FULL LENGTH PAPER

Series A

Efficient first-order methods for convex minimization: a constructive approach

Yoel Drori[1] · Adrien B. Taylor[2]

Abstract

We describe a novel constructive technique for devising efficient first-order methods for a wide range of large-scale convex minimization settings, including smooth, non-smooth, and strongly convex minimization. The technique builds upon a certain variant of the conjugate gradient method to construct a family of methods such that (a) all methods in the family share the same worst-case guarantee as the base conjugate gradient method, and (b) the family includes a fixed-step first-order method. We demonstrate the effectiveness of the approach by deriving optimal methods for the smooth and non-smooth cases, including new methods that forego knowledge of the problem parameters at the cost of a one-dimensional line search per iteration, and a universal method for the union of these classes that requires a three-dimensional search per iteration. In the strongly convex case, we show how numerical tools can be used to perform the construction, and show that the resulting method offers an improved worst-case bound compared to Nesterov's celebrated fast gradient method.

Yoel Drori and Adrien B. Taylor have contributed equally to this work.

✉ Adrien B. Taylor
  adrien.taylor@inria.fr

  Yoel Drori
  dyoel@google.com

[1] Google LLC, 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA

[2] INRIA, Département d'Informatique de l'ENS, École Normale Supérieure, CNRS, PSL Research University, Paris, France

1 Introduction

Convex optimization plays a central role in many fields of applications, including optimal control, machine learning and signal processing. In particular, when a large number of variables are involved within a convex optimization problem, the use of first-order methods is more and more widespread due to their typically very attractive low computational cost per iteration. This low computational cost comes, however, at a price: first-order methods often suffer from potentially slow convergence speeds, making them appropriate mostly for obtaining low to medium accuracy solutions. Nevertheless, first-order methods remain the methods of choice in many applications and currently receive a lot of attention from the optimization community, which constantly aims at improving them.

An effective and fruitful approach used for analyzing and comparing first-order methods is the study of their worst-case behavior through the black-box model. In this setting, methods are only allowed to gain information on the objective through an oracle, which provides the value and the gradient of the objective at selected points. Historically, this approach was largely motivated by the seminal work of Nemirovski and Yudin [34], and later by the work of Nesterov [36]. These works established lower and upper bounds on the worst-case performances of first-order methods on several important classes of problems and initiated the search for optimal algorithms, which exhibit the best possible worst-case performances (up to a constant factor) for the class of problems they were designed to solve.

In this work, we consider the generic task of designing first-order methods for convex minimization. The suggested approach starts from a conceptual method that does not have an efficient implementation. Then, we show, from the analysis of this method, that one can construct efficient implementations that benefits from the same worst-case guarantees. This design approach has been considered several times in the past, see [25,32] and many more. Here, unlike the alluded works, the chosen conceptual method is a very fundamental method capable of handling diverse families of problems, making the design approach applicable to a variety of settings. The conceptual algorithm we choose is a variant of the conjugate-gradient method, whose analysis therefore occupies an important place in the sequel.

1.1 Related works

A large number of generic techniques for developing optimization methods were proposed in the past years. Below we give a short overview of such techniques tailored for convex optimization; we do not attempt to give a comprehensive list.

1.1.1 Links with subspace-search methods

One core idea underlying several classical optimization algorithms, and also strongly related to the technique proposed below, is the use of subspace-searches. Among them, conjugate gradient methods (see e.g., [18,55]), which can be seen as methods performing minimization steps on increasingly larger subspaces, have a prominent place.

Related to that, the original optimal methods for smooth convex minimization, developed by Nemirovski [30] and Nemirovski and Yudin [33] (see e.g., the review in [29]), requires two and three-dimensional subspace minimization at each iteration and is therefore reminiscent of subspace-search methods. For smooth convex minimization, those methods can be seen as predecessors to the celebrated Nesterov's fast gradient method [35], which achieves the same *optimal* convergence rate without relying on those exact subspace minimizations steps. Motivated by similarities between the subspace-search methods of Nemirovski and Yudin [30,33] and Nesterov's fast gradient method [35], we propose a generic technique which transforms subspace-search Conjugate Gradient-like methods to fixed-step methods that have equal or better worst-case performances. This technique was also premised in two previous works by the authors:

- In [12, Remark 3.1], Drori remarked that the lower bound for smooth convex unconstrained minimization was achieved by a greedy method (referenced to as the *ideal first-order method*),
- In [6, Section 4.1], de Klerk et al. study the worst-case complexity of steepest descent with exact line search applied to strongly convex functions. As suggested by an anonymous referee in [6], the worst-case certificates were also valid for the gradient method with an appropriate fixed-step size.

Links between fast gradient methods and conjugate gradient methods were also recently analyzed in [21].

