\beta) + \text{const}$ with

$$d_*(s) = \ln\left(\sum_{i=0}^{m} e^{s^{(i)}}\right).$$

Therefore we have the following important expressions for so-called *choice probabilities*:

$$\mathbf{Prob}\left[s^{(i)} + \epsilon^{(i)} = \max_{1 \le j \le m} [s^{(j)} + \epsilon^{(j)}]\right] \;=\; \nabla U_\beta(s)^{(i)}$$

$$= \; \nabla d_*(s/\beta)^{(i)} \;=\; e^{s^{(i)}/\beta} \cdot \left[\sum_{j=1}^{m} e^{s^{(j)}/\beta}\right]^{-1}, \quad i = 1, \ldots, m,$$

(compare with (7.6)). The smaller β is, the closer is our choice strategy to deterministic maximum.

Using the logit model, we can adopt the following behavioral explanation of the rules of Step 4 in (7.7).

- During the period $k+1$ we regularly compare the levels of accumulated features in relative scale defined by the vector of minimal standards b. Each audit defines the worst characteristic, for which the corresponding level is minimal.
- The above computations are done with additive errors[3], which satisfy the logit model.

[3]Note that these errors can occur in a natural way, due to inexact data, for example. However, even if the actual data and measurements are exact, we need to introduce an *artificial randomization* of the results.

- To the end of the period we obtain the frequencies $y_{k+1}^{(i)}$ of detecting the level of corresponding characteristic as the worst one. Then, we apply the following prices:

$$u_{k+1}^{(i)} \quad = \quad \frac{1}{b^{(i)}} \, y_{k+1}^{(i)}, \quad i = 1, \ldots, m. \tag{7.9}$$

Note that the conditions for convergence of BD-method can be derived from the right-hand side of inequality (7.8). Assume for example, that we have a sequence of equal investments λ. Further, during each period we compare average accumulation of characteristics using the logit model with tolerance parameter μ. This means that in (7.7) we take

$$\beta_{k+1} = \mu \cdot (k+1), \quad \lambda_k = \lambda, \quad k \geq 0.$$

Defining $\beta_0 = \mu$, we can estimate the right-hand side of inequality (7.8) as follows:

$$\frac{\mu}{\lambda} \ln m + \frac{\lambda L^2}{2\mu} \cdot \frac{2 + \ln k}{k+1}, \quad k \geq 1.$$

Thus, the quality of our property is stabilized on the level $\tau^* + \frac{\mu}{\lambda} \ln m$. In order to have convergence to the optimal balance, the inaccuracy level μ of audit testing must gradually go to zero.

7.2 Preliminary planning versus dynamic adjustment

Let us consider a multi-period optimization problem in the following form. For k consecutive periods, $k = 0, \ldots, N$, our expenses will be defined by different convex objective functions

$$f_0(x), \ldots, f_N(x), \quad x \in Q,$$

where Q is a closed convex set. Thus, if we choose to apply for the period k some strategy $x_k \in Q$, then the total expenses are given by

$$\Psi(X) = \sum_{k=0}^{N} f_k(x_k), \quad X = \{x_k\}_{k=0}^N.$$

In this model there are several possibilities for defining a rational strategy.

1. If all functions $\{f_k(x)\}_{k=0}^N$ are known in advance, then we can define an optimal *dynamic* strategy X^* as

$$x_k \; = \; x_k^* \; \stackrel{\text{def}}{=} \; \arg\min_x \left\{ f_k(x) : \; x \in Q \right\}, \quad k = 0, \ldots, N. \tag{7.10}$$

Clearly, the value of objective function Ψ at this sequence is as minimal as possible. However, this strategy has several drawbacks. First of all, it requires a considerable amount of preliminary computations since it is necessary to solve $(N+1)$ different optimization problems. The objective function of each problem must be known in advance. Finally, we have no control on the distance between two consecutive strategies x_k and x_{k+1}. If the distance is big, the required change may result in additional expenses which are not included in the model.

2. A less ambitious strategy consists in employing an optimal solution of the following problem:

$$x^* \; \stackrel{\text{def}}{=} \; \arg\min_x \left\{ \bar{f}_N(x) \stackrel{\text{def}}{=} \frac{1}{N+1} \sum_{k=0}^{N} f_k(x) : \; x \in Q \right\}. \tag{7.11}$$

At each period we apply the same variant $x_k = x^*$, $k = 0, \ldots, N$; we call such a sequence X_* the optimal *static* strategy. Clearly, the computation and implementation of this strategy is much easier than that of X^*. However, we still need to know all objective functions in advance.

3. The last possibility consists in launching an adjustment process which generates somehow a sequence X taking into account the observed information on the objective function of the current period. No objective functions are available in advance.

