Mathematical Programming (2023) 199:305–341
https://doi.org/10.1007/s10107-022-01833-4

FULL LENGTH PAPER

Series A

# Subgradient ellipsoid method for nonsmooth convex problems

**Anton Rodomanov**[1] · **Yurii Nesterov**[2]

Received: 25 June 2021 / Accepted: 30 April 2022 / Published online: 14 June 2022


## Abstract

In this paper, we present a new ellipsoid-type algorithm for solving nonsmooth problems with convex structure. Examples of such problems include nonsmooth convex minimization problems, convex-concave saddle-point problems and variational inequalities with monotone operator. Our algorithm can be seen as a combination of the standard Subgradient and Ellipsoid methods. However, in contrast to the latter one, the proposed method has a reasonable convergence rate even when the dimensionality of the problem is sufficiently large. For generating accuracy certificates in our algorithm, we propose an efficient technique, which ameliorates the previously known recipes (Nemirovski in Math Oper Res 35(1):52–78, 2010).

**Keywords** Subgradient method · Ellipsoid method · Accuracy certificates · Separating oracle · Convex optimization · Nonsmooth optimization · Saddle-point problems · Variational inequalities

**Mathematics Subject Classification** 90C25 · 90C47 · 68Q25

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 788368).

✉ Anton Rodomanov
anton.rodomanov@uclouvain.be

Yurii Nesterov
yurii.nesterov@uclouvain.be

[1] Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Catholic University of Louvain (UCL), Louvain-la-Neuve, Belgium

[2] Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), Louvain-la-Neuve, Belgium

## 1 Introduction

The Ellipsoid Method is a classical algorithm in Convex Optimization. It was proposed in 1976 by Yudin and Nemirovski [23] as the modified method of centered cross-sections and then independently rediscovered a year later by Shor [21] in the form of the subgradient method with space dilation. However, the popularity came to the Ellipsoid Method only when Khachiyan used it in 1979 for proving his famous result on polynomial solvability of Linear Programming [10]. Shortly after, several polynomial algorithms, based on the Ellipsoid Method, were developed for some combinatorial optimization problems [9]. For more details and historical remarks on the Ellipsoid Method, see [2,3,14].

Despite its long history, the Ellipsoid Method still has some issues which have not been fully resolved or have been resolved only recently. One of them is the computation of accuracy certificates which is important for generating approximate solutions to dual problems or for solving general problems with convex structure (saddle-point problems, variational inequalities, etc.). For a long time, the procedure for calculating an accuracy certificate in the Ellipsoid Method required solving an auxiliary piecewise linear optimization problem (see, e.g., sect. 5 and 6 in [14]). Although this auxiliary computation did not use any additional calls to the oracle, it was still computationally expensive and, in some cases, could take even more time than the Ellipsoid Method itself. Only recently an efficient alternative has been proposed [16].

Another issue with the Ellipsoid Method is related to its poor dependency on the dimensionality of the problem. Consider, e.g., the minimization problem

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function and $Q := \{x \in \mathbb{R}^n : \|x\| \le R\}$ is the Euclidean ball of radius $R > 0$. The Ellipsoid Method for solving (1) can be written as follows (see, e.g., sect. 3.2.8 in [19]):

$$\begin{aligned} x_{k+1} &:= x_k - \frac{1}{n+1} \frac{W_k g_k}{\langle g_k, W_k g_k \rangle^{1/2}}, \\ W_{k+1} &:= \frac{n^2}{n^2-1} \left( W_k - \frac{2}{n+1} \frac{W_k g_k g_k^T W_k}{\langle g_k, W_k g_k \rangle} \right), \qquad k \ge 0, \end{aligned} \tag{2}$$

where $x_0 := 0$, $W_0 := R^2 I$ ($I$ is the identity matrix) and $g_k := f'(x_k)$ is an arbitrary nonzero subgradient if $x_k \in Q$, and $g_k$ is an arbitrary separator[1] of $x_k$ from $Q$ if $x_k \notin Q$.

To solve problem (1) with accuracy $\epsilon > 0$ (in terms of the function value), the Ellipsoid Method needs

$$O\left(n^2 \ln \frac{MR}{\epsilon}\right) \tag{3}$$

[1] More precisely, $g_k$ must be a non-zero vector such that $\langle g_k, x_k - x \rangle \ge 0$ for all $x \in Q$. In particular, for the Euclidean ball, one can take $g_k := x_k$.

iterations, where $M > 0$ is the Lipschitz constant of $f$ on $Q$ (see theorem 3.2.11 in [19]). Looking at this estimate, we can see an immediate drawback: it directly depends on the dimension and becomes useless when $n \to \infty$. In particular, we cannot guarantee any reasonable rate of convergence for the Ellipsoid Method when the dimensionality of the problem is sufficiently big.

Note that the aforementioned drawback is an artifact of the method itself, not its analysis. Indeed, when $n \to \infty$, iteration (2) reads

$$x_{k+1} := x_k, \quad W_{k+1} := W_k, \quad k \geq 0.$$

Thus, the method stays at the same point and does not make any progress.

On the other hand, the simplest Subgradient Method for solving (1) possesses the "dimension-independent" $O(M^2R^2/\epsilon^2)$ iteration complexity bound (see, e.g., sect. 3.2.3 in [19]). Comparing this estimate with (3), we see that the Ellipsoid Method is significantly faster than the Subgradient Method only when $n$ is not too big compared to $MR/\epsilon$ and significantly slower otherwise. Clearly, this situation is strange because the former algorithm does much more work at every iteration by "improving" the "metric" $W_k$ which is used for measuring the norm of the subgradients.

In this paper, we propose a new ellipsoid-type algorithm for solving nonsmooth problems with convex structure, which does not have the discussed above drawback. Our algorithm can be seen as a combination of the Subgradient and Ellipsoid methods and its convergence rate is basically as good as the best of the corresponding rates of these two methods (up to some logarithmic factors). In particular, when $n \to \infty$, the convergence rate of our algorithm coincides with that of the Subgradient Method.

### Contents

This paper is organized as follows. In Sect. 2.1, we review the general formulation of a problem with convex structure and the associated with it notions of *accuracy certificate* and *residual*. Our presentation mostly follows [16] with examples taken from [18]. Then, in Sect. 2.2, we introduce the notions of *accuracy semicertificate* and *gap* and discuss their relation with those of accuracy certificate and residual.

In Sect. 3, we present the general algorithmic scheme of our methods. To measure the convergence rate of this scheme, we introduce the notion of *sliding gap* and establish some preliminary bounds on it.

In Sect. 4, we discuss different choices of parameters in our general scheme. First, we show that, by setting some of the parameters to zero, we obtain the standard Subgradient and Ellipsoid methods. Then we consider a couple of other less trivial choices which lead to two new algorithms. The principal of these new algorithms is the latter one, which we call the *Subgradient Ellipsoid Method*. We demonstrate that the convergence rate of this algorithm is basically as good as the best of those of the Subgradient and Ellipsoid methods.

In Sect. 5, we show that, for both our new methods, it is possible to efficiently generate accuracy semicertificates whose gap is upper bounded by the sliding gap.

We also compare our approach with the recently proposed technique from [16] for building accuracy certificates for the standard Ellipsoid Method.

In Sect. 6, we discuss how to efficiently implement our general scheme and the procedure for generating accuracy semicertificates. In particular, we show that the time and memory requirements of our scheme are the same as in the standard Ellipsoid Method.

Finally, in Sect. 7, we discuss some open questions.

### Notation and generalities

In this paper, $\mathbb{E}$ denotes an arbitrary $n$-dimensional real vector space. Its dual space, composed of all linear functionals on $\mathbb{E}$, is denoted by $\mathbb{E}^*$. The value of $s \in \mathbb{E}^*$, evaluated at $x \in \mathbb{E}$, is denoted by $\langle s, x \rangle$. See [19, sect. 4.2.1] for the supporting discussion of abstract real vector spaces in Optimization.

Let us introduce in the spaces $\mathbb{E}$ and $\mathbb{E}^*$ a pair of conjugate Euclidean norms. To this end, let us fix a self-adjoint positive definite linear operator $B : \mathbb{E} \to \mathbb{E}^*$ and define

$$\|x\| := \langle Bx, x \rangle^{1/2}, \quad x \in \mathbb{E}, \qquad \|s\|_* := \langle s, B^{-1}s \rangle^{1/2}, \quad s \in \mathbb{E}^*.$$

Note that, for any $s \in \mathbb{E}^*$ and $x \in \mathbb{E}$, we have the Cauchy-Schwarz inequality

$$|\langle s, x \rangle| \leq \|s\|_* \|x\|,$$

which becomes an equality if and only if $s$ and $Bx$ are collinear. In addition to $\|x\|$ and $\|\cdot\|_*$, we often work with other Euclidean norms defined in the same way but using another reference operator instead of $B$. In this case, we write $\|\cdot\|_G$ and $\|\cdot\|_G^*$, where $G : \mathbb{E} \to \mathbb{E}^*$ is the corresponding self-adjoint positive definite linear operator.

Sometimes, in the formulas, involving products of linear operators, it is convenient to treat $x \in \mathbb{E}$ as a linear operator from $\mathbb{R}$ to $\mathbb{E}$, defined by $x\alpha := \alpha x$, and $x^*$ as a linear operator from $\mathbb{E}^*$ to $\mathbb{R}$, defined by $x^* s := \langle s, x \rangle$. Likewise, any $s \in \mathbb{E}^*$ can be treated as a linear operator from $\mathbb{R}$ to $\mathbb{E}^*$, defined by $s\alpha := \alpha s$, and $s^*$ as a linear operator from $\mathbb{E}$ to $\mathbb{R}$, defined by $s^* x := \langle s, x \rangle$. Then, $xx^*$ and $ss^*$ are rank-one self-adjoint linear operators from $\mathbb{E}^*$ to $\mathbb{E}$ and from $\mathbb{E}$ to $\mathbb{E}^*$ respectively, acting as follows: $(xx^*)s = \langle s, x \rangle x$ and $(ss^*)x = \langle s, x \rangle s$ for any $x \in \mathbb{E}$ and $s \in \mathbb{E}^*$.

For a self-adjoint linear operator $G : \mathbb{E} \to \mathbb{E}^*$, by $\operatorname{tr} G$ and $\det G$, we denote the trace and determinant of $G$ with respect to our fixed operator $B$:

$$\operatorname{tr} G := \operatorname{tr}(B^{-1}G), \qquad \det G := \det(B^{-1}G).$$

Note that, in these definitions, $B^{-1}G$ is a linear operator from $\mathbb{E}$ to $\mathbb{E}$, so $\operatorname{tr}(B^{-1}G)$ and $\det(B^{-1}G)$ are the standard well-defined notions of trace and determinant of a linear operator acting on the same space. For example, they can be defined as the trace and determinant of the matrix representation of $B^{-1}G$ with respect to an arbitrary chosen basis in $\mathbb{E}$ (the result is independent of the particular choice of basis). Alternatively, $\operatorname{tr} G$ and $\det G$ can be equivalently defined as the sum and product, respectively, of the eigenvalues of $G$ with respect to $B$.

For a point $x \in \mathbb{E}$ and a real $r > 0$, by

$$B(x, r) := \{y \in \mathbb{E} : \|x\| \le r\},$$

we denote the closed Euclidean ball with center $x$ and radius $r$.

Given two solids[2] $Q$, $Q_0 \subseteq \mathbb{E}$, we can define the *relative volume* of $Q$ with respect to $Q_0$ by $\mathrm{vol}(Q/Q_0) := \mathrm{vol}\, Q^e / \mathrm{vol}\, Q_0^e$, where $e$ is an arbitrary basis in $\mathbb{E}$, $Q^e$, $Q_0^e \subseteq \mathbb{R}^n$ are the coordinate representations of the sets $Q$, $Q_0$ in the basis $e$ and vol is the Lebesgue measure in $\mathbb{R}^n$. Note that the relative volume is independent of the particular choice of the basis $e$. Indeed, for any other basis $f$, we have $Q^e = T_f^e Q^f$, $Q_0^e = T_f^e Q_0^f$, where $T_f^e$ is the $n \times n$ change-of-basis matrix, so $\mathrm{vol}\, Q^e = (\det T_f^e)(\mathrm{vol}\, Q^f)$, $\mathrm{vol}\, Q_0^e = (\det T_f^e)(\mathrm{vol}\, Q_0^f)$ and hence $\mathrm{vol}\, Q^e / \mathrm{vol}\, Q_0^e = \mathrm{vol}\, Q^f / \mathrm{vol}\, Q_0^f$.

For us, it will be convenient to define the *volume* of a solid $Q \subseteq \mathbb{E}$ as the relative volume of $Q$ with respect to the unit ball:

$$\mathrm{vol}\, Q := \mathrm{vol}(Q/B(0, 1)).$$

For an ellipsoid $W := \{x \in \mathbb{E} : \langle Gx, x\rangle \le 1\}$, where $G : \mathbb{E} \to \mathbb{E}^*$ is a self-adjoint positive definite linear operator, we have $\mathrm{vol}\, W = (\det G)^{-1/2}$.

## 2 Convex problems and accuracy certificates

### 2.1 Description and examples

In this paper, we consider numerical algorithms for solving *problems with convex structure*. The main examples of such problems are convex minimization problems, convex-concave saddle-point problems, convex Nash equilibrium problems, and variational inequalities with monotone operators.

The general formulation of a problem with convex structure involves two objects:

– Solid $Q \subseteq \mathbb{E}$ (called the *feasible set*), represented by the *Separation Oracle*: given any point $x \in \mathbb{E}$, this oracle can check whether $x \in \mathrm{int}\, Q$, and if not, it reports a vector $g_Q(x) \in \mathbb{E}^* \setminus \{0\}$ which separates $x$ from $Q$:

$$\langle g_Q(x), x - y\rangle \ge 0, \quad \forall y \in Q. \tag{4}$$

– Vector field $g : \mathrm{int}\, Q \to \mathbb{E}^*$, represented by the *First-Order Oracle*: given any point $x \in \mathrm{int}\, Q$, this oracle returns the vector $g(x)$.

In what follows, we only consider the problems satisfying the following condition:

$$\exists x^* \in Q : \quad \langle g(x), x - x^*\rangle \ge 0, \quad \forall x \in \mathrm{int}\, Q. \tag{5}$$

[2] Hereinafter, a *solid* is any convex compact set with nonempty interior.

***Remark 1*** A careful reader may note that the notation $x^*$ overlaps with our general notation for the linear operator generated by a point $x$ (see Sect. 1). However, there should be no risk of confusion since the precise meaning of $x^*$ can usually be easily inferred from the context.

A numerical algorithm for solving a problem with convex structure starts at some point $x_0 \in \mathbb{E}$. At each step $k \geq 0$, it queries the oracles at the current *test point* $x_k$ to obtain the new information about the problem, and then somehow uses this new information to form the next test point $x_{k+1}$. Depending on whether $x_k \in \operatorname{int} Q$, the $k$th step of the algorithm is called *productive* or *nonproductive*.

The total information, obtained by the algorithm from the oracles after $k \geq 1$ steps, comprises its *execution protocol* which consists of:

- The test points $x_0, \ldots, x_{k-1} \in \mathbb{E}$.
- The set of productive steps $I_k := \{0 \leq i \leq k-1 : x_i \in \operatorname{int} Q\}$.
- The vectors $g_0, \ldots, g_{k-1} \in \mathbb{E}^*$ reported by the oracles: $g_i := g(x_i)$, if $i \in I_k$, and $g_i := g_Q(x_i)$, if $i \notin I_k$, $0 \leq i \leq k-1$.