1.1.2 Links with fast gradient methods

Fast gradient schemes for minimizing smooth convex and smooth strongly convex functions originated in the seminal works of Nesterov [35,36] (fast gradient methods), Polyak [40,41] (heavy-ball method), Nemirovski [30] and Nemirovski and Yudin [33]. Despite its fundamental nature, acceleration remained an obscure phenomenon relying on an algebraic trick for years, and many authors have recently developed new explanations for this behavior. For smooth strongly convex minimization, recent popular works include geometrical approaches such as a shrinking ball scheme [5], a new method based on lower quadratic approximations [15], analyses relying on stability theory for discrete-time and/or continuous-time dynamical systems [16,26,53] (specifically for fast gradient in [19]), or even as a specific integration scheme of the gradient flow [45]. In the context of smooth convex minimization, another recent trend include parallels with differential equations [46]—some of the previous approaches for the smooth strongly convex case also apply without strong convexity.

1.1.3 Links with systematic and computer-assisted approaches to worst-case analyses

This work takes place within the current effort for the development of systematic/computer-guided analyses and design of optimization algorithms. Among them, a systematic approach to lower bounds (which focuses on quadratic cases) is presented in by Arjevani et al. [1], a systematic use of control theory (via integral quadratic

constraints) for developing upper bounds is presented by Lessard et al. [26], and the performance estimation approach, which aims at finding worst-case bounds was originally developed in [13] (see also surveys in [11] and [47]).

Those methodologies are mostly presented as tools for performing worst-cases analyses (see the numerous examples in [11,19,47,48,50,51]), however, such techniques were also recently used to develop new methods with improved worst-case complexities. Among others, such an approach was used in [13,22] to devise a fixed-step method that attains the best possible worst-case performance for smooth convex minimization [12], and later in [14] to obtain a variant of Kelley's cutting plane method with the best possible worst-case guarantee for non-smooth convex minimization. Also, the control-theoretic approach presented by Lessard et al. [26] was used in [52] for developing a new accelerated method for smooth strongly convex minimization, called the triple momentum method.

1.2 Paper organization and main contributions

The paper is organized as follows. First, Sect. 2 introduces elementary facts and definitions that are used throughout this work. Then, Sect. 3 presents a specific variant of the conjugate gradient method, which we refer to as the *Greedy First-Order Method* (GFOM), along with the corresponding tools for analyzing it. Following that, Sect. 4 proposes a procedure for constructing fixed-step first-order methods that benefits from the same worst-case guarantees as that of GFOM. The procedure is applied on the class of non-smooth and smooth convex functions, and is shown to produce families of first-order methods achieving the best-possible worst-case bounds in both settings. Section 5 is devoted the strongly-convex case, where no analytical solution is known to the problem that arises from the design procedure. We show that the resulting numerically-defined algorithm attains both an efficient implementation and an improved worst-case bounds as compared to a standard fast gradient method of Nesterov [36, Section 2.2]. Finally, we conclude and discuss extensions in Sect. 6.

1.3 Notations

Consider the convex minimization problem

$$f_* := \min_{x \in \mathbb{R}^d} f(x), \tag{OPT}$$

with $f \in \mathcal{F}(\mathbb{R}^d)$, for some class $\mathcal{F}(\mathbb{R}^d)$ of closed, convex and proper (c.c.p.) functions $f : \mathbb{R}^d \to \mathbb{R}$. For notational convenience, we use the notation $f \in \mathcal{F}$ when the dimension d is left unspecified, the notation val (OPT) when referring to the optimal value of the problem f_*, and by x_* to denote an element in argmin f.

Additionally, we denote by $x_i \in \mathbb{R}^d$ the iterates produced by the different optimization schemes, and use $f'(x_i)$ to denote an element in the subdifferential $\partial f(x_i)$. When $g_i \in \mathbb{R}^d$ is an arbitrary vector, we use the standard notation $g_i \in \partial f(x_i)$ to specify the requirement that g_i is a subdifferential of f at x_i. The set $\{x_i\}_{i \in I_N^*}$ containing the

first N iterates and an optimal point is often used, where the index set I_N^* is defined as follows:

$$I_N^* := \{*, 0, \ldots, N\}. \tag{1}$$

Additionally, we use the standard notation $\langle \cdot, \cdot \rangle : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ to denote the Euclidean inner product, and the corresponding induced norm $\|\cdot\|$. Given a positive semidefinite matrix $A \succeq 0$, we also use the notation $\langle \cdot, \cdot \rangle_A = \langle A\cdot, \cdot \rangle$ and the corresponding induced semi-norm $\|\cdot\|_A$. Finally, we use the notation $(\cdot \odot \cdot) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ to denote the symmetric outer product, that is, for any $x, y \in \mathbb{R}^d$:

$$x \odot y = \frac{1}{2} \left(xy^\top + yx^\top \right),$$

resulting in the following useful identity: $\langle x, y \rangle_A = \mathrm{Tr}\left(A(x \odot y) \right)$.

2 Basic definitions

This section briefly introduces some definitions that we use in the forthcoming analyses. We start by introducing, for the sake of convenience, a shorthand notation for the set of inputs expected by the algorithms under consideration. We continue by introducing a notation allowing us to implicitly define a function based on some local first-order information.