Clearly, we cannot expect too much from the last approach. Nevertheless, let us compare its efficiency with efficiency of sequence X_*. We assume that the sequence X is generated by the following variant of SDA-method (2.21).

Initialization: Set $s_0 = 0 \in E^*$. Choose $\gamma > 0$.
Iterations ($k = 0, \ldots, N$)**:** **1.** Compute $g_k \in \partial f_k(x_k)$ and set $s_{k+1} = s_k + g_k$. **2.** Choose $\beta_{k+1} = \gamma \hat{\beta}_{k+1}$ and set $x_{k+1} = \pi_{\beta_{k+1}}(-s_{k+1})$.

(7.12)

Let us assume that for any $k = 0, \ldots, N$ we have

$$\|g_k\|_* \leq L \quad \forall g_k \in \partial f_k(x), \ \forall x \in Q.$$

Then, in view of Theorem 2 we have the following bounds on the gap:

$$\tfrac{1}{N+1}\delta_N(D) \ \leq \ \tfrac{\hat{\beta}_{N+1}}{N+1}\left(\gamma d(x^*) + \tfrac{L}{2\sigma\gamma}\right).$$

On the other hand, for $D \geq d(x^*)$ we have

$$\begin{aligned}
\tfrac{1}{N+1}\delta_N(D) &= \tfrac{1}{N+1}\max_x \left\{ \sum_{k=0}^N \langle g_k, x_k - x\rangle : \ x \in \mathcal{F}_D \right\} \\
&\geq \tfrac{1}{N+1}\max_x \left\{ \sum_{k=0}^N [f_k(x_k) - f_k(x)]\rangle : \ x \in \mathcal{F}_D \right\} \\
&= \tfrac{1}{N+1}\sum_{k=0}^N f_k(x_k) - \min_x \{\bar{f}_N(x) : \ x \in Q\} \\
&= \tfrac{1}{N+1}\Psi(X) - \bar{f}_N(x^*).
\end{aligned}$$

Thus,

$$\tfrac{1}{N+1}\Psi(X) \ \leq \ \bar{f}_N(x^*) + \tfrac{\hat{\beta}_{N+1}}{N+1}\left(\gamma d(x^*) + \tfrac{L}{2\sigma\gamma}\right), \tag{7.13}$$

and we come to a rather intriguing conclusion:

For big N, the average efficiency of the dynamic adjustment process (7.12) becomes comparable with the average efficiency of the optimal static strategy computed by preliminary planning and based on full information on the future objective functions.

8 Discussion

Let us discuss the proposed schemes of dual averaging and compare them with traditional optimization methods.

1. Divergent series. There are many possibilities for defining the sequences $\{\lambda_k\}_{k=0}^{\infty}$ and $\{\beta_k\}_{k=0}^{\infty}$, which ensure an appropriate rate of decrease of $\frac{1}{S_k}\delta_k(D)$. The choice

$$\lambda_k = 1, \quad \beta_k = \gamma\hat{\beta}_k, \quad k \geq 0,$$

results in $\frac{1}{S_k}\delta_k(D) = O(\frac{1}{\sqrt{k}})$. This is the best possible rate of convergence of black-box methods as applied to nonsmooth minimization problems of unbounded dimension [8].

Let us check what can be achieved by another choices. Consider, for example,

$$\lambda_k = (k+1)^p, \quad \beta_k = \gamma(k+1)^q, \quad k \geq 0.$$

Note that for $\lambda > -1$ we have

$$\sum_{i=0}^{k}(i+1)^{\lambda} \;=\; O\left(\tfrac{1}{1+\lambda}(k+1)^{1+\lambda}\right).$$

In view of inequality (2.15), we need to compare the following objects:

$$\begin{aligned}
\beta_{k+1} &\approx (k+1)^q, \\
\sum_{i=0}^{k}\lambda_i &\approx \tfrac{1}{1+p}(k+1)^{1+p}, \\
\sum_{i=0}^{k}\tfrac{\lambda_i^2}{\beta_i} &\approx \tfrac{1}{1+2p-q}(k+1)^{1+2p-q}.
\end{aligned}$$

Thus, the rate of convergence of DA-methods is defined by

$$\max\left\{(1+p)(k+1)^{q-p-1}, \tfrac{1+p}{1+2p-q}(k+1)^{p-q}\right\}.$$

Hence, the optimal relation between the powers is $q - p = \frac{1}{2}$. Then we can bound the gap function by

$$O\left(\max\left\{1, \tfrac{1}{p+0.5}\right\}\tfrac{p+1}{\sqrt{k+1}}\right). \tag{8.1}$$