An *accuracy certificate*, associated with the above execution protocol, is a nonnegative vector $\lambda := (\lambda_0, \ldots, \lambda_{k-1})$ such that $S_k(\lambda) := \sum_{i \in I_k} \lambda_i > 0$ (and, in particular, $I_k \neq \emptyset$). Given any solid $\Omega$, containing $Q$, we can define the following *residual* of $\lambda$ on $\Omega$:

$$\epsilon_k(\lambda) := \max_{x \in \Omega} \frac{1}{S_k(\lambda)} \sum_{i=0}^{k-1} \lambda_i \langle g_i, x_i - x \rangle, \tag{6}$$

which is easily computable whenever $\Omega$ is a simple set (e.g., a Euclidean ball). Note that

$$\epsilon_k(\lambda) \geq \max_{x \in Q} \frac{1}{S_k(\lambda)} \sum_{i=0}^{k-1} \lambda_i \langle g_i, x_i - x \rangle \geq \max_{x \in Q} \frac{1}{S_k(\lambda)} \sum_{i \in I_k} \lambda_i \langle g_i, x_i - x \rangle \tag{7}$$

and, in particular, $\epsilon_k(\lambda) \geq 0$ in view of (5).

In what follows, we will be interested in the algorithms, which can produce accuracy certificates $\lambda^{(k)}$ with $\epsilon_k(\lambda^{(k)}) \to 0$ at a certain rate. This is a meaningful goal because, for all known instances of problems with convex structure, the residual $\epsilon_k(\lambda)$ upper bounds a certain natural inaccuracy measure for the corresponding problem. Let us briefly review some standard examples (for more examples, see [16,18] and the references therein).

***Example 1*** **(Convex minimization problem)** Consider the problem

$$f^* := \min_{x \in Q} f(x), \tag{8}$$

where $Q \subseteq \mathbb{E}$ is a solid and $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ is closed convex and finite on $\operatorname{int} Q$.

The First-Order Oracle for (8) is $g(x) := f'(x)$, $x \in \operatorname{int} Q$, where $f'(x)$ is an arbitrary subgradient of $f$ at $x$. Clearly, (5) holds for $x^*$ being any solution of (8).

One can verify that, in this example, the residual $\epsilon_k(\lambda)$ upper bounds the functional residual: for $\hat{x}_k := \frac{1}{S_k(\lambda)} \sum_{i \in I_k} \lambda_i x_i$ or $x_k^* := \operatorname{argmin}\{f(x) : x \in X_k\}$, where $X_k := \{x_i : i \in I_k\}$, we have $f(\hat{x}_k) - f^* \le \epsilon_k(\lambda)$ and $f(x_k^*) - f^* \le \epsilon_k(\lambda)$.

Moreover, $\epsilon_k(\lambda)$, in fact, upper bounds the primal-dual gap for a certain dual problem for (8). Indeed, let $f_*: \mathbb{E}^* \to \mathbb{R} \cup \{+\infty\}$ be the conjugate function of $f$. Then, we can represent (8) in the following dual form:

$$f^* = \min_{x \in Q} \max_{s \in \operatorname{dom} f_*} [\langle s, x \rangle - f_*(s)] = \max_{s \in \operatorname{dom} f_*} [-f_*(s) - \xi_Q(-s)], \tag{9}$$

where $\operatorname{dom} f_* := \{s \in \mathbb{E}^* : f_*(s) < +\infty\}$ and $\xi_Q(-s) := \max_{x \in Q} \langle -s, x \rangle$. Denote $s_k := \frac{1}{S_k(\lambda)} \sum_{i \in I_k} \lambda_i g_i$. Then, using (7) and the convexity of $f$ and $f_*$, we obtain

$$\begin{aligned} \epsilon_k(\lambda) &\ge \frac{1}{S_k(\lambda)} \sum_{i \in I_k} \lambda_i \langle g_i, x_i \rangle + \xi_Q(-s_k) \\ &= \frac{1}{S_k(\lambda)} \sum_{i \in I_k} \lambda_i [f(x_i) + f_*(g_i)] + \xi_Q(-s_k) \\ &\ge f(\hat{x}_k) + f_*(s_k) + \xi_Q(-s_k). \end{aligned}$$

Thus, $\hat{x}_k$ and $s_k$ are $\epsilon_k(\lambda)$-approximate solutions (in terms of function value) to problems (8) and (9), respectively. Note that the same is true if we replace $\hat{x}_k$ with $x_k^*$.

***Example 2*** **(Convex-concave saddle-point problem)** Consider the following problem: Find $(u^*, v^*) \in U \times V$ such that

$$f(u^*, v) \le f(u^*, v^*) \le f(u, v^*), \quad \forall (u, v) \in U \times V, \tag{10}$$

where $U$, $V$ are solids in some finite-dimensional vector spaces $\mathbb{E}_u$, $\mathbb{E}_v$, respectively, and $f : U \times V \to \mathbb{R}$ is a continuous function which is *convex-concave*, i.e., $f(\cdot, v)$ is convex and $f(u, \cdot)$ is concave for any $u \in U$ and any $v \in V$.

In this example, we set $\mathbb{E} := \mathbb{E}_u \times \mathbb{E}_v$, $Q := U \times V$ and use the First-Order Oracle

$$g(x) := (f_u'(x), -f_v'(x)), \quad x := (u, v) \in \operatorname{int} Q,$$

where $f_u'(x)$ is an arbitrary subgradient of $f(\cdot, v)$ at $u$ and $f_v'(y)$ is an arbitrary supergradient of $f(u, \cdot)$ at $v$. Then, for any $x := (u, v) \in \operatorname{int} Q$ and any $x' := (u', v') \in Q$,

$$\langle g(x), x - x' \rangle = \langle f_u'(x), u - u' \rangle - \langle f_v'(x), v - v' \rangle \ge f(u, v') - f(u', v). \tag{11}$$

In particular, (5) holds for $x^* := (u^*, v^*)$ in view of (10).

Let $\phi\colon U \to \mathbb{R}$ and $\psi\colon V \to \mathbb{R}$ be the functions

$$\phi(u) := \max_{v \in V} f(u, v), \qquad \psi(v) := \min_{u \in U} f(u, v).$$

In view of (10), we have $\psi(v) \le f(u^*, v^*) \le \phi(u)$ for all $(u, v) \in U \times V$. Therefore, the difference $\phi(u) - \psi(v)$ (called the *primal-dual gap*) can be used for measuring the quality of an approximate solution $x := (u, v) \in Q$ to problem (10).

Denoting $\hat{x}_k := \frac{1}{S_k(\lambda)} \sum_{i \in I_k} \lambda_i x_i =: (\hat{u}_k, \hat{v}_k)$ and using (7), we obtain

$$\begin{aligned}
\epsilon_k(\lambda) &\ge \max_{x \in Q} \frac{1}{S_k(\lambda)} \sum_{i \in I_k} \lambda_i \langle g_i, x_i - x \rangle \\
&\ge \max_{u \in U, v \in V} \frac{1}{S_k(\lambda)} \sum_{i \in I_k} \lambda_i [f(u_i, v) - f(u, v_i)] \\
&\ge \max_{u \in U, v \in V} [f(\hat{u}_k, v) - f(u, \hat{v}_k)] = \phi(\hat{u}_k) - \psi(\hat{v}_k),
\end{aligned}$$

where the second inequality is due to (11) and the last one follows from the convexity-concavity of $f$. Thus, the residual $\epsilon_k(\lambda)$ upper bounds the primal-dual gap for the approximate solution $\hat{x}_k$.

***Example 3*** **(Variational inequality with monotone operator)** Let $Q \subseteq \mathbb{E}$ be a solid and let $V\colon Q \to \mathbb{E}^*$ be a continuous operator which is *monotone*, i.e., $\langle V(x) - V(y), x - y \rangle \ge 0$ for all $x, y \in Q$. The goal is to solve the following (weak) *variational inequality*:

$$\text{Find } x^* \in Q\colon \quad \langle V(x), x - x^* \rangle \ge 0, \quad \forall x \in Q. \tag{12}$$

Since $V$ is continuous, this problem is equivalent to its strong variant: find $x^* \in Q$ such that $\langle V(x^*), x - x^* \rangle \ge 0$ for all $x \in Q$.

A standard tool for measuring the quality of an approximate solution to (12) is the *dual gap function*, introduced in [1]:

$$f(x) := \max_{y \in Q} \langle V(y), x - y \rangle, \qquad x \in Q.$$

It is easy to see that $f$ is a convex nonnegative function which equals 0 exactly at the solutions of (12).

In this example, the First-Order Oracle is defined by $g(x) := V(x)$ for any $x \in \operatorname{int} Q$. Denote $\hat{x}_k := \frac{1}{S_k(\lambda)} \sum_{i \in I_k} \lambda_i x_i$. Then, using (7) and the monotonicity of $V$, we obtain

$$\begin{aligned}
\epsilon_k(\lambda) &\ge \max_{x \in Q} \frac{1}{S_k(\lambda)} \sum_{i \in I_k} \lambda_i \langle V(x_i), x_i - x \rangle \\
&\ge \max_{x \in Q} \frac{1}{S_k(\lambda)} \sum_{i \in I_k} \lambda_i \langle V(x), x_i - x \rangle = f(\hat{x}_k).
\end{aligned}$$

Thus, $\epsilon_k(\lambda)$ upper bounds the dual gap function for the approximate solution $\hat{x}_k$.

### 2.2 Establishing convergence of residual

For the algorithms, considered in this paper, instead of accuracy certificates and residuals, it turns out to be more convenient to speak about closely related notions of *accuracy semicertificates* and *gaps*, which we now introduce.

As before, let $x_0, \ldots, x_{k-1}$ be the test points, generated by the algorithm after $k \geq 1$ steps, and let $g_0, \ldots, g_{k-1}$ be the corresponding oracle outputs. An *accuracy semicertificate*, associated with this information, is a nonnegative vector $\lambda := (\lambda_0, \ldots, \lambda_{k-1})$ such that $\Gamma_k(\lambda) := \sum_{i=0}^{k-1} \lambda_i \|g_i\|_* > 0$. Given any solid $\Omega$, containing $Q$, the *gap* of $\lambda$ on $\Omega$ is defined in the following way:

$$\delta_k(\lambda) := \max_{x \in \Omega} \frac{1}{\Gamma_k(\lambda)} \sum_{i=0}^{k-1} \lambda_i \langle g_i, x_i - x \rangle. \tag{13}$$

Comparing these definitions with those of accuracy certificate and residual, we see that the only difference between them is that now we use a different "normalizing" coefficient: $\Gamma_k(\lambda)$ instead of $S_k(\lambda)$. Also, in the definitions of semicertificate and gap, we do not make any distinction between productive and nonproductive steps. Note that $\delta_k(\lambda) \geq 0$.

Let us demonstrate that by making the gap sufficiently small, we can make the corresponding residual sufficiently small as well. For this, we need the following standard assumption about our problem with convex structure (see, e.g., [16]).

**Assumption 1** The vector field $g$, reported by the First-Order Oracle, is semibounded:

$$\langle g(x), y - x \rangle \leq V, \quad \forall x \in \operatorname{int} Q, \ \forall y \in Q.$$

A classical example of a semibounded field is a bounded one: if there is $M \geq 0$, such that $\|g(x)\|_* \leq M$ for all $x \in \operatorname{int} Q$, then $g$ is semibounded with $V := MD$, where $D$ is the diameter of $Q$. However, there exist other examples. For instance, if $g$ is the subgradient field of a convex function $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$, which is finite and continuous on $Q$, then $g$ is semibounded with $V := \max_Q f - \min_Q f$ (variation of $f$ on $Q$); however, $g$ is not bounded if $f$ is not Lipschitz continuous (e.g., $f(x) := -\sqrt{x}$ on $Q := [0, 1]$). Another interesting example is the subgradient field $g$ of a $\nu$-self-concordant barrier $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ for the set $Q$; in this case, $g$ is semibounded with $V := \nu$ (see, e.g., [19, Theorem 5.3.7]), while $f(x) \to +\infty$ at the boundary of $Q$.

**Lemma 1** *Let $\lambda$ be a semicertificate such that $\delta_k(\lambda) < r$, where $r$ is the largest of the radii of Euclidean balls contained in $Q$. Then, $\lambda$ is a certificate and*

$$\epsilon_k(\lambda) \leq \frac{\delta_k(\lambda)}{r - \delta_k(\lambda)} V.$$

***Proof*** Denote $\delta_k := \delta_k(\lambda)$, $\Gamma_k := \Gamma_k(\lambda)$, $S_k := S_k(\lambda)$. Let $\bar{x} \in Q$ be such that $B(\bar{x}, r) \subseteq Q$. For each $0 \le i \le k-1$, let $z_i$ be a maximizer of $z \mapsto \langle g_i, z - \bar{x} \rangle$ on $B(\bar{x}, r)$. Then, for any $0 \le i \le k-1$, we have $\langle g_i, \bar{x} - x_i \rangle = \langle g_i, z_i - x_i \rangle - r\|g_i\|_*$ with $z_i \in Q$. Therefore,

$$\sum_{i=0}^{k-1} \lambda_i \langle g_i, \bar{x} - x_i \rangle = \sum_{i=0}^{k-1} \lambda_i \langle g_i, z_i - x_i \rangle - r\Gamma_k \le S_k V - r\Gamma_k, \tag{14}$$

where the inequality follows from the separation property (4) and Assumption 1.

Let $x \in \Omega$ be arbitrary. Denoting $y := \big(\delta_k \bar{x} + (r - \delta_k)x\big)/r \in \Omega$, we obtain

$$\begin{aligned}(r - \delta_k) \sum_{i=0}^{k-1} \lambda_i \langle g_i, x_i - x \rangle &= r \sum_{i=0}^{k-1} \lambda_i \langle g_i, x_i - y \rangle + \delta_k \sum_{i=0}^{k-1} \lambda_i \langle g_i, \bar{x} - x_i \rangle \\ &\le r\delta_k\Gamma_k + \delta_k \sum_{i=0}^{k-1} \lambda_i \langle g_i, \bar{x} - x_i \rangle \le \delta_k S_k V, \end{aligned} \tag{15}$$

where the inequalities follow from the definition (13) of $\delta_k$ and (14), respectively.

It remains to show that $\lambda$ is a certificate, i.e., $S_k > 0$. But this is simple. Indeed, if $S_k = 0$, then, taking $x := \bar{x}$ in (15) and using (14), we get $0 \ge \sum_{i=0}^{k-1} \lambda_i \langle g_i, x_i - \bar{x} \rangle \ge r\Gamma_k$, which contradicts our assumption that $\lambda$ is a semicertificate, i.e., $\Gamma_k > 0$. □

According to Lemma 1, from the convergence rate of the gap $\delta_k(\lambda^{(k)})$ to zero, we can easily obtain the corresponding convergence rate of the residual $\epsilon_k(\lambda^{(k)})$. In particular, to ensure that $\epsilon_k(\lambda^{(k)}) \le \epsilon$ for some $\epsilon > 0$, it suffices to make $\delta_k(\lambda^{(k)}) \le \delta(\epsilon) := \epsilon r/(\epsilon + V)$. For this reason, in the rest of this paper, we can focus our attention on studying the convergence rate only for the gap.

## 3 General algorithmic scheme

Consider the general scheme presented in Algorithm 1. This scheme works with an arbitrary oracle $\mathcal{G}\colon \mathbb{E} \to \mathbb{E}^*$ satisfying the following condition:

$$\exists x^* \in B(x_0, R): \quad \langle \mathcal{G}(x), x - x^* \rangle \ge 0, \quad \forall x \in \mathbb{E}. \tag{16}$$

The point $x^*$ from (16) is typically called a *solution* of our problem. For the general problem with convex structure, represented by the First-Order Oracle $g$ and the Separation Oracle $g_Q$ for the solid $Q$, the oracle $\mathcal{G}$ is usually defined as follows: $\mathcal{G}(x) := g(x)$, if $x \in \operatorname{int} Q$, and $\mathcal{G}(x) := g_Q(x)$, otherwise. To ensure that (16) holds, the constant $R$ needs to be chosen sufficiently big so that $Q \subseteq B(x_0, R)$.