Definition 1 A pair (f, x_0) is called an $(\mathcal{F}(\mathbb{R}^d), R_x)$-*input* if $\mathcal{F}(\mathbb{R}^d)$ is a class of closed, convex and proper (c.c.p.) functions over $\mathbb{R}^d$, R_x is a nonnegative constant, $f \in \mathcal{F}(\mathbb{R}^d)$, $x_0 \in \mathbb{R}^d$, and $\|x_0 - x_*\| \leq R_x$ holds for some $x_* \in \mathrm{argmin}\, f(x)$.

Definition 2 Let $\mathcal{F}$ be a class of c.c.p. functions. A set of triplets $S = \{(x_i, g_i, f_i)\}_{i \in I}$ (for some index set I) is called $\mathcal{F}$-*interpolable* if there exists a function $f \in \mathcal{F}$ such that $g_i \in \partial f(x_i)$ and $f_i = f(x_i)$ for all $i \in I$.

For many classes of functions $\mathcal{F}$, the condition "S is $\mathcal{F}$-interpolable" can be expressed as a finite set of constraints on the elements of S. In these cases, we refer to this set of constraints as *interpolation conditions* for the class $\mathcal{F}$. See [48,50] for a list of known interpolation conditions for different classes of functions, along with the corresponding proofs.

For the sake of completeness, we include below the interpolation conditions for the class of strongly-convex smooth functions and for the class of non-smooth convex functions. These classes are used in the examples presented in Sects. 4 and 5.

Theorem 1 ([50, Theorem 4]) *Let I be a finite index set and let $\mathcal{F}_{\mu,L}$ denote the set of L-smooth and μ-strongly convex functions for some $0 \leq \mu < L \leq \infty$. A set $\{(x_i, g_i, f_i)\}_{i \in I}$ is $\mathcal{F}_{\mu,L}$-interpolable if and only if for all $i, j \in I$*

$$f_i \geq f_j + \langle g_j, x_i - x_j \rangle + \frac{1}{2(1 - \mu/L)}$$
$$\cdot \left(\frac{1}{L} \|g_i - g_j\|^2 + \mu \|x_i - x_j\|^2 - 2\frac{\mu}{L} \langle x_i - x_j, g_i - g_j \rangle \right). \qquad (2)$$

Theorem 2 ([48, Theorem 4]) *Let I be a finite index set and let $\mathcal{C}_M$ denote the set of Lipschitz c.c.p. functions whose gradient is bounded in norm by M for some $0 \leq M \leq \infty$. A set $\{(x_i, g_i, f_i)\}_{i \in I}$ is $\mathcal{C}_M$-interpolable if and only if for all $i, j \in I$*

$$f_i \geq f_j + \langle g_j, x_i - x_j \rangle,$$
$$\|g_i\| \leq M. \qquad (3)$$

Finally, we introduce the following technical property, which will be heavily used in establishing tightness results in the sequel.

Definition 3 A class of functions $\mathcal{F}(\mathbb{R}^d)$ is said to be *contraction-preserving* if for any $\mathcal{F}$-interpolable set $S = \{(x_i, g_i, f_i)\}_{i \in I}$, where I is some finite index set, and any $\{\hat{x}_i\}_{i \in I} \subset \mathbb{R}^d$ satisfying

$$\|\hat{x}_i - \hat{x}_j\| \leq \|x_i - x_j\|, \text{ and } \langle g_j, \hat{x}_i - \hat{x}_j \rangle = \langle g_j, x_i - x_j \rangle, \quad \forall i, j \in I, \qquad (4)$$

we have that $\hat{S} = \{(\hat{x}_i, g_i, f_i)\}_{i \in I}$ is $\mathcal{F}$-interpolable.

Two important examples of contraction preserving classes were discussed above, namely the class of smooth (possibly strongly) convex functions and the class of non-smooth convex functions.

Proposition 1 *The class of L-smooth, μ-strongly convex functions $\mathcal{F}_{\mu,L}$ with $0 \leq \mu < L \leq \infty$ is contraction preserving.*

Proof Let I be some index set, let $S = \{(x_i, g_i, f_i)\}_{i \in I}$ be a $\mathcal{F}_{\mu,L}$-interpolable set, and suppose $\hat{S} = \{(\hat{x}_i, g_i, f_i)\}_{i \in I}$ with $\hat{x}_i$ satisfying (4). Then from Theorem 1, we have $\forall i, j \in I$:

$$f_i \geq f_j + \langle g_j, x_i - x_j \rangle + \frac{1}{2(1 - \mu/L)}$$
$$\cdot \left(\frac{1}{L} \|g_i - g_j\|^2 + \mu \|x_i - x_j\|^2 - 2\frac{\mu}{L} \langle x_i - x_j, g_i - g_j \rangle \right),$$
$$\geq f_j + \langle g_j, \hat{x}_i - \hat{x}_j \rangle + \frac{1}{2(1 - \mu/L)}$$
$$\cdot \left(\frac{1}{L} \|g_i - g_j\|^2 + \mu \|\hat{x}_i - \hat{x}_j\|^2 - 2\frac{\mu}{L} \langle \hat{x}_i - \hat{x}_j, g_i - g_j \rangle \right),$$

hence $\hat{S}$ is $\mathcal{F}_{\mu,L}$-interpolable, as those inequalities are necessary and sufficient for smooth strongly convex interpolation (see Theorem