To conclude, for any $p > -\frac{1}{2}$ we can guarantee an optimal order of the rate of convergence. When $p \downarrow -\frac{1}{2}$, then the constant in (8.1) explodes. Nevertheless, for $p = -\frac{1}{2}$ we can guarantee the following rate:

$$O\left(\tfrac{1+\ln(k+1)}{\sqrt{k+1}}\right).$$

Note that with this variant of parameters, we get

$$\lambda_k = \frac{1}{\sqrt{k+1}}, \quad \beta_k = \gamma, \quad k \geq 0. \tag{8.2}$$

Therefore, the scheme (2.14) reduces to an optimal variant standard divergent-series scheme. In SDA-method (2.21) we choose

$$\lambda_k = 1, \quad \beta_k \approx \gamma\sqrt{k+1}, \quad k \geq 0.$$

Note that in the smoothing technique [10], which is based on much faster methods, the strategy for averaging of support functions is much more aggressive:

$$\lambda_k \approx k+1, \quad k \geq 0.$$

It is interesting that during many years the standard Convex Optimization theory recommended the *worst possible* choice of parameters (8.2) (see, for example, Section 3.2.3 [9]).

2. Bundle methods. The scheme (3.4) has a slight similarity with Bundle Method. However, the differences of these approaches are essential. Indeed,

- in Bundle Method the prox-center is moved after each "essential" iteration, and in (3.4) the center is fixed;
- in Bundle Method the scaling parameter μ_k is increasing for finding a point with better value of objective function. In method (3.4) this parameter is decreasing, which ensures a faster move of the minimization sequence;
- finally, the optimal worst-case complexity bound $O(\frac{1}{\epsilon^2})$ is not established yet for Bundle Method. For method (3.4) it follows from (3.3).

Acknowledgement. The author would like to thank F. Chudak and B. Rangarajan for their very useful suggestions.

References

[1] S. P. Andersen, A. de Palma, and J.-F. Thisse. Discrete choice theory of product differentiation. MIT Press, Cambridge, 1992.

[2] K. Anstreicher and L. Wolsey. On dual solutions in subgradient optimization. Unpublished manuscript (1993).

[3] F. Barahona and R. Anbil. The volume algorithm: Producing primal solutions with a subgradient method. *Mathematical Programming A*, 87 (2000) 385–399.

[4] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters* 31 (2003) 167–175.

[5] A. Ben-Tal, T. Margalit, and A. Nemirovski. The Ordered Subsets Mirror Descent Optimization Method with Applications to Tomography. *SIOPT* 12 (2001) #1 79–108.

[6] Yu. M. Ermoliev. Methods for Solving Nonlinear Extremal Problems. *Kibernetika* 4 (1966) 1-17.

[7] J.-L. Goffin. On convergence rates of subgradient optimization methods. *Mathematical Programming* 13 (1977) 329-347.

[8] A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. J. Wiley & Sons, 1983.

[9] Yu.Nesterov. *Introductory lectures on convex optimization. Basic course.* Kluwer, Boston, 2004.

[10] Yu. Nesterov. Smooth minimization of nonsmooth functions. CORE Discussion Paper #2003/12, CORE 2003. Published by *Mathematical Programming A*.

[11] Yu. Nesterov. Excessive gap technique in nonsmooth convex minimization. CORE Discussion Paper #2003/35, CORE 2003. Published by *SIOPT*.

[12] Yu. Nesterov. Dual extrapolation and its applications for solving variational inequalities and related problems. CORE Discussion Paper #2003/68, CORE 2003.

[13] Yu. Nesterov and J.-Ph. Vial. Confidence level solutions for stochastic programming. CORE Discussion Paper #2000/13, CORE 2000.

[14] B. T. Polyak. A general method of solving extremum problems. *Soviet Mathematics Doklady* 8 (1967) 593–597.

[15] N. Z. Shor. *Minimization Methods for Nondifferentiable Functions.* Springer-Verlag, Berlin, 1985.

9 Appendix: Strongly convex functions

Let Q be a closed convex set in E.

Definition 1 *A continuous function $d(x)$ is called* strongly convex *on Q if $Q \subseteq \operatorname{dom} d$ and there exists a constant $\sigma > 0$ such that for all $x, y \in Q$ and $\alpha \in [0,1]$ we have*

$$d(\alpha x + (1-\alpha)y) \leq \alpha d(x) + (1-\alpha)d(y) - \tfrac{1}{2}\sigma\alpha(1-\alpha)\|x-y\|^2. \tag{9.3}$$

We call σ the convexity parameter *of $d(\cdot)$.*

Consider the following minimization problem:

$$\min_{x \in Q} d(x). \tag{9.4}$$

Lemma 6 *If $d(\cdot)$ is strongly convex, then the problem (9.4) is solvable and its solution x^* is unique. Moreover, for any $x \in Q$ we have*

$$d(x) \geq d(x^*) + \tfrac{1}{2}\sigma\|x - x^*\|^2. \tag{9.5}$$

Proof:
Since $d(\cdot)$ is continuous on $Q \subseteq \operatorname{dom} d$, for existence of a solution of the problem (9.4) it is enough to show that the level sets of this function are bounded.