**Algorithm 1: General Scheme of Subgradient Ellipsoid Method**

**Input:** Point $x_0 \in \mathbb{E}$ and scalar $R > 0$.

**Initialization:** Define the functions $\ell_0(x) := 0$, $\omega_0(x) := \frac{1}{2}\|x\|^2$.

**For** $k \geq 0$ **iterate:**

1. Query the oracle to obtain $g_k := \mathcal{G}(x_k)$.
2. Compute $U_k := \max_{x \in \Omega_k \cap L_k^-} \langle g_k, x_k - x \rangle$, where

$$\Omega_k := \{x \in \mathbb{E} : \omega_k(x) \leq \tfrac{1}{2}R^2\}, \qquad L_k^- := \{x \in \mathbb{E} : \ell_k(x) \leq 0\}.$$

3. Choose some coefficients $a_k, b_k \geq 0$ and update the functions

$$\begin{aligned} \ell_{k+1}(x) &:= \ell_k(x) + a_k \langle g_k, x - x_k \rangle, \\ \omega_{k+1}(x) &:= \omega_k(x) + \tfrac{1}{2} b_k (U_k - \langle g_k, x_k - x \rangle) \langle g_k, x - x_k \rangle. \end{aligned} \tag{17}$$

4. Set $x_{k+1} := \operatorname{argmin}_{x \in \mathbb{E}} [\ell_{k+1}(x) + \omega_{k+1}(x)]$.

Note that, in Algorithm 1, $\omega_k$ are strictly convex quadratic functions and $\ell_k$ are affine functions. Therefore, the sets $\Omega_k$ are certain ellipsoids and $L_k^-$ are certain halfspaces (possibly degenerate).

Let us show that Algorithm 1 is a cutting-plane scheme in which the sets $\Omega_k \cap L_k^-$ are the localizers of the solution $x^*$.

**Lemma 2** *In Algorithm 1, for all $k \geq 0$, we have $x^* \in \Omega_k \cap L_k^-$ and $\hat{Q}_{k+1} \subseteq \Omega_{k+1} \cap L_{k+1}^-$, where $\hat{Q}_{k+1} := \{x \in \Omega_k \cap L_k^- : \langle g_k, x - x_k \rangle \leq 0\}$.*

***Proof*** Let us prove the claim by induction. Clearly, $\Omega_0 = B(x_0, R)$, $L_0^- = \mathbb{E}$, hence $\Omega_0 \cap L_0^- = B(x_0, R) \ni x^*$ by (16). Suppose we have already proved that $x^* \in \Omega_k \cap L_k^-$ for some $k \geq 0$. Combining this with (16), we obtain $x^* \in \hat{Q}_{k+1}$, so it remains to show that $\hat{Q}_{k+1} \subseteq \Omega_{k+1} \cap L_{k+1}^-$. Let $x \in \hat{Q}_{k+1}$ ($\subseteq \Omega_k \cap L_k^-$) be arbitrary. Note that $0 \leq \langle g_k, x_k - x \rangle \leq U_k$. Hence, by (17), $\ell_{k+1}(x) \leq \ell_k(x) \leq 0$ and $\omega_{k+1}(x) \leq \omega_k(x) \leq \frac{1}{2}R^2$, which means that $x \in \Omega_{k+1} \cap L_{k+1}^-$. □

Next, let us establish an important representation of the ellipsoids $\Omega_k$ via the functions $\ell_k$ and the test points $x_k$. For this, let us define $G_k := \nabla^2 \omega_k(0)$ for each $k \geq 0$. Observe that these operators satisfy the following simple relations (cf. (17)):

$$G_0 = B, \qquad G_{k+1} = G_k + b_k g_k g_k^*, \quad k \geq 0. \tag{18}$$

Also, let us define the sequence $R_k > 0$ by the recurrence

$$R_0 = R, \qquad R_{k+1}^2 = R_k^2 + (a_k + \tfrac{1}{2} b_k U_k)^2 \frac{\|g_k\|_{G_k}^2}{1 + b_k \|g_k\|_{G_k}^2}, \quad k \geq 0. \tag{19}$$

**Lemma 3** *In Algorithm 1, for all $k \geq 0$, we have*

$$\Omega_k = \{x \in \mathbb{E} : -\ell_k(x) + \tfrac{1}{2}\|x - x_k\|^2_{G_k} \leq \tfrac{1}{2}R_k^2\}.$$

*In particular, for all $k \geq 0$ and all $x \in \Omega_k \cap L_k^-$, we have $\|x - x_k\|_{G_k} \leq R_k$.*

***Proof*** Let $\psi_k \colon \mathbb{E} \to \mathbb{R}$ be the function $\psi_k(x) := \ell_k(x) + \omega_k(x)$. Note that $\psi_k$ is a quadratic function with Hessian $G_k$ and minimizer $x_k$. Hence, for any $x \in \mathbb{E}$, we have

$$\psi_k(x) = \psi_k^* + \tfrac{1}{2}\|x - x_k\|^2_{G_k}, \tag{20}$$

where $\psi_k^* := \min_{x \in \mathbb{E}} \psi_k(x)$.

Let us compute $\psi_k^*$. Combining (17), (18) and (20), for any $x \in \mathbb{E}$, we obtain

$$\begin{aligned}
\psi_{k+1}(x) &= \psi_k(x) + (a_k + \tfrac{1}{2}b_k U_k)\langle g_k, x - x_k\rangle + \tfrac{1}{2}b_k\langle g_k, x - x_k\rangle^2 \\
&= \psi_k^* + \tfrac{1}{2}\|x - x_k\|^2_{G_k} + (a_k + \tfrac{1}{2}b_k U_k)\langle g_k, x - x_k\rangle + \tfrac{1}{2}b_k\langle g_k, x - x_k\rangle^2 \\
&= \psi_k^* + \tfrac{1}{2}\|x - x_k\|^2_{G_{k+1}} + (a_k + \tfrac{1}{2}b_k U_k)\langle g_k, x - x_k\rangle,
\end{aligned} \tag{21}$$

Therefore,

$$\begin{aligned}
\psi_{k+1}^* &= \psi_k^* - \tfrac{1}{2}(a_k + \tfrac{1}{2}b_k U_k)^2\|g_k\|^2_{G_{k+1}} \\
&= \psi_k^* - \tfrac{1}{2}(a_k + \tfrac{1}{2}b_k U_k)^2 \frac{\|g_k\|^2_{G_k}}{1 + b_k\|g_k\|^2_{G_k}},
\end{aligned} \tag{22}$$

where the last identity follows from the fact that $G_{k+1}^{-1}g_k = G_k^{-1}g_k/(1 + b_k\|g_k\|^2_{G_k})$ (since $G_{k+1}G_k^{-1}g_k = (1 + b_k\|g_k\|^2_{G_k})g_k$ in view of (18)). Since (22) is true for any $k \geq 0$ and since $\psi_0^* = 0$, we thus obtain, in view of (19),

$$\psi_k^* = \tfrac{1}{2}(R^2 - R_k^2). \tag{23}$$

Let $x \in \Omega_k$ be arbitrary. Using the definition of $\psi_k(x)$ and (23), we obtain

$$-\ell_k(x) + \tfrac{1}{2}\|x - x_k\|^2_{G_k} = \omega_k(x) - \psi_k^* = \omega_k(x) + \tfrac{1}{2}(R_k^2 - R^2).$$

Thus, $x \in \Omega_k \iff \omega_k(x) \leq \tfrac{1}{2}R^2 \iff -\ell_k(x) + \tfrac{1}{2}\|x - x_k\|^2_{G_k} \leq \tfrac{1}{2}R_k^2$. In particular, for any $x \in \Omega_k \cap L_k^-$, we have $\ell_k(x) \leq 0$ and $\|x - x_k\|_{G_k} \leq R_k$. □

Lemma 3 has several consequences. First, we see that the localizers $\Omega_k \cap L_k^-$ are contained in the ellipsoids $\{x : \|x - x_k\|_{G_k} \leq R_k\}$ whose centers are the test points $x_k$.

Second, we get a uniform upper bound on the function $-\ell_k$ on the ellipsoid $\Omega_k$: $-\ell_k(x) \leq \tfrac{1}{2}R_k^2$ for all $x \in \Omega_k$. This observation leads us to the following definition

of the *sliding gap*:

$$\Delta_k := \max_{x \in \Omega_k} \frac{1}{\Gamma_k}[-\ell_k(x)] = \max_{x \in \Omega_k} \frac{1}{\Gamma_k} \sum_{i=0}^{k-1} a_i \langle g_i, x_i - x \rangle, \quad k \geq 1, \tag{24}$$

provided that $\Gamma_k := \sum_{i=0}^{k-1} a_i \|g_i\|_* > 0$. According to our observation, we have

$$\Delta_k \leq \frac{R_k^2}{2\Gamma_k}. \tag{25}$$

At the same time, $\Delta_k \geq 0$ in view of Lemma 2 and 16

Comparing the definition (24) of the sliding gap $\Delta_k$ with the definition (13) of the gap $\delta_k(a^{(k)})$ for the semicertificate $a^{(k)} := (a_0, \ldots, a_{k-1})$, we see that they are almost identical. The only difference between them is that the solid $\Omega_k$, over which the maximum is taken in the definition of the sliding gap, depends on the iteration counter $k$. This seems to be unfortunate because we cannot guarantee that *each* $\Omega_k$ contains the feasible set $Q$ (as required in the definition of gap) even if so does the initial solid $\Omega_0 = B(x_0, R)$. However, this problem can be dealt with. Namely, in Sect. 5, we will show that the semicertificate $a^{(k)}$ can be efficiently converted into another semicertificate $\lambda^{(k)}$ for which $\delta_k(\lambda^{(k)}) \leq \Delta_k$ when taken over the initial solid $\Omega := \Omega_0$. Thus, the sliding gap $\Delta_k$ is a meaningful measure of convergence rate of Algorithm 1, and it makes sense to call the coefficients $a^{(k)}$ a *preliminary semicertificate*.

Let us now demonstrate that, for a suitable choice of the coefficients $a_k$ and $b_k$ in Algorithm 1, we can ensure that the sliding gap $\Delta_k$ converges to zero.

***Remark 2*** From now on, in order to avoid taking into account some trivial degenerate cases, it will be convenient to make the following minor technical assumption:

$$\text{In Algorithm 1, } g_k \neq 0 \text{ for all } k \geq 0.$$

Indeed, when the oracle reports $g_k = 0$ for some $k \geq 0$, it usually means that the test point $x_k$, at which the oracle was queried, is, in fact, an exact solution to our problem. For example, if the standard oracle for a problem with convex structure has reported $g_k = 0$, we can terminate the method and return the certificate $\lambda := (0, \ldots, 0, 1)$ for which the residual $\epsilon_k(\lambda) = 0$.

Let us choose the coefficients $a_k$ and $b_k$ in the following way:

$$a_k := \frac{\alpha_k R + \frac{1}{2}\theta\gamma R_k}{\|g_k\|_{G_k}^*}, \qquad b_k := \frac{\gamma}{\|g_k\|_{G_k}^2}, \qquad k \geq 0, \tag{26}$$

where $\alpha_k, \theta, \gamma \geq 0$ are certain coefficients to be chosen later.

According to (25), to estimate the convergence rate of the sliding gap, we need to estimate the rate of growth of the coefficients $R_k$ and $\Gamma_k$ from above and below, respectively. Let us do this.

**Lemma 4** *In Algorithm 1 with parameters (26), for all $k \geq 0$, we have*

$$R_k^2 \leq [q_c(\gamma)]^k C_k R^2, \tag{27}$$

*where $q_c(\gamma) := 1 + \frac{c\gamma^2}{2(1+\gamma)}$, $c := \frac{1}{2}(\tau+1)(\theta+1)^2$, $C_k := 1 + \frac{\tau+1}{\tau}\sum_{i=0}^{k-1}\alpha_i^2$ and $\tau > 0$ can be chosen arbitrarily. Moreover, if $\alpha_k = 0$ for all $k \geq 0$, then, $R_k^2 = [q_c(\gamma)]^k R^2$ for all $k \geq 0$ with $c := \frac{1}{2}(\theta+1)^2$.*

***Proof*** By the definition of $U_k$ and Lemma 3, we have

$$U_k = \max_{x \in \Omega_k \cap L_k^-} \langle g_k, x_k - x\rangle \leq \max_{\|x - x_k\|_{G_k} \leq R_k} \langle g_k, x_k - x\rangle = R_k \|g_k\|_{G_k}^*. \tag{28}$$

At the same time, $U_k \geq 0$ in view of Lemma 2 and (16). Hence,

$$\begin{aligned}(a_k + \tfrac{1}{2}b_k U_k)^2 \frac{\|g_k\|_{G_k}^2}{1 + b_k\|g_k\|_{G_k}^2} &\leq (a_k + \tfrac{1}{2}b_k R_k \|g_k\|_{G_k}^*)^2 \frac{\|g_k\|_{G_k}^2}{1 + b_k\|g_k\|_{G_k}^2} \\ &= \frac{1}{1+\gamma}\big(\alpha_k R + \tfrac{1}{2}(\theta+1)\gamma R_k\big)^2,\end{aligned}$$

where the identity follows from (26). Combining this with (19), we obtain

$$R_{k+1}^2 \leq R_k^2 + \frac{1}{1+\gamma}\big(\alpha_k R + \tfrac{1}{2}(\theta+1)\gamma R_k\big)^2. \tag{29}$$

Note that, for any $\xi_1, \xi_2 \geq 0$ and any $\tau > 0$, we have

$$(\xi_1 + \xi_2)^2 = \xi_1^2 + 2\xi_1\xi_2 + \xi_2^2 \leq \frac{\tau+1}{\tau}\xi_1^2 + (\tau+1)\xi_2^2 = (\tau+1)\Big(\frac{1}{\tau}\xi_1^2 + \xi_2^2\Big)$$

(look at the minimum of the right-hand side in $\tau$). Therefore, for arbitrary $\tau > 0$,

$$R_{k+1}^2 \leq R_k^2 + \frac{\tau+1}{1+\gamma}\Big(\frac{1}{\tau}\alpha_k^2 R^2 + \tfrac{1}{4}(\theta+1)^2\gamma^2 R_k^2\Big) = q R_k^2 + \beta_k R^2,$$

where we denote $q := q_c(\gamma) \geq 1$ and $\beta_k := \frac{\tau+1}{\tau(1+\gamma)}\alpha_k^2$. Dividing both sides by $q^{k+1}$, we get

$$\frac{R_{k+1}^2}{q^{k+1}} \leq \frac{R_k^2}{q^k} + \frac{\beta_k R^2}{q^{k+1}}.$$

Since this is true for any $k \geq 0$, we thus obtain, in view of (19), that

$$\frac{R_k^2}{q^k} \leq \frac{R_0^2}{q^0} + R^2 \sum_{i=0}^{k-1} \frac{\beta_i}{q^{i+1}} = \Big(1 + \sum_{i=0}^{k-1}\frac{\beta_i}{q^{i+1}}\Big)R^2,$$

Multiplying both sides by $q^k$ and using that $\frac{\beta_i}{q^{i+1}} \le \frac{\tau+1}{\tau}\alpha_i^2$, we come to (27).

When $\alpha_k = 0$ for all $k \ge 0$, we have $\ell_k = 0$ and $L_k^- = \mathbb{E}$ for all $k \ge 0$. Therefore, by Lemma 3, $\Omega_k = \{x : \|x - x_k\|_{G_k} \le R_k\}$ and hence (28) is, in fact, an equality. Consequently, (29) becomes $R_{k+1}^2 = R_k^2 + \frac{c\gamma^2}{2(1+\gamma)}R_k^2 = q_c(\gamma)R_k^2$, where $c := \frac{1}{2}(\theta+1)^2$. □

***Remark 3*** From the proof, one can see that the quantity $C_k$ in Lemma 4 can be improved up to $C_k' := 1 + \frac{\tau+1}{\tau(1+\gamma)}\sum_{i=0}^{k-1}\frac{\alpha_i^2}{[q_c(\gamma)]^{i+1}}$.

**Lemma 5** *In Algorithm 1 with parameters (26), for all $k \ge 1$, we have*

$$\Gamma_k \ge R\Big(\sum_{i=0}^{k-1}\alpha_i + \tfrac{1}{2}\theta\sqrt{\gamma n\big[(1+\gamma)^{k/n} - 1\big]}\Big). \tag{30}$$

***Proof*** By the definition of $\Gamma_k$ and (26), we have

$$\Gamma_k = \sum_{i=0}^{k-1} a_i\|g_i\|_* = R\sum_{i=0}^{k-1}\alpha_i\rho_i + \tfrac{1}{2}\theta\gamma\sum_{i=0}^{k-1}R_i\rho_i,$$

where $\rho_i := \|g_i\|_*/\|g_i\|_{G_i}^*$. Let us estimate each sum from below separately.

For the first sum, we can use the trivial bound $\rho_i \ge 1$, which is valid for any $i \ge 0$ (since $G_i \ge B$ in view of (18)). This gives us $\sum_{i=0}^{k-1}\alpha_i\rho_i \ge \sum_{i=0}^{k-1}\alpha_i$.

Let us estimate the second sum. According to (19), for any $i \ge 0$, we have $R_i \ge R$. Hence, $\sum_{i=0}^{k-1}R_i\rho_i \ge R\sum_{i=0}^{k-1}\rho_i \ge R\left(\sum_{i=0}^{k-1}\rho_i^2\right)^{1/2}$ and it remains to lower bound $\sum_{i=0}^{k-1}\rho_i^2$. By 18 and 26, $G_0 = B$ and $G_{i+1} = G_i + \gamma g_i g_i^*/\|g_i\|_{G_i}^2$ for all $i \ge 0$. Therefore,

$$\begin{aligned}\sum_{i=0}^{k-1}\rho_i^2 &= \frac{1}{\gamma}\sum_{i=0}^{k-1}(\operatorname{tr}G_{i+1} - \operatorname{tr}G_i) = \frac{1}{\gamma}(\operatorname{tr}G_k - \operatorname{tr}B) = \frac{1}{\gamma}(\operatorname{tr}G_k - n)\\ &\ge \frac{n}{\gamma}\big[(\det G_k)^{1/n} - 1\big] = \frac{n}{\gamma}\big[(1+\gamma)^{k/n} - 1\big],\end{aligned}$$

where we have applied the arithmetic-geometric mean inequality. Combining the obtained estimates, we get (30). □

## 4 Main instances of general scheme

Let us now consider several possibilities for choosing the coefficients $\alpha_k$, $\theta$ and $\gamma$ in (26).

### 4.1 Subgradient method

The simplest possibility is to choose

$$\alpha_k > 0, \qquad \theta := 0, \qquad \gamma := 0.$$

In this case, $b_k = 0$ for all $k \geq 0$, so $G_k = B$ and $\omega_k(x) = \omega_0(x) = \frac{1}{2}\|x\|^2$ for all $x \in \mathbb{E}$ and all $k \geq 0$ (see (17) and (18)). Consequently, the new test points $x_{k+1}$ in Algorithm 1 are generated according to the following rule:

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{E}} \Big[ \sum_{i=0}^{k} a_i \langle g_i, x - x_i \rangle + \tfrac{1}{2}\|x\|^2 \Big],$$

where $a_i = \alpha_i R / \|g_i\|_*$. Thus, Algorithm 1 is the Subgradient Method: $x_{k+1} = x_k - a_k g_k$.

In this example, each ellipsoid $\Omega_k$ is simply a ball: $\Omega_k = B(x_0, R)$ for all $k \geq 0$. Hence, the sliding gap $\Delta_k$, defined in (24), does not "slide" and coincides with the gap of the semicertificate $a := (a_0, \ldots, a_{k-1})$ on the solid $B(x_0, R)$:

$$\Delta_k = \max_{x \in B(x_0, R)} \frac{1}{\Gamma_k} \sum_{i=0}^{k-1} a_i \langle g_i, x_i - x \rangle.$$

In view of Lemmas 4 and 5, for all $k \geq 1$, we have

$$R_k^2 \leq \Big(1 + \sum_{i=0}^{k-1} \alpha_i^2\Big) R^2, \qquad \Gamma_k \geq R \sum_{i=0}^{k-1} \alpha_i$$

(tend $\tau \to +\infty$ in Lemma 4). Substituting these estimates into (25), we obtain the following well-known estimate for the gap in the Subgradient Method:

$$\Delta_k \leq \frac{1 + \sum_{i=0}^{k-1} \alpha_i^2}{2 \sum_{i=0}^{k-1} \alpha_i} R.$$

The standard strategies for choosing the coefficients $\alpha_i$ are as follows (see, e.g., sect. 3.2.3 in [19]):

1. We fix in advance the number of iterations $k \geq 1$ of the method and use *constant* coefficients $\alpha_i := \frac{1}{\sqrt{k}}$, $0 \leq i \leq k-1$. This corresponds to the so-called *Short-Step* Subgradient Method, for which we have

$$\Delta_k \leq \frac{R}{\sqrt{k}}.$$

2. Alternatively, we can use *time-varying* coefficients $\alpha_i := \frac{1}{\sqrt{i+1}}$, $i \geq 0$. This approach does not require us to fix in advance the number of iterations $k$. However, the corresponding convergence rate estimate becomes slightly worse:

$$\Delta_k \leq \frac{\ln k + 2}{2\sqrt{k}} R.$$

(Indeed, $\sum_{i=0}^{k-1} \alpha_i^2 = \sum_{i=1}^{k} \frac{1}{i} \leq \ln k + 1$, while $\sum_{i=0}^{k-1} \alpha_i \geq \sqrt{k}$.)

***Remark 4*** If we allow projections onto the feasible set, then, for the resulting Subgradient Method with time-varying coefficients $\alpha_i$, one can establish the $O(1/\sqrt{k})$ convergence rate for the "truncated" gap

$$\Delta_{k_0,k} := \max_{x \in B(x_0,R)} \frac{1}{\Gamma_{k_0,k}} \sum_{i=k_0}^{k} a_i \langle g_i, x_i - x \rangle,$$

where $\Gamma_{k_0,k} := \sum_{i=k_0}^{k} a_i \|g_i\|_*$, $k_0 := \lceil k/2 \rceil$. For more details, see sect. 5.2.1 in [2] or sect. 3.1.1 in [12].

### 4.2 Standard ellipsoid method

Another extreme choice is the following one:

$$\alpha_k := 0, \qquad \theta := 0, \qquad \gamma > 0. \tag{31}$$

For this choice, we have $a_k = 0$ for all $k \geq 0$. Hence, $\ell_k = 0$ and $L_k^- = \mathbb{E}$ for all $k \geq 0$. Therefore, the localizers in this method are the following ellipsoids (see Lemma 3):

$$\Omega_k \cap L_k^- = \Omega_k = \{x \in \mathbb{E} : \|x - x_k\|_{G_k} \leq R_k\}, \qquad k \geq 0. \tag{32}$$

Observe that, in this example, $\Gamma_k \equiv \sum_{i=0}^{k-1} a_i \|g_i\|_* = 0$ for all $k \geq 1$, so there is no preliminary semicertificate and the sliding gap is undefined. However, we can still ensure the convergence to zero of a certain meaningful measure of optimality, namely, the *average radius* of the localizers $\Omega_k$:

$$\operatorname{avrad} \Omega_k := (\operatorname{vol} \Omega_k)^{1/n}, \qquad k \geq 0. \tag{33}$$

Indeed, let us define the following functions for any real $c$, $p > 0$:

$$q_c(\gamma) := 1 + \frac{c\gamma^2}{2(1+\gamma)}, \qquad \zeta_{p,c}(\gamma) := \frac{[q_c(\gamma)]^p}{1+\gamma}, \qquad \gamma > 0. \tag{34}$$

According to Lemma 4, for any $k \geq 0$, we have

$$R_k^2 = [q_{1/2}(\gamma)]^k R^2. \tag{35}$$

At the same time, in view of (18) and (26), $\det G_k = \prod_{i=0}^{k-1}(1+b_i\|g_i\|_{G_i}^2) = (1+\gamma)^k$ for all $k \geq 0$. Combining this with (32)–(34), we obtain, for any $k \geq 0$, that

$$\operatorname{avrad} \Omega_k = \frac{R_k}{(\det G_k)^{1/(2n)}} = \frac{[q_{1/2}(\gamma)]^{k/2} R}{(1+\gamma)^{k/(2n)}} = [\zeta_{n,1/2}(\gamma)]^{k/(2n)} R. \tag{36}$$

Let us now choose $\gamma$ which minimizes $\operatorname{avrad} \Omega_k$. For such computations, the following auxiliary result is useful (see Sect. A for the proof).

**Lemma 6** *For any $c \geq 1/2$ and any $p \geq 2$, the function $\zeta_{p,c}$, defined in (34), attains its minimum at a unique point*

$$\gamma_c(p) := \frac{2}{\sqrt{c^2p^2 - (2c-1)} + cp - 1} \in \left[\frac{1}{cp}\frac{2}{cp}\right] \tag{37}$$

*with the corresponding value $\zeta_{p,c}\big(\gamma_c(p)\big) \leq e^{-1/(2cp)}$.*

Applying Lemma 6 to 36, we see that the optimal value of $\gamma$ is

$$\gamma := \gamma_{1/2}(n) = \frac{2}{n/2 + n/2 - 1} = \frac{2}{n-1}, \tag{38}$$

for which $\zeta_{n,1/2}(\gamma) \leq e^{-1/n}$. With this choice of $\gamma$, we obtain, for all $k \geq 0$, that

$$\operatorname{avrad} \Omega_k \leq e^{-k/(2n^2)} R. \tag{39}$$

One can check that Algorithm 1 with parameters (26), (31) and (38) is, in fact, the standard Ellipsoid Method (see Remark 6).

### 4.3 Ellipsoid method with preliminary semicertificate

As we have seen, we cannot measure the convergence rate of the standard Ellipsoid Method using the sliding gap because there is no preliminary semicertificate in this method. Let us present a modification of the standard Ellipsoid Method which does not have this drawback but still enjoys the same convergence rate as the original method (up to some absolute constants).

For this, let us choose the coefficients in the following way:

$$\alpha_k := 0, \qquad \theta := \sqrt{2} - 1 \ (\approx 0.41), \qquad \gamma > 0. \tag{40}$$

Then, in view of Lemma 4, for all $k \geq 0$, we have

$$R_k^2 = [q_1(\gamma)]^k R^2, \tag{41}$$

Also, by Lemma 5, $\Gamma_k \geq \frac{1}{2}\theta R\sqrt{\gamma n[(1+\gamma)^{k/n}-1]}$ for all $k \geq 1$. Thus, for each $k \geq 1$, we obtain the following estimate for the sliding gap (see (25)):

$$\Delta_k \leq \frac{[q_1(\gamma)]^k R}{\theta\sqrt{\gamma n[(1+\gamma)^{k/n}-1]}} = \frac{1}{\theta\kappa_k(\gamma,n)}[\zeta_{2n,1}(\gamma)]^{k/(2n)}R, \tag{42}$$

where $\kappa_k(\gamma,n) := \sqrt{\gamma n(1-\frac{1}{(1+\gamma)^{k/n}})}$ and $\zeta_{2n,1}(\gamma)$ is defined in (34).

Note that the main factor in estimate (42) is $[\zeta_{2n,1}(\gamma)]^{k/(2n)}$. Let us choose $\gamma$ by minimizing this expression. Applying Lemma 6, we obtain

$$\gamma := \gamma_1(2n) \in \left[\frac{1}{2n}\frac{1}{n}\right]. \tag{43}$$

**Theorem 1** *In Algorithm 1 with parameters (26), (40), (43), for all $k \geq 1$,*

$$\Delta_k \leq 6e^{-k/(8n^2)}R.$$

***Proof*** Suppose $k \geq n^2$. According to Lemma 6, we have $\zeta_{2n,1}(\gamma) \leq e^{-1/(4n)}$. Hence, by (42), $\Delta_k \leq \frac{1}{\theta\kappa_k(\gamma,n)}e^{-k/(8n^2)}R$. It remains to estimate from below $\theta\kappa_k(\gamma,n)$.

Since $k \geq n^2$, we have $(1+\gamma)^{k/n} \geq (1+\gamma)^n \geq 1+\gamma n$. Hence, $\kappa_k(\gamma,n) \geq \frac{\gamma n}{\sqrt{1+\gamma n}}$. Note that the function $\tau \mapsto \frac{\tau}{\sqrt{1+\tau}}$ is increasing on $\mathbb{R}_+$. Therefore, using (43), we obtain $\kappa_k(\gamma,n) \geq \frac{1/2}{\sqrt{1+1/2}} = \frac{1}{\sqrt{6}}$. Thus, $\theta\kappa_k(\gamma,n) \geq \frac{\sqrt{2}-1}{\sqrt{6}} \geq \frac{1}{6}$ for our choice of $\theta$.

Now suppose $k \leq n^2$. Then, $6e^{-k/(8n^2)} \geq 6e^{-1/8} \geq 5$. Therefore, it suffices to prove that $\Delta_k \leq 5R$ or, in view of (24), that $\langle g_i, x_i - x\rangle \leq 5R\|g_i\|_*$, where $x \in \Omega_k \cap L_k^-$ and $0 \leq i \leq k-1$ are arbitrary. Note that $\langle g_i, x_i - x\rangle \leq \|g_i\|_{G_i}^*\|x_i - x\|_{G_i} \leq \|g_i\|_*\|x_i - x\|_{G_i}$ since $G_i \succeq B$ (see (18)). Hence, it remains to prove that $\|x_i - x\|_{G_i} \leq 5R$.

Recall from (18) and (19) that $G_i \preceq G_k$ and $R_i \leq R_k$. Therefore,

$$\begin{aligned}\|x_i - x\|_{G_i} &\leq \|x_i - x^*\|_{G_i} + \|x^* - x\|_{G_i} \leq \|x_i - x^*\|_{G_i} + \|x^* - x\|_{G_k}\\ &\leq \|x_i - x^*\|_{G_i} + \|x_k - x^*\|_{G_k} + \|x_k - x\|_{G_k} \leq R_i + 2R_k \leq 3R_k,\end{aligned}$$

where the penultimate inequality follows from Lemma 2 and 3. According to (41), $R_k = [q_1(\gamma)]^{k/2}R \leq [q_1(\gamma)]^{n^2/2}R$ (recall that $q_1(\gamma) \geq 1$). Thus, it remains to show that $3[q_1(\gamma)]^{n^2/2} \leq 5$. But this is immediate. Indeed, by (34) and (43), we have $[q_1(\gamma)]^{n^2/2} \leq e^{n^2\gamma^2/(4(1+\gamma))} \leq e^{1/4}$, so $3[q_1(\gamma)]^{n^2/2} \leq 3e^{1/4} \leq 5$. □

### 4.4 Subgradient ellipsoid method

The previous algorithm still shares the drawback of the original Ellipsoid Method, namely, it does not work when $n \to \infty$. To eliminate this drawback, let us choose $\alpha_k$ similarly to how this is done in the Subgradient Method.

Consider the following choice of parameters:

$$\alpha_i := \beta_i \sqrt{\frac{\theta}{\theta+1}}, \qquad \theta := \sqrt[3]{2} - 1 \ (\approx 0.26), \qquad \gamma := \gamma_1(2n) \in \left[\frac{1}{2n}, \frac{1}{n}\right], \quad (44)$$

where $\beta_i > 0$ are certain coefficients (to be specified later) and $\gamma_1(2n)$ is defined in (37).