Consider the set $S = \{x \in Q : \ d(x) \leq M\} \neq \emptyset$. Assume that this set is not bounded. Then there exists an unbounded sequence $\{x_k\}_{k=0}^{\infty} \subset S$. Without loss of generality we can assume that

$$\|x_k - x_0\| \geq 1, \quad k \geq 1.$$

Define $\alpha_k = 1/\|x_k - x_0\| \in [0,1]$. Note that $\alpha_k \to 0$. Consider the sequence $\{y_k\}_{k=1}^{\infty} \subset Q$ defined as follows:

$$y_k = \alpha_k x_k + (1-\alpha_k)x_0.$$

Note that $\|y_k - x_0\| = 1$, $k \geq 1$. On the other hand, since $d(\cdot)$ is strongly convex, we have

$$\begin{aligned} d(y_k) &\leq \alpha_k d(x_k) + (1-\alpha_k)d(x_0) - \tfrac{1}{2}\sigma\alpha_k(1-\alpha_k)\|x_k - x_0\|^2 \\ &\leq M - \tfrac{1}{2}\sigma(1-\alpha_k)\|x_k - x_0\| \to -\infty. \end{aligned}$$

This contradicts to continuity of d.

In view of inequality (9.3), the solution of the problem (9.4) is unique. It remains to prove inequality (9.5). Indeed, for any $x \in Q$ and any $\alpha \in (0,1)$ we have

$$\begin{aligned} \alpha d(x) + (1-\alpha)d(x^*) &\geq d(\alpha x + (1-\alpha)x^*) + \tfrac{1}{2}\sigma\alpha(1-\alpha)\|x - x^*\|^2 \\ &\geq d(x^*) + \tfrac{1}{2}\sigma\alpha(1-\alpha)\|x - x^*\|^2. \end{aligned}$$

Thus, $d(x) \geq d(x^*) + \frac{1}{2}\sigma(1-\alpha)\|x - x^*\|^2$ and we get (9.5) as $\alpha \to 0$. □

The following sufficient condition of strong convexity is often useful.

Lemma 7 *Assume that function $d(\cdot)$ is twice differentiable on $\operatorname{rint} Q \subset \operatorname{dom} d$. If for some $\sigma > 0$ we have*

$$\langle \nabla^2 d(x)h, h\rangle \geq \sigma \|h\|^2, \quad \forall x \in \operatorname{int} Q,\ \forall h \in E,$$

then $d(\cdot)$ is strongly convex on Q with convexity parameter σ.

Proof:
Let u and v belong to $\operatorname{rint} Q$. Consider the univariate function

$$\phi(\tau) = d(u + \tau(v-u)) - d(u) - \tau\langle \nabla d(u), v-u\rangle, \quad \tau \in [0,1].$$

Note that $\phi(\tau)$ is convex, $\phi(0) = \phi'(0) = 0$, and

$$\phi''(\tau) = \langle \nabla^2 d(u+\tau(v-u))(v-u), v-u\rangle \geq \sigma\|v-u\|^2, \quad \tau \in [0,1].$$

Hence, $\phi(1) \geq \frac{\sigma}{2}\|v-u\|^2$. That is

$$d(v) \ \geq \ d(u) + \langle \nabla d(u), v-u\rangle + \tfrac{\sigma}{2}\|v-u\|^2, \quad u, v \in \operatorname{rint} Q.$$

Thus, for any $x, y \in \operatorname{rint} Q$ and $x_\alpha = \alpha x + (1-\alpha)y$, $\alpha \in [0,1]$, we have

$$\begin{aligned} d(x) &\geq d(x_\alpha) + \langle \nabla d(x_\alpha), x - x_\alpha\rangle + \tfrac{\sigma}{2}\|x - x_\alpha\|^2, \\ d(y) &\geq d(x_\alpha) + \langle \nabla d(x_\alpha), y - x_\alpha\rangle + \tfrac{\sigma}{2}\|y - x_\alpha\|^2. \end{aligned}$$

Adding these inequalities with coefficients α and $1-\alpha$, we get (9.3). Since $d(\cdot)$ is continuous, we extend this inequality to all x and y from Q. □