**Theorem 2** *In Algorithm 1 with parameters (26) and (44), where $\beta_0 \geq 1$, we have, for all $k \geq 1$,*

$$\Delta_k \leq \begin{cases} \frac{2}{\sum_{i=0}^{k-1} \beta_i}(1 + \sum_{i=0}^{k-1} \beta_i^2)R, & \text{if } k \leq n^2, \\ 6e^{-k/(8n^2)}(1 + \sum_{i=0}^{k-1} \beta_i^2)R, & \text{if } k \geq n^2. \end{cases} \quad (45)$$

***Proof*** Applying Lemma 4 with $\tau := \theta$ and using (44), we obtain

$$R_k^2 \leq [q_1(\gamma)]^k C_k R^2, \qquad C_k = 1 + \sum_{i=0}^{k-1} \beta_i^2. \quad (46)$$

At the same time, by Lemma 5, we have

$$\Gamma_k \geq R\left(\sqrt{\frac{\theta}{\theta+1}} \sum_{i=0}^{k-1} \beta_i + \tfrac{1}{2}\theta\sqrt{\gamma n[(1+\gamma)^{k/n} - 1]}\right). \quad (47)$$

Note that $\frac{1}{2}\theta\sqrt{\gamma n} \leq \frac{1}{2}\theta \leq \sqrt{\theta/(\theta+1)}$ by (44). Since $\beta_0 \geq 1$, we thus obtain

$$\begin{aligned} \Gamma_k &\geq \tfrac{1}{2}R\theta\sqrt{\gamma n}\left(1 + \sqrt{(1+\gamma)^{k/n} - 1}\right) \geq \tfrac{1}{2}R\theta\sqrt{\gamma n}(1+\gamma)^{k/(2n)} \\ &\geq \tfrac{1}{2\sqrt{2}}R\theta(1+\gamma)^{k/(2n)} \geq \tfrac{1}{12}R(1+\gamma)^{k/(2n)}, \end{aligned} \quad (48)$$

where the last two inequalities follow from (44). Therefore, by (25), (46) and (48),

$$\Delta_k \leq \frac{R_k^2}{2\Gamma_k} \leq 6\frac{[q_1(\gamma)]^k}{(1+\gamma)^{k/(2n)}} C_k R = 6[\zeta_{2n,1}(\gamma)]^{k/(2n)} C_k R,$$

where $\zeta_{2n,1}(\gamma)$ is defined in (34). Observe that, for our choice of $\gamma$, by Lemma 6, we have $\zeta_{2n,1}(\gamma) \leq e^{-1/(4n)}$. This proves the second estimate[3] in (45).

On the other hand, dropping the second term in (47), we can write

$$\Gamma_k \geq R\sqrt{\frac{\theta}{\theta+1}} \sum_{i=0}^{k-1} \beta_i. \quad (49)$$

[3] In fact, we have proved the second estimate in (45) for all $k \geq 1$ (not only for $k \geq n^2$).

Suppose $k \le n^2$. Then, from (34) and (44), it follows that

$$[q_1(\gamma)]^k \le [q_1(\gamma)]^{n^2} \le e^{\gamma^2 n^2/(2(1+\gamma))} \le \sqrt{e}.$$

Hence, by (46), $R_k \le \sqrt{e} C_k R^2$. Combining this with (25) and (49), we obtain

$$\Delta_k \le \frac{1}{2}\sqrt{\frac{e(\theta+1)}{\theta}} \frac{1}{\sum_{i=0}^{k-1} \beta_i} C_k R.$$

By numerical evaluation, one can verify that, for our choice of $\theta$, we have $\frac{1}{2}\sqrt{\frac{e(\theta+1)}{\theta}} \le 2$. This proves the first estimate in (45). □

Exactly as in the Subgradient Method, we can use the following two strategies for choosing the coefficients $\beta_i$:

1. We fix in advance the number of iterations $k \ge 1$ of the method and use constant coefficients $\beta_i := \frac{1}{\sqrt{k}}$, $0 \le i \le k-1$. In this case,

$$\Delta_k \le \begin{cases} 4R/\sqrt{k}, & \text{if } k \le n^2, \\ 12Re^{-k/(8n^2)}, & \text{if } k \ge n^2. \end{cases} \tag{50}$$

2. We use time-varying coefficients $\beta_i := \frac{1}{\sqrt{i+1}}$, $i \ge 0$. In this case,

$$\Delta_k \le \begin{cases} 2(\ln k + 2)R/\sqrt{k}, & \text{if } k \le n^2, \\ 6(\ln k + 2)Re^{-k/(8n^2)}, & \text{if } k \ge n^2. \end{cases}$$

Let us discuss convergence rate estimate (50). Up to absolute constants, this estimate is exactly the same as in the Subgradient Method when $k \le n^2$ and as in the Ellipsoid Method when $k \ge n^2$. In particular, when $n \to \infty$, we recover the convergence rate of the Subgradient Method.

To provide a better interpretation of the obtained results, let us compare the convergence rates of the Subgradient and Ellipsoid methods:

$$\begin{aligned} \text{Subgradient Method:} &\quad 1/\sqrt{k} \\ \text{Ellipsoid Method:} &\quad e^{-k/(2n^2)}. \end{aligned}$$

To compare these rates, let us look at their squared ratio:

$$\rho_k := \left(\frac{1/\sqrt{k}}{e^{-k/(2n^2)}}\right)^2 = \frac{e^{k/n^2}}{k}.$$

Let us find out for which values of $k$ the rate of the Subgradient Method is better than that of the Ellipsoid Method and vice versa. We assume that $n \ge 2$.

Note that the function $\tau \mapsto e^\tau/\tau$ is strictly decreasing on $(0, 1]$ and strictly increasing on $[1, +\infty)$ (indeed, its derivative equals $e^\tau(\tau - 1)/\tau^2$). Hence, $\rho_k$ is strictly decreasing in $k$ for $1 \le k \le n^2$ and strictly increasing in $k$ for $k \ge n^2$. Since $n \ge 2$, we have $\rho_2 = e^{2/n^2}/2 \le e^{1/2}/2 \le 1$. At the same time, $\rho_k \to +\infty$ when $k \to \infty$. Therefore, there exists a unique integer $K_0 \ge 2$ such that $\rho_k \le 1$ for all $k \le K_0$ and $\rho_k \ge 1$ for all $k \ge K_0$.

Let us estimate $K_0$. Clearly, for any $n^2 \le k \le n^2 \ln(2n)$, we have

$$\rho_k \le \frac{e^{n^2 \ln(2n)/n^2}}{n^2 \ln(2n)} = \frac{2}{n \ln(2n)} \le 1,$$

while, for any $k \ge 3n^2 \ln(2n)$, we have

$$\rho_k \ge \frac{e^{3n^2 \ln(2n)}/n^2}{3n^2 \ln(2n)} = \frac{(2n)^3}{3n^2 \ln(2n)} = \frac{8n}{3 \ln(2n)} \ge 1.$$

Hence,

$$n^2 \ln(2n) \le K_0 \le 3n^2 \ln(2n).$$

Thus, up to an absolute constant, $n^2 \ln(2n)$ is the switching moment, starting from which the rate of the Ellipsoid Method becomes better than that of the Subgradient Method.

Returning to our obtained estimate (50), we see that, ignoring absolute constants and ignoring the "small" region of the values of $k$ between $n^2$ and $n^2 \ln n$, our convergence rate is basically the best of the corresponding convergence rates of the Subgradient and Ellipsoid methods.

## 5 Constructing accuracy semicertificate

Let us show how to convert a preliminary accuracy semicertificate, produced by Algorithm 1, into a semicertificate whose gap on the initial solid is upper bounded by the sliding gap. The key ingredient here is the following auxiliary algorithm which was first proposed in [16] for building accuracy certificates in the standard Ellipsoid Method.

### 5.1 Augmentation algorithm

Let $k \ge 0$ be an integer and let $Q_0, \dots, Q_k$ be solids in $\mathbb{E}$ such that

$$\hat{Q}_i := \{x \in Q_i : \langle g_i, x - x_i \rangle \le 0\} \subseteq Q_{i+1}, \qquad 0 \le i \le k - 1, \tag{51}$$

where $x_i \in \mathbb{E}$, $g_i \in \mathbb{E}^*$. Further, suppose that, for any $s \in \mathbb{E}^*$ and any $0 \le i \le k-1$, we can compute a *dual multiplier* $\mu \ge 0$ such that

$$\max_{x \in \hat{Q}_i} \langle s, x \rangle = \max_{x \in Q_i} [\langle s, x \rangle + \mu \langle g_i, x_i - x \rangle] \tag{52}$$

(provided that certain regularity conditions hold). Let us abbreviate any solution $\mu$ of this problem by $\mu(s, Q_i, x_i, g_i)$.

Consider now the following routine.

**Algorithm 2: Augmentation Algorithm**

**Input:** $s_k \in \mathbb{E}^*$.

**Iterate for** $i = k-1, \ldots, 0$**:**
1. Compute $\mu_i := \mu(s_{i+1}, Q_i, x_i, g_i)$.
2. Set $s_i := s_{i+1} - \mu_i g_i$.

**Lemma 7** *Let $\mu_0, \ldots, \mu_{k-1} \ge 0$ be generated by Algorithm 2. Then,*

$$\max_{x \in Q_0} \Big[ \langle s_k, x \rangle + \sum_{i=0}^{k-1} \mu_i \langle g_i, x_i - x \rangle \Big] \le \max_{x \in Q_k} \langle s_k, x \rangle.$$

***Proof*** Indeed, at every iteration $i = k-1, \ldots, 0$, we have

$$\begin{aligned} \max_{x \in Q_{i+1}} \langle s_{i+1}, x \rangle &\ge \max_{x \in \hat{Q}_i} \langle s_{i+1}, x \rangle = \max_{x \in Q_i} [\langle s_{i+1}, x \rangle + \mu_i \langle g_i, x_i - x \rangle] \\ &= \max_{x \in Q_i} \langle s_i, x \rangle + \mu_i \langle g_i, x_i \rangle. \end{aligned}$$

Summing up these inequalities for $i = 0, \ldots, k-1$, we obtain

$$\max_{x \in Q_k} \langle s_k, x \rangle \ge \max_{x \in Q_0} \langle s_0, x \rangle + \sum_{i=0}^{k-1} \mu_i \langle g_i, x_i \rangle = \max_{x \in Q_0} \Big[ \langle s_k, x \rangle + \sum_{i=0}^{k-1} \langle g_i, x_i - x \rangle \Big],$$

where the identity follows from the fact that $s_0 = s_k - \sum_{i=0}^{k-1} \mu_i g_i$. □

### 5.2 Methods with preliminary certificate

Let us apply the Augmentation Algorithm for building an accuracy semicertificate for Algorithm 1. We only consider those instances for which $\Gamma_k := \sum_{i=0}^{k-1} a_i \|g_i\|_* > 0$

so that the sliding gap $\Delta_k$ is well-defined:

$$\Delta_k := \max_{x \in \Omega_k} \frac{1}{\Gamma_k}[-\ell_k(x)] = \max_{x \in \Omega_k \cap L_k^-} \frac{1}{\Gamma_k}[-\ell_k(x)]$$
$$= \max_{x \in \Omega_k \cap L_k^-} \frac{1}{\Gamma_k} \sum_{i=0}^{k-1} a_i \langle g_i, x_i - x \rangle.$$

Recall that the vector $a := (a_0, \ldots, a_{k-1})$ is called a preliminary semicertificate.

For technical reasons, it will be convenient to add the following termination criterion into Algorithm 1:

$$\text{Terminate Algorithm 1 at Step 2 if } U_k \le \delta \|g_k\|_*, \tag{53}$$

where $\delta > 0$ is a fixed constant. Depending on whether this termination criterion has been satisfied at iteration $k$, we call it a *terminal* or *nonterminal* iteration, respectively.

***Remark 5*** In practice, one can set $\delta$ to an arbitrarily small value (within machine precision) if the desired target accuracy is unknown. As can be seen from the subsequent discussion, the main purpose of the termination criterion (53) is to ensure that $U_k$ never becomes equal to zero during the iterations of Algorithm 1. This guarantees the existence of dual multiplier in (52) for any $s \in \mathbb{E}^*$ at every nonterminal iteration. The case $U_k = 0$ corresponds to the degenerate situation when Algorithm 1 has "accidentally" found an exact solution.

Let $k \ge 1$ be an iteration of Algorithm 1. According to Lemma 2, the sets $Q_i := \Omega_i \cap L_i^-$ satisfy (51). Since the method has not been terminated during the course of the previous iterations, we have[4] $U_i > 0$ for all $0 \le i \le k-1$. Therefore, for any $0 \le i \le k-1$, there exists $x \in Q_i$ such that $\langle g_i, x - x_i \rangle < 0$. This guarantees the existence of dual multiplier in (52).

Let us apply Algorithm 2 to $s_k := -\sum_{i=0}^{k-1} a_i g_i$ in order to obtain dual multipliers $\mu := (\mu_0, \ldots, \mu_{k-1})$. From Lemma 7, it follows that

$$\max_{x \in B(x_0, R)} \sum_{i=0}^{k-1} (a_i + \mu_i) \langle g_i, x_i - x \rangle \le \max_{x \in Q_k} \sum_{i=0}^{k-1} a_i \langle g_i, x_i - x \rangle = \Gamma_k \Delta_k,$$

(note that $Q_0 = \Omega_0 \cap L_0^- = B(x_0, R)$). Thus, defining $\lambda := a + \mu$, we obtain $\Gamma_k(\lambda) \equiv \sum_{i=0}^{k-1} \lambda_i \|g_i\|_* \ge \sum_{i=0}^{k-1} a_i \|g_i\|_* \equiv \Gamma_k > 0$ and

$$\delta_k(\lambda) \equiv \max_{x \in B(x_0, R)} \frac{1}{\Gamma_k(\lambda)} \sum_{i=0}^{k-1} \lambda_i \langle g_i, x_i - x \rangle \le \frac{\Gamma_k}{\Gamma_k(\lambda)} \Delta_k \le \Delta_k,$$

[4] Recall that $g_i \ne 0$ for all $i \ge 0$ by (2).

Thus, $\lambda$ is a semicertificate whose gap on $B(x_0, R)$ is bounded by the sliding gap $\Delta_k$.

If $k \geq 0$ is a terminal iteration, then, by the termination criterion and the definition of $U_k$ (see Algorithm 1), we have $\max_{x \in \Omega_k \cap L_k^-} \frac{1}{\|g_k\|_*} \langle g_k, x_k - x \rangle \leq \delta$. In this case, we apply Algorithm 2 to $s_k := -g_k$ to obtain dual multipliers $\mu_0, \ldots, \mu_{k-1}$. By the same reasoning as above but with the vector $(0, \ldots, 0, 1)$ instead of $(a_0, \ldots, a_{k-1})$, we can obtain that $\delta_{k+1}(\lambda) \leq \delta$, where $\lambda := (\mu_0, \ldots, \mu_{k-1}, 1)$.

### 5.3 Standard ellipsoid method

In the standard Ellipsoid Method, there is no preliminary semicertificate. Therefore, we cannot apply the above procedure. However, in this method, it is still possible to generate an accuracy semicertificate, although the corresponding procedure is slightly more involved. Let us now briefly describe this procedure and discuss how it differs from the previous approach. For details, we refer the reader to [16].

Let $k \geq 1$ be an iteration of the method. There are two main steps. The first step is to find a direction $s_k$, in which the "width" of the ellipsoid $\Omega_k$ (see (32)) is minimal:

$$s_k := \operatorname*{argmin}_{\|s\|_*=1} \max_{x,y \in \Omega_k} \langle s, x - y \rangle = \operatorname*{argmin}_{\|s\|_*=1} \Big[ \max_{x \in \Omega_k} \langle s, x \rangle - \min_{x \in \Omega_k} \langle s, x \rangle \Big].$$

It is not difficult to see that $s_k$ is given by the unit eigenvector[5] of the operator $G_k$, corresponding to the largest eigenvalue. For the corresponding minimal "width" of the ellipsoid, we have the following bound via the average radius:

$$\max_{x,y \in \Omega_k} \langle s_k, x - y \rangle \leq \rho_k, \tag{54}$$

where $\rho_k := 2 \operatorname{avrad} \Omega_k$. Recall that $\operatorname{avrad} \Omega_k \leq e^{-k/(2n^2)} R$ in view of (39).

At the second step, we apply Algorithm 2 two times with the sets $Q_i := \Omega_i$: first, to the vector $s_k$ to obtain dual multipliers $\mu := (\mu_0, \ldots, \mu_{k-1})$ and then to the vector $-s_k$ to obtain dual multipliers $\mu' := (\mu'_0, \ldots, \mu'_{k-1})$. By Lemma 7 and (54), we have

$$\max_{x \in B(x_0,R)} \Big[ \langle s_k, x - x_k \rangle + \sum_{i=0}^{k-1} \mu_i \langle g_i, x_i - x \rangle \Big] \leq \max_{x \in \Omega_k} \langle s_k, x - x_k \rangle \leq \rho_k,$$

$$\max_{x \in B(x_0,R)} \Big[ \langle s_k, x_k - x \rangle + \sum_{i=0}^{k-1} \mu'_i \langle g_i, x_i - x \rangle \Big] \leq \max_{x \in \Omega_k} \langle s_k, x_k - x \rangle \leq \rho_k$$

(note that $Q_0 = \Omega_0 = B(x_0, R)$). Consequently, for $\lambda := \mu + \mu'$, we obtain

$$\max_{x \in B(x_0,R)} \sum_{i=0}^{k-1} \lambda_i \langle g_i, x_i - x \rangle \leq 2\rho_k.$$

[5] Here eigenvectors and eigenvalues are defined with respect to the operator $B$ inducing the norm $\|x\|$.

Finally, one can show that

$$\Gamma_k(\lambda) \equiv \sum_{i=0}^{k-1} \lambda_i \|g_i\|_* \geq \frac{r - \rho_k}{D},$$

where $D$ is the diameter of $Q$ and $r$ is the maximal of the radii of Euclidean balls contained in $Q$. Thus, whenever $\rho_k < r$, $\lambda$ is a semicertificate with the following gap on $B(x_0, R)$:

$$\delta_k(\lambda) \equiv \max_{x \in B(x_0,R)} \frac{1}{\Gamma_k(\lambda)} \sum_{i=0}^{k-1} \lambda_i \langle g_i, x_i - x \rangle \leq \frac{2\rho_k D}{r - \rho_k}.$$

Compared to the standard Ellipsoid Method, we see that, in the Subgradient Ellipsoid methods, the presence of the preliminary semicertificate removes the necessity in finding the minimal-"width" direction and requires only one run of the Augmentation Algorithm.

## 6 Implementation details

### 6.1 Explicit representations

In the implementation of Algorithm 1, instead of the operators $G_k$, it is better to work with their inverses $H_k := G_k^{-1}$. Applying the Sherman-Morrison formula to (18), we obtain the following update rule for $H_k$:

$$H_{k+1} = H_k - \frac{b_k H_k g_k g_k^* H_k}{1 + b_k \langle g_k, H_k g_k \rangle}, \quad k \geq 0. \tag{55}$$

Let us now obtain an explicit formula for the next test point $x_{k+1}$. This has already been partly done in the proof of Lemma 3. Indeed, recall that $x_{k+1}$ is the minimizer of the function $\psi_{k+1}(x)$. From (21), we see that $x_{k+1} = x_k - (a_k + \frac{1}{2} b_k U_k) H_{k+1} g_k$. Combining it with (55), we obtain

$$x_{k+1} = x_k - \frac{a_k + \frac{1}{2} b_k U_k}{1 + b_k \langle g_k, H_k g_k \rangle} H_k g_k, \quad k \geq 0. \tag{56}$$

Finally, one can obtain the following explicit representations for $L_k^-$ and $\Omega_k$:

$$L_k^- = \{x \in \mathbb{E} : \langle c_k, x \rangle \leq \sigma_k\}, \qquad \Omega_k = \{x \in \mathbb{E} : \|x - z_k\|^2_{H_k^{-1}} \leq D_k\}, \tag{57}$$

where, for any $k \geq 0$,

$$\begin{aligned} c_0 := 0, \quad \sigma_0 := 0, \qquad c_{k+1} := c_k + a_k g_k, \quad \sigma_{k+1} := \sigma_k + a_k \langle g_k, x_k \rangle, \\ z_k := x_k - H_k c_k, \quad D_k := R_k^2 + 2(\sigma_k - \langle c_k, x_k \rangle) + \langle c_k, H_k c_k \rangle. \end{aligned} \tag{58}$$

Indeed, recalling the definition of functions $\ell_k$, we see that $\ell_k(x) = \langle c_k, x\rangle - \sigma_k$ for all $x \in \mathbb{E}$. Therefore, $L_k^- \equiv \{x : \ell_k(x) \le 0\} = \{x : \langle c_k, x\rangle \le \sigma_k\}$. Further, by Lemma 3, $\Omega_k = \{x : \langle c_k, x\rangle + \frac{1}{2}\|x - x_k\|_{G_k}^2 \le \frac{1}{2}R_k^2 + \sigma_k\}$. Note that $\langle c_k, x\rangle + \frac{1}{2}\|x - x_k\|_{G_k}^2 = \frac{1}{2}\|x - z_k\|_{G_k}^2 + \langle c_k, x_k\rangle - \frac{1}{2}\|c_k\|_{G_k}^2$ for any $x \in \mathbb{E}$. Hence, $\Omega_k = \{x : \frac{1}{2}\|x - z_k\|_{G_k}^2 \le \frac{1}{2}D_k\}$.

***Remark 6*** Now we can justify the claim made in Sect. 4.2 that Algorithm 1 with parameters (26), (31) and (38) is the standard Ellipsoid Method. Indeed, from (26) and (32), we see that $b_k = \frac{\gamma}{\langle g_k, H_k g_k\rangle}$ and $U_k = R_k\langle g_k, H_k g_k\rangle^{1/2}$. Also, in view of (38), $\frac{\gamma}{1+\gamma} = \frac{2}{n+1}$. Hence, by (56) and (55),

$$\begin{aligned} x_{k+1} &= x_k - \frac{R_k}{n+1}\frac{H_k g_k}{\langle g_k, H_k g_k\rangle^{1/2}}, \\ H_{k+1} &= H_k - \frac{2}{n+1}\frac{H_k g_k g_k^* H_k}{\langle g_k, H_k g_k\rangle}, \qquad k \ge 0. \end{aligned} \tag{59}$$

Further, according to (35) and (38), for any $k \ge 0$, we have $R_k^2 = q^k R^2$, where $q = 1 + \frac{1}{(n-1)(n+1)} = \frac{n^2}{n^2-1}$. Thus, method (59) indeed coincides[6] with the standard Ellipsoid Method (2) under the change of variables $W_k := R_k^2 H_k$.

### 6.2 Computing support function

To calculate $U_k$ in Algorithm 1, we need to compute the following quantity (see (57)):

$$U_k = \max_x\{\langle g_k, x_k - x\rangle : \|x - z_k\|_{H_k^{-1}}^2 \le D_k, \langle c_k, x\rangle \le \sigma_k\}.$$

Let us discuss how to do this.

First, let us introduce the following support function to simplify our notation:

$$\xi(H, s, a, \beta) := \max_x\{\langle s, x\rangle : \|x\|_{H^{-1}}^2 \le 1, \langle a, x\rangle \le \beta\},$$

where $H : \mathbb{E}^* \to \mathbb{E}$ is a self-adjoint positive definite linear operator, $s, a \in \mathbb{E}^*$ and $\beta \in \mathbb{R}$. In this notation, assuming that $D_k > 0$, we have

$$U_k = \langle g_k, x_k - z_k\rangle + \xi(D_k H_k, -g_k, c_k, \sigma_k - \langle c_k, z_k\rangle).$$

Let us show how to compute $\xi(H, s, a, \beta)$. Dualizing the linear constraint, we obtain

$$\xi(H, s, a, \beta) = \min_{\tau \ge 0}\big[\|s - \tau a\|_{H^{-1}}^* + \tau\beta\big], \tag{60}$$

[6] Note that, in (2), we identify the spaces $\mathbb{E}, \mathbb{E}^*$ with $\mathbb{R}^n$ in such a way that $\langle\cdot,\cdot\rangle$ coincides with the standard dot-product and $\|x\|$ coincides with the standard Euclidean norm. Therefore, $B$ becomes the identity matrix and $g_k^*$ becomes $g_k^T$.

provided that there exists some $x \in \mathbb{E}$ such that $\|x\|_{H^{-1}} < 1$, $\langle a, x\rangle \le \beta$ (Slater condition). One can show that (60) has the following solution (see Lemma 10):

$$\tau(H, s, a, \beta) := \begin{cases} 0, & \text{if } \langle a, Hs\rangle \le \beta \|s\|^*_{H^{-1}}, \\ u(H, s, a, \beta), & \text{otherwise,} \end{cases} \tag{61}$$

where $u(H, s, a, \beta)$ is the unconstrained minimizer of the objective function in (60).

Let us present an explicit formula for $u(H, s, a, \beta)$. For future use, it will be convenient to write down this formula in a slightly more general form for the following multidimensional[7] variant of problem (60):

$$\min_{u \in \mathbb{R}^m} \big[\|s - Au\|^*_{H^{-1}} + \langle u, b\rangle\big], \tag{62}$$

where $s \in \mathbb{E}^*$, $H : \mathbb{E}^* \to \mathbb{E}$ is a self-adjoint positive definite linear operator, $A : \mathbb{R}^m \to \mathbb{E}^*$ is a linear operator with trivial kernel and $b \in \mathbb{R}^m$, $\langle b, (A^*HA)^{-1}b\rangle < 1$. It is not difficult to show that problem (62) has the following unique solution (see Lemma 9):

$$\begin{aligned} u(H, s, A, b) &:= (A^*HA)^{-1}(A^*s - rb), \\ r &:= \sqrt{\frac{\langle s, Hs\rangle - \langle s, A(A^*HA)^{-1}A^*s\rangle}{1 - \langle b, (A^*HA)^{-1}b\rangle}}. \end{aligned} \tag{63}$$

Note that, in order for the above approach to work, we need to guarantee that the sets $\Omega_k$ and $L_k^-$ satisfy a certain regularity condition, namely, $\operatorname{int} \Omega_k \cap L_k^- \ne \emptyset$. This condition can be easily fulfilled by adding into Algorithm 1 the termination criterion (53).

**Lemma 8** *Consider Algorithm 1 with termination criterion (53). Then, at each iteration $k \ge 0$, at the beginning of Step 2, we have* $\operatorname{int} \Omega_k \cap L_k^- \ne \emptyset$. *Moreover, if $k$ is a nonterminal iteration, we also have $\langle g_k, x - x_k\rangle \le 0$ for some $x \in \operatorname{int} \Omega_k \cap L_k^-$.*

***Proof*** Note that $\operatorname{int} \Omega_0 \cap L_0^- = \operatorname{int} B(x_0, R) \ne \emptyset$. Now suppose $\operatorname{int} \Omega_k \cap L_k^- \ne \emptyset$ for some nonterminal iteration $k \ge 0$. Denote $P_k^- := \{x \in \mathbb{E} : \langle g_k, x - x_k\rangle \le 0\}$. Since iteration $k$ is nonterminal, $U_k > 0$ and hence $\Omega_k \cap L_k^- \cap \operatorname{int} P_k^- \ne \emptyset$. Combining it with the fact that $\operatorname{int} \Omega_k \cap L_k^- \ne \emptyset$, we obtain $\operatorname{int} \Omega_k \cap L_k^- \cap \operatorname{int} P_k^- \ne \emptyset$ and, in particular, $\operatorname{int} \Omega_k \cap L_k^- \cap P_k^- \ne \emptyset$. At the same time, slightly modifying the proof of Lemma 2 (using that $\operatorname{int} \Omega_i = \{x \in \mathbb{E} : \omega_i(x) < \frac{1}{2}R^2\}$ for any $i \ge 0$ since $\omega_i$ is a strictly convex quadratic function), it is not difficult to show that $\operatorname{int} \Omega_k \cap L_k^- \cap P_k^- \subseteq \operatorname{int} \Omega_{k+1} \cap L_{k+1}^-$. Thus, $\operatorname{int} \Omega_{k+1} \cap L_{k+1}^- \ne \emptyset$, and we can continue by induction. □

### 6.3 Computing dual multipliers

Recall from Sect. 5 that the procedure for generating an accuracy semicertificate for Algorithm 1 requires one to repeatedly carry out the following operation: given $s \in \mathbb{E}^*$

[7] Hereinafter, we identify $(\mathbb{R}^m)^*$ with $\mathbb{R}^m$ in such a way that $\langle \cdot, \cdot \rangle$ is the standard dot product.

and some iteration number $i \geq 0$, compute a dual multiplier $\mu \geq 0$ such that

$$\max_{x \in \Omega_i \cap L_i^-} \{\langle s, x\rangle : \langle g_i, x - x_i\rangle \leq 0\} = \max_{x \in \Omega_i \cap L_i^-} \big[\langle s, x\rangle + \mu \langle g_i, x_i - x\rangle\big].$$

This can be done as follows.

First, using (57), let us rewrite the above primal problem more explicitly:

$$\max_x \{\langle s, x\rangle : \|x - z_i\|^2_{H_i^{-1}} \leq D_i, \langle c_i, x\rangle \leq \sigma_i, \langle g_i, x - x_i\rangle \leq 0\}.$$

Our goal is to dualize the second linear constraint and find the corresponding multiplier. However, for the sake of symmetry, it is better to dualize both linear constraints, find the corresponding multipliers and then keep only the second one.

Let us simplify our notation by introducing the following problem:

$$\max_x \{\langle s, x\rangle : \|x\|_{H^{-1}} \leq 1, \langle a_1, x\rangle \leq b_1, \langle a_2, x\rangle \leq b_2\}, \tag{64}$$

where $H : \mathbb{E}^* \to \mathbb{E}$ is a self-adjoint positive definite linear operator, $s, a_1, a_2 \in \mathbb{E}^*$ and $b_1, b_2 \in \mathbb{R}$. Clearly, our original problem can be transformed into this form by setting $H := D_i H_i$, $a_1 := c_i$, $a_2 := g_i$, $b_1 := \sigma_i - \langle c_i, z_i\rangle$, $b_2 := \langle g_i, x_i - z_i\rangle$. Note that this transformation does not change the dual multipliers.

Dualizing the linear constraints in (64), we obtain the following dual problem:

$$\min_{\mu \in \mathbb{R}^2_+} \big[\|s - \mu_1 a_1 - \mu_2 a_2\|^*_{H^{-1}} + \mu_1 b_1 + \mu_2 b_2\big], \tag{65}$$

which is solvable provided the following Slater condition holds:

$$\exists x \in \mathbb{E} : \|x\|_{H^{-1}} < 1, \langle a_1, x\rangle \leq b_1, \langle a_2, x\rangle \leq b_2. \tag{66}$$

Note that (66) can be ensured by adding termination criterion (53) into Algorithm 1 (see Lemma 8).

A solution of (65) can be found using Algorithm 3. In this routine, $\tau(\cdot)$, $\xi(\cdot)$ and $u(\cdot)$ are the auxiliary operations, defined in Sect. 6.2, and $A := (a_1, a_2)$ is the linear operator $Au := u_1 a_1 + u_2 a_2$ acting from $\mathbb{R}^2$ to $\mathbb{E}^*$. The correctness of Algorithm 3 is proved in Theorem 3.

**Algorithm 3: Computing Dual Multipliers**

1. Compute $\tau_1 := \tau(H, s, a_1, b_1)$ and $\tau_2 := \tau(H, s, a_2, b_2)$.
Compute $\xi_1 := \xi(H, a_2, a_1, b_1)$ and $\xi_2 := \xi(H, a_1, a_2, b_2)$.
2. If $\xi_1 \leq b_2$, return $(\tau_1, 0)$. Else if $\xi_2 \leq b_1$, return $(0, \tau_2)$.
3. Else if $\langle a_2, H(s - \tau_1 a_1)\rangle \leq b_2 \|s - \tau_1 a_1\|^*_{H^{-1}}$, return $(\tau_1, 0)$.
Else if $\langle a_1, H(s - \tau_2 a_2)\rangle \leq b_1 \|s - \tau_2 a_2\|^*_{H^{-1}}$, return $(0, \tau_2)$.
4. Else return $u := u(H, s, A, b)$, where $A := (a_1, a_2)$, $b := (b_1, b_2)^T$.

### 6.4 Time and memory requirements

Let us discuss the time and memory requirements of Algorithm 1, taking into account the previously mentioned implementation details.

The main objects in Algorithm 1, which need to be stored and updated between iterations, are the test points $x_k$, matrices $H_k$, scalars $R_k$, vectors $c_k$ and scalars $\sigma_k$, see (19), (55), 56 and (58) for the corresponding updating formulas. To store all these objects, we need $O(n^2)$ memory.

Consider now what happens at each iteration $k$. First, we compute $U_k$. For this, we calculate $z_k$ and $D_k$ according to (58) and then perform the calculations described in Sect. 6.2. The most difficult operation there is computing the matrix-vector product, which takes $O(n^2)$ time. After that, we calculate the coefficients $a_k$ and $b_k$ according to (26), where $\alpha_k$, $\theta$ and $\gamma$ are certain scalars, easily computable for all main instances of Algorithm 1 (see Sects. 4.1–4.4). The most expensive step there is computing the norm $\|g_k\|_{G_k}^*$, which can be done in $O(n^2)$ operations by evaluating the product $H_k g_k$. Finally, we update our main objects, which takes $O(n^2)$ time.

Thus, each iteration of Algorithm 1 has $O(n^2)$ time and memory complexities, exactly as in the standard Ellipsoid Method.

Now let us analyze the complexity of the auxiliary procedure from Sect. 5 for converting a preliminary semicertificate into a semicertificate. The main operation in this procedure is running Algorithm 2, which iterates "backwards", computing some dual multiplier $\mu_i$ at each iteration $i = k-1, \ldots, 0$. Using the approach from Sect. 6.3, we can compute $\mu_i$ in $O(n^2)$ time, provided that the objects $x_i, g_i, H_i, z_i, D_i, c_i, \sigma_i$ are stored in memory. Note, however, that, in contrast to the "forward" pass, when iterating "backwards", there is no way to efficiently recompute all these objects without storing in memory a certain "history" of the main process from iteration 0 up to $k$. The simplest choice is to keep in this "history" all the objects mentioned above, which requires $O(kn^2)$ memory. A slightly more efficient idea is to keep the matrix-vector products $H_i g_i$ instead of $H_i$ and then use (55) to recompute $H_i$ from $H_{i+1}$ in $O(n^2)$ operations. This allows us to reduce the size of the "history" down to $O(kn)$ while still keeping the $O(kn^2)$ total time complexity of the auxiliary procedure. Note that these estimates are exactly the same as those for the best currently known technique for generating accuracy certificates in the standard Ellipsoid Method [16]. In particular, if we generate a semicertificate only once at the very end, then the time complexity of our procedure is comparable to that of running the standard Ellipsoid Method without computing any certificates. Alternatively, as suggested in [16], one can generate semicertificates, say, every 2, 4, 8, 16, ... iterations. Then, the total "overhead" of the auxiliary procedure for generating semicertificates will be comparable to the time complexity of the method itself.

## 7 Conclusion

In this paper, we have addressed one of the issues of the standard Ellipsoid Method, namely, its poor convergence for problems of large dimension $n$. For this, we have

proposed a new algorithm which can be seen as the combination of the Subgradient and Ellipsoid methods.

Our developments can be considered as a first step towards constructing universal methods for nonsmooth problems with convex structure. Such methods could significantly improve the practical efficiency of solving various applied problems.

Note that there are still some open questions. First, the convergence estimate of our method with time-varying coefficients contains an extra factor proportional to the logarithm of the iteration counter. We have seen that this logarithmic factor has its roots yet in the Subgradient Method. However, as discussed in Remark 4, for the Subgradient Method, this issue can be easily resolved by allowing projections onto the feasible set and working with "truncated" gaps. An even better alternative, which does not require any of this machinery, is to use Dual Averaging [18] instead of the Subgradient Method. It is an interesting question whether one can combine the Dual Averaging with the Ellipsoid Method similarly to how we have combined the Subgradient and Ellipsoid methods.

Second, the convergence rate estimate, which we have obtained for our method, is not continuous in the dimension $n$. Indeed, for small values of the iteration counter $k$, this estimate behaves as that of the Subgradient Method and then, at some moment (around $n^2$), it switches to the estimate of the Ellipsoid Method. As discussed at the end of Sect. 4.4, there exists some "small" gap between these two estimates around the switching moment. Nevertheless, the method itself is continuous in $n$ and does not contain any explicit switching rules. Therefore, there should be some continuous convergence rate estimate for our method, and it is an open question to find it.

Another interesting question is to understand what happens with the proposed method on other (less general) classes of convex problems than those, considered in this paper. For example, it is well-known that, on smooth and/or strongly convex problems, (sub)gradient methods have much better convergence rates than on the general nonsmooth problems. We expect that similar conclusions should also be valid for the proposed Subgradient Ellipsoid Method. However, to achieve the acceleration, it may be necessary to introduce some modifications in the algorithm such as using different step sizes. We leave this direction for future research.

Finally, apart from the Ellipsoid Method, there exist other "dimension-dependent" methods (e.g., the Center-of-Gravity Method[8] [13,20], the Inscribed Ellipsoid Method [22], the Circumscribed Simplex Method [6], etc.). Similarly, the Subgradient Method is not the only "dimension-independent" method and there exist numerous alternatives which are better suited for certain problem classes (e.g., the Fast Gradient Method [17] for Smooth Convex Optimization or methods for Stochastic Programming [7,8,11,15]). Of course, it is interesting to consider different combinations of the aforementioned "dimension-dependent" and "dimension-independent" methods. In this regard, it is also worth mentioning the works [4,5], where the authors propose new variants of gradient-type methods for smooth strongly convex minimization problems inspired by the geometric construction of the Ellipsoid Method.

**Acknowledgements** We would like to thank the anonymous reviewers for their valuable time and efforts spent on reviewing this manuscript. Their feedback was very useful.

[8] Although this method is not practical, it is still interested from an academic point of view.

## A Proof of Lemma 6

***Proof*** Everywhere in the proof, we assume that the parameter $c$ is fixed and drop all the indices related to it.

Let us show that $\zeta_p$ is a convex function. Indeed, the function $\omega\colon \mathbb{R}\times\mathbb{R}_{++}\to\mathbb{R}$, defined by $\omega(x,t) := \frac{x^2}{t}$, is convex. Hence, the function $q$, defined in (34), is also convex. Further, since $\omega$ is increasing in its first argument on $\mathbb{R}_+$, the function $\omega_p\colon \mathbb{R}_+\times\mathbb{R}_{++}\to\mathbb{R}$, defined by $\omega_p(x,t) := \frac{x^p}{t}$, is also convex as the composition of $\omega$ with the mapping $(x,t)\mapsto(x^{p/2},t)$, whose first component is convex (since $p\geq 2$) and the second one is affine. Note that $\omega_p$ is increasing in its first argument. Hence, $\zeta_p$ is indeed a convex function as the composition of $\omega_p$ with the mapping $\gamma\mapsto\big(q(\gamma),1+\gamma\big)$, whose first part is convex and the second one is affine.

Differentiating, for any $\gamma>0$, we obtain

$$\zeta_p'(\gamma)=\frac{p[q(\gamma)]^{p-1}q'(\gamma)(1+\gamma)-[q(\gamma)]^p}{(1+\gamma)^2}=\frac{[q(\gamma)]^{p-1}\big(pq'(\gamma)(1+\gamma)-q(\gamma)\big)}{(1+\gamma)^2}.$$

Therefore, the minimizers of $\zeta_p$ are exactly solutions to the following equation:

$$pq'(\gamma)(1+\gamma)=q(\gamma). \tag{67}$$

Note that $q'(\gamma)=\frac{c[2\gamma(1+\gamma)-\gamma^2]}{2(1+\gamma)^2}=\frac{c\gamma(2+\gamma)}{2(1+\gamma)^2}$ (see (34)). Hence, (67) can be written as $cp\gamma(2+\gamma)=2(1+\gamma)+c\gamma^2$ or, equivalently, $c(p-1)\gamma^2+2(cp-1)\gamma=2$. Clearly, $\gamma=0$ is not a solution of this equation. Making the change of variables $\gamma=\frac{2}{u}$, $u\neq 0$, we come the quadratic equation $u^2-2(cp-1)u=2c(p-1)$ or, equivalently, to $[u-(cp-1)]^2=2c(p-1)+(cp-1)^2=c^2p^2-(2c-1)$. This equation has two solutions: $u_1:=cp-1+\sqrt{c^2p^2-(2c-1)}$ and $u_2:=cp-1-\sqrt{c^2p^2-(2c-1)}$. Note that $u_2\geq cp-1-\sqrt{c^2p^2+1}\geq cp-1-(cp+1)=-2$. Hence, $\gamma_2:=\frac{2}{u_2}\leq -1$ cannot be a minimizer of $\zeta_p$. Consequently, only $u_1$ is an acceptable solution (note that $u_1>0$ in view of our assumptions on $c$ and $p$). Thus, (37) is proved.

Let us show that $\gamma(p)$ belongs to the interval specified in (37). For this, we need to prove that $1\leq cp\gamma(p)\leq 2$. Note that the function $h_a(t):=\frac{t}{\sqrt{t^2-a}+t-1}$, where $a\geq 0$, is decreasing in $t$. Indeed, $\frac{1}{h_a(t)}=\sqrt{1-\frac{a}{t^2}}-\frac{1}{t}+1$ is an increasing function in $t$. Hence, $cp\gamma(p)=2h_{2c-1}(cp)\geq 2\lim_{t\to\infty}h_{2c-1}(t)=1$. On the other hand, using that $p\geq 2$ and denoting $\alpha:=2c\geq 1$, we get $cp\gamma(p)=2h_{\alpha-1}(cp)\leq 2g(\alpha)$, where

$g(\alpha) := h_{\alpha-1}(\alpha) = \frac{\alpha}{\sqrt{\alpha^2-\alpha+1}+\alpha-1}$. Note that $g$ is decreasing in $\alpha$. Indeed, denoting $\tau := \frac{1}{\alpha} \in (0, 1]$, we obtain $\frac{1}{g(\alpha)} = \sqrt{1-\tau+\tau^2} - \tau + 1$, which is a decreasing function in $\tau$. Thus, $cp\gamma(p) \le 2g(1) = 2$.

It remains to prove that $\zeta_p(\gamma(p)) \le e^{-1/(2cp)}$. Let $\phi\colon [2, +\infty) \to \mathbb{R}$ be the function

$$\phi(p) := -\ln \zeta_p\big(\gamma(p)\big) = \ln\big(1+\gamma(p)\big) - p\ln q\big(\gamma(p)\big). \tag{68}$$

We need to show that $\phi(p) \ge 1/(2cp)$ for all $p \ge 2$ or, equivalently, that the function $\chi\colon \big(0, \frac{1}{2}\big] \to \mathbb{R}$, defined by $\chi(\tau) := \phi(\frac{1}{\tau})$, satisfies $\chi(\tau) \ge \frac{\tau}{2c}$ for all $\tau \in \big(0, \frac{1}{2}\big]$. For this, it suffices to show that $\chi$ is convex, $\lim_{\tau\to 0}\chi(\tau) = 0$ and $\lim_{\tau\to 0}\chi'(\tau) = \frac{1}{2c}$. Differentiating, we see that $\chi'(\tau) = -\frac{1}{\tau^2}\phi'(\frac{1}{\tau})$ and $\chi''(\tau) = \frac{2}{\tau^3}\phi'(\frac{1}{\tau}) + \frac{1}{\tau^4}\phi''(\frac{1}{\tau})$ for all $\tau \in \big(0, \frac{1}{2}\big]$. Thus, we need to justify that

$$2\phi'(p) + p\phi''(p) \ge 0 \tag{69}$$

for all $p \ge 2$ and that

$$\lim_{p\to\infty}\phi(p) = 0, \qquad \lim_{p\to\infty}[-p^2\phi'(p)] = \frac{1}{2c}. \tag{70}$$

Let $p \ge 2$ be arbitrary. Differentiating and using (67), we obtain

$$\begin{aligned}
\phi'(p) &= \frac{\gamma'(p)}{1+\gamma(p)} - \ln q\big(\gamma(p)\big) - \frac{pq'\big(\gamma(p)\big)\gamma'(p)}{q\big(\gamma(p)\big)} = -\ln q\big(\gamma(p)\big),\\
\phi''(p) &= -\frac{q'(\gamma(p))\gamma'(p)}{q(\gamma(p))} = -\frac{\gamma'(p)}{p\big(1+\gamma(p)\big)}.
\end{aligned} \tag{71}$$

Therefore,

$$2\phi'(p) + p\phi''(p) = -2\ln q(\gamma(p)) - \frac{\gamma'(p)}{1+\gamma(p)} \ge -\frac{c\gamma^2(p)+\gamma'(p)}{1+\gamma(p)},$$

where the inequality follows from (34) and the fact that $\ln(1+\tau) \le \tau$ for any $\tau > -1$. Thus, to show (69), we need to prove that $-\gamma'(p) \ge c\gamma^2(p)$ or, equivalently, $\frac{d}{dp}\frac{1}{\gamma(p)} \ge c$. But this is immediate. Indeed, using (37), we obtain $\frac{d}{dp}\frac{1}{\gamma(p)} = \frac{c}{2}\left(\frac{cp}{\sqrt{c^2p^2-(2c-1)}}+1\right) \ge c$ since the function $\tau \mapsto \frac{\tau}{\sqrt{\tau^2-1}}$ is decreasing. Thus, (69) is proved.

It remains to show (70). From (37), we see that $\gamma(p) \to 0$ and $p\gamma(p) \to \frac{1}{c}$ as $p \to \infty$. Hence, using (34), we obtain

$$\lim_{p\to\infty} p^2\ln q\big(\gamma(p)\big) = \lim_{p\to\infty}\frac{cp^2\gamma^2(p)}{2\big(1+\gamma(p)\big)} = \frac{c}{2}\lim_{p\to\infty}p^2\gamma^2(p) = \frac{1}{2c}.$$

Consequently, in view of (68) and (71), we have

$$\lim_{p\to\infty} \phi(p) = \lim_{p\to\infty} \big[\ln\big(1+\gamma(p)\big) - p\ln q\big(\gamma(p)\big)\big] = 0,$$
$$\lim_{p\to\infty} [-p^2\phi'(p)] = \lim_{p\to\infty} p^2 \ln q\big(\gamma(p)\big) = \frac{1}{2c},$$

which is exactly (70). □

## B Support function and dual multipliers: proofs

For brevity, everywhere in this section, we write $\|x\|$ and $\|\cdot\|_*$ instead of $\|\cdot\|_{H^{-1}}$ and $\|\cdot\|_{H^{-1}}^*$, respectively. We also denote $B_0 := \{x \in \mathbb{E} : \|x\| \le 1\}$.

### B.1 Auxiliary operations

**Lemma 9** *Let $s \in \mathbb{E}^*$, let $A\colon \mathbb{R}^m \to \mathbb{E}^*$ be a linear operator with trivial kernel and let $b \in \mathbb{R}^m$, $\langle b, (A^*HA)^{-1}b\rangle < 1$. Then, problem (62) has a unique solution given by (63).*

***Proof*** Note that the sublevel sets of the objective function in (62) are bounded:

$$\|s-Au\|_*+\langle u, b\rangle \ge \|Au\|_*-\|s\|_*+\langle u, b\rangle \ge (1-\langle b, (A^*HA)^{-1}b\rangle^{1/2})\|Au\|_*-\|s\|_*$$

for all $u \in \mathbb{R}^m$. Hence, problem (62) has a solution.

Let $u \in \mathbb{R}^m$ be a solution of problem (62). If $s = Au$, then $u = (A^*HA)^{-1}A^*s$, which coincides with the solution given by (63) (note that, in this case, $r = 0$).

Now suppose $s \ne Au$. Then, from the first-order optimality condition, we obtain that $b = A^*(s - Au)/\rho$, where $\rho := \|s - Au\|_* > 0$. Hence, $u = (A^*HA)^{-1}(A^*s - \rho b)$ and

$$\begin{aligned}\rho^2 &= \|s - Au\|_*^2 = \|s\|_*^2 - 2\langle A^*s, u\rangle + \langle A^*HAu, u\rangle \\ &= \|s\|_*^2 - 2\langle A^*s, (A^*HA)^{-1}(A^*s - \rho b)\rangle + \langle A^*s - \rho b, (A^*HA)^{-1}(A^*s - \rho b)\rangle \\ &= \|s\|_*^2 - \langle s, A(A^*HA)^{-1}A^*s\rangle + \rho^2\langle b, (A^*HA)^{-1}b\rangle.\end{aligned}$$

Thus, $\rho = r$ and $u = u(H, s, A, b)$ given by (63). □

**Lemma 10** *Let $s, a \in \mathbb{E}^*$, $\beta \in \mathbb{R}$ be such that $\langle a, x\rangle \le \beta$ for some $x \in \operatorname{int} B_0$. Then, problem (60) has a solution given by (61). Moreover, this solution is unique if $\beta < \|a\|_*$.*

***Proof*** Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be the function $\phi(\tau) := \|s - \tau a\|_* + \tau\beta$. By our assumptions, $\beta > -\|a\|_*$ if $a \ne 0$ and $\beta \ge 0$ if $a = 0$. If additionally $\beta < \|a\|_*$, then $|\beta| < \|a\|_*$.

If $s = 0$, then $\phi(\tau) = \tau(\|a\|_* + \beta) \ge \phi(0)$ for all $\tau \ge 0$, so 0 is a solution of (60). Clearly, this solution is unique when $\beta < \|a\|_*$ because then $|\beta| < \|a\|_*$.

From now on, suppose $s \neq 0$. Then, $\phi$ is differentiable at 0 with $\phi'(0) = \beta - \langle a, s\rangle/\|s\|_*$. If $\langle a, s\rangle \leq \beta\|s\|_*$, then $\phi'(0) \geq 0$, so 0 is a solution of (60). Note that this solution is unique if $\langle a, s\rangle < \beta\|s\|_*$ because then $\phi'(0) > 0$, i.e., $\phi$ is strictly increasing on $\mathbb{R}_+$.

Suppose $\langle a, s\rangle > \beta\|s\|_*$. Then, $\beta < \|a\|_*$ and thus $|\beta| < \|a\|_*$. Note that, for any $\tau \geq 0$, we have $\phi(\tau) \geq \tau(\|a\|_* + \beta) - \|s\|_*$. Hence, the sublevel sets of $\phi$, intersected with $\mathbb{R}_+$, are bounded, so problem (60) has a solution. Since $\phi'(0) < 0$, any solution of (60) is strictly positive and so must be a solution of problem (62) for $A := a$ and $b := \beta$. But, by Lemma 9, the latter solution is unique and equals $u(H, s, a, \beta)$.

We have proved that (61) is indeed a solution of (60). Moreover, when $\langle a, s\rangle \neq \beta\|s\|_*$, we have shown that this solution is unique. It remains to prove the uniqueness of solution when $\langle a, s\rangle = \beta\|s\|_*$, assuming additionally that $\beta < \|a\|_*$. But this is simple. Indeed, by our assumptions, $|\beta| < \|a\|_*$, so $|\langle a, s\rangle| = |\beta|\|s\|_* < \|a\|_*\|s\|_*$. Hence, $a$ and $s$ are linearly independent. But then $\phi$ is strictly convex, and thus its minimizer is unique. □

### B.2 Computation of dual multipliers

In this section, we prove the correctness of Algorithm 3.

For $s \in \mathbb{E}^*$, let $X(s)$ be the subdifferential of $\|\cdot\|_*$ at the point $s$:

$$X(s) := \begin{cases} \{Hs/\|s\|_*\}, & \text{if } s \neq 0, \\ B_0, & \text{if } s = 0. \end{cases} \tag{72}$$

Clearly, $X(s) \subseteq B_0$ for any $s \in \mathbb{E}^*$. When $s \neq 0$, we denote the unique element of $X(s)$ by $x(s)$.

Let us formulate a convenient optimality condition.

**Lemma 11** *Let $A$ be the linear operator from $\mathbb{R}^m$ to $\mathbb{E}^*$, defined by $Au := \sum_{i=1}^m u_i a_i$, where $a_1, \ldots, a_m \in \mathbb{E}^*$, and let $b \in \mathbb{R}^m$, $s \in \mathbb{E}^*$. Then, $\mu^* \in \mathbb{R}^m_+$ is a minimizer of $\psi(\mu) := \|s - A\mu\|_* + \langle \mu, b\rangle$ over $\mathbb{R}^m_+$ if and only if $X(s - A\mu^*) \cap L_1(\mu_1^*) \ldots L_m(\mu_m^*) \neq \emptyset$, where, for each $1 \leq i \leq m$ and $\tau > 0$, we denote $L_i(\tau) := \{x \in \mathbb{E} : \langle a_i, x\rangle \leq b_i\}$, if $\tau = 0$, and $L_i(\tau) := \{x \in \mathbb{E} : \langle a_i, x\rangle = b_i\}$, if $\tau > 0$.*

***Proof*** Indeed, the standard optimality condition for a convex function over the nonnegative orthant is as follows: $\mu^* \in \mathbb{R}^m_+$ is a minimizer of $\psi$ on $\mathbb{R}^m_+$ if and only if there exists $g^* \in \partial\psi(\mu^*)$ such that $g_i^* \geq 0$ and $g_i^*\mu_i^* = 0$ for all $1 \leq i \leq m$. It remains to note that $\partial\psi(\mu^*) = b - A^*X(s - A\mu^*)$. □

**Theorem 3** *Algorithm 3 is well-defined and returns a solution of (65).*

***Proof*** i. For each $i = 1, 2$ and $\tau \geq 0$, denote $L_i^- := \{x \in \mathbb{E} : \langle a_i, x\rangle \leq b_i\}$, $L_i := \{x \in \mathbb{E} : \langle a_i, x\rangle = b_i\}$, $L_i(\tau) := L_i^-$, if $\tau = 0$, and $L_i(\tau) := L_i$, if $\tau > 0$.

ii. From (66) and Lemma 10, it follows that Step 1 is well-defined and, for each $i = 1, 2$, $\tau_i$ is a solution of (60) with parameters $(s, a_i, b_i)$. Hence, by Lemma 11,

$$X(s - \tau_i a_i) \cap L_i(\tau_i) \neq \emptyset, \quad i = 1, 2. \tag{73}$$

iii. Consider Step 2. Note that the condition $\xi_1 \le b_2$ is equivalent to $B_0 \cap L_1^- \subseteq L_2^-$ since $\xi_1 = \max_{x \in B_0 \cap L_1^-} \langle a_2, x \rangle$. If $B_0 \cap L_1^- \subseteq L_2^-$, then, by (73), $X(s - \tau_1 a_1) \cap L_1(\tau_1) \cap L_2^- = X(s - \tau_1 a_1) \cap L_1(\tau_1) \ne \emptyset$, so, by Lemma 11, $(\tau_1, 0)$ is indeed a solution of (65).

Similarly, if $\xi_2 \le b_1$, then $B_0 \cap L_2^- \subseteq L_1^-$ and $(0, \tau_2)$ is a solution of (65).

iv. From now on, we can assume that $B_0 \cap L_1^- \cap \operatorname{int} L_2^+ \ne \emptyset$, $B_0 \cap L_2^- \cap \operatorname{int} L_1^+ \ne \emptyset$, where $\operatorname{int} L_i^+ := \{x \in \mathbb{E} : \langle a_i, x \rangle > b_i\}$, $i = 1, 2$. Combining this with (66), we obtain[9]

$$\operatorname{int} B_0 \cap L_1 \cap L_2^- \ne \emptyset, \qquad \operatorname{int} B_0 \cap L_2 \cap L_1^- \ne \emptyset. \tag{74}$$

Suppose $\langle a_2, H(s - \tau_1 a_1) \rangle \le b_2 \|s - \tau_1 a_1\|_*$ at Step 3. 1) If $s \ne \tau_1 a_1$, then $X(s - \tau_1 a_1)$ is a singleton, $x(s - \tau_1 a_1) = H(s - \tau_1 a_1)/\|s - \tau_1 a_1\|_*$, so we obtain $x(s - \tau_1 a_1) \in L_2^-$. Combining this with (73), we get $x(s - \tau_1 a_1) \in L_1(\tau_1) \cap L_2^-$. 2) If $s = \tau_1 a_1$, then $X(s - \tau_1 a_1) \cap L_1(\tau_1) \cap L_2^- = B_0 \cap L_1(\tau_1) \cap L_2^- \ne \emptyset$ in view of the first claim in (74) (recall that $L_1 \subseteq L_1(\tau_1)$). Thus, in any case, $X(s - \tau a_1) \cap L_1(\tau_1) \cap L_2^- \ne \emptyset$, and so, by Lemma 11, $(\tau_1, 0)$ is a solution of (65).

Similarly, one can consider the case when $\langle a_1, H(s - \tau_2 a_2) \rangle \le b_1 \|s - \tau_2 a_2\|_*$ at Step 3.

Suppose we have reached Step 4. From now on, we can assume that

$$X(s - \tau_1 a_1) \cap L_1(\tau_1) \cap \operatorname{int} L_2^+ \ne \emptyset, \qquad X(s - \tau_2 a_2) \cap L_2(\tau_2) \cap \operatorname{int} L_1^+ \ne \emptyset. \tag{75}$$

Indeed, since both conditions at Step 3 have not been satisfied, $s \ne \tau_i a_i$, $i = 1, 2$, and $x(s - \tau_1 a_1) \notin L_2^-$, $x(s - \tau_2 a_2) \notin L_1^-$. Also, by (73), $x(s - \tau_i a_i) \in L_i(\tau_i)$, $i = 1, 2$.

Let $\mu \in \mathbb{R}^2_+$ be any solution of (65). By Lemma 11, $X(s - A\mu) \cap L_1(\mu_1) \cap L_2(\mu_2) \ne \emptyset$. Note that we cannot have $\mu_2 = 0$. Indeed, otherwise, we get $X(s - \mu_1 a_1) \cap L_1(\mu_1) \cap L_2^- \ne \emptyset$, so $\mu_1$ must be a solution of (60) with parameters $(s, a_1, b_1)$. But, by Lemma 10, such a solution is unique (in view of the second claim in (75), $\langle a_1, x \rangle > b_1$ for some $x \in B_0$, so $b_1 < \|a_1\|_*$). Hence, $\mu_1 = \tau_1$, and we obtain a contradiction with (75). Similarly, we can show that $\mu_1 \ne 0$. Consequently, $\mu_1, \mu_2 > 0$, which means that $\mu$ is a solution of (62).

Thus, at this point, any solution of (65) must be a solution of (62). In view of Lemma 9, to finish the proof, it remains to show that the vectors $a_1$, $a_2$ are linearly independent and $\langle b, (A^* H A)^{-1} b \rangle < 1$. But this is simple. Indeed, from (75), it follows that

$$\text{either} \quad B_0 \cap L_1 \cap \operatorname{int} L_2^+ \ne \emptyset \quad \text{or} \quad B_0 \cap L_2 \cap \operatorname{int} L_1^+ \ne \emptyset \tag{76}$$

since $\tau_1$ and $\tau_2$ cannot both be equal to 0. Combining (76) and (74), we see that $\operatorname{int} B_0 \cap L_1 \cap L_2 \ne \emptyset$ and, in particular, $L_1 \cap L_2 \ne \emptyset$. Hence, $a_1$, $a_2$ are linearly independent (otherwise, $L_1 = L_2$, which contradicts (76)). Taking any $x \in \operatorname{int} B_0 \cap L_1 \cap L_2$, we

[9] Take an appropriate convex combination of two points from the specified nonempty convex sets.

obtain $\|x\| < 1$ and $A^*x = b$, hence $\langle b, (A^*HA)^{-1}b\rangle = \langle A^*x, (A^*HA)^{-1}A^*x\rangle \leq \|x\|^2 < 1$, where we have used $A(A^*HA)^{-1}A^* \preceq H^{-1}$. □

## References


1. Auslender, A.: Résolution numérique d'inégalités variationnelles. RAIRO **7**(2), 67–72 (1973)
2. Ben-Tal, A., Nemirovski, A.: Lectures on modern convex optimization. Lecture notes (2021)
3. Bland, R., Goldfarb, D., Todd, M.: The ellipsoid method: a survey. Oper. Res. **29**(6), 1039–1091 (1981)
4. Bubeck, S., Lee, Y.T.: Black-box optimization with a politician. In: International Conference on Machine Learning, pp. 1624–1631. PMLR (2016)
5. Bubeck, S., Lee, Y.T., Singh, M.: A geometric alternative to Nesterov's accelerated gradient descent. arXiv preprint arXiv:1506.08187 (2015)
6. Bulatov, V., Shepot'ko, L.: Method of centers of orthogonal simplexes for solving convex programming problems. Methods Optim. Appl. (1982)
7. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**(7) (2011)
8. Dvurechensky, P., Gasnikov, A.: Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. J. Optim. Theory Appl. **171**(1), 121–145 (2016)
9. Grötschel, M., Lovász, L., Schrijver, A.: The ellipsoid method and its consequences in combinatorial optimization. Combinatorica **1**(2), 169–197 (1981)
10. Khachiyan, L.: A polynomial algorithm in linear programming. Soviet Math. Dokl. **244**(5), 1093–1096 (1979)
11. Lan, G.: An optimal method for stochastic composite optimization. Math. Program. **133**(1), 365–397 (2012)
12. Lan, G.: First-Order and Stochastic Optimization Methods for Machine Learning. Springer, Switzerland (2020)
13. Levin, A.: An algorithm for minimizing convex functions. Soviet Math. Dokl. **160**(6), 1244–1247 (1965)
14. Nemirovski, A.: Information-based complexity of convex programming. Lecture notes (1995)
15. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2009)
16. Nemirovski, A., Onn, S., Rothblum, U.G.: Accuracy certificates for computational problems with convex structure. Math. Oper. Res. **35**(1), 52–78 (2010)
17. Nesterov, Y.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$. Soviet Math. Dokl. **269**, 543–547 (1983)
18. Nesterov, Y.: Primal-dual subgradient methods for convex problems. Math. Program. **120**(1), 221–259 (2009)
19. Nesterov, Y.: Lectures on convex optimization. Springer, Berlin (2018)
20. Newman, D.: Location of the maximum on unimodal surfaces. J. ACM (JACM) **12**(3), 395–398 (1965)
21. Shor, N.: Cut-off method with space extension in convex programming problems. Cybernetics **13**(1), 94–96 (1977)
22. Tarasov, S., Khachiyan, L., Erlikh, I.: The method of inscribed ellipsoids. Soviet Math. Dokl. **37**(1), 226–230 (1988)
23. Yudin, D., Nemirovskii, A.: Informational complexity and efficient methods for the solution of convex extremal problems. Matekon **13**(2), 22–45 (1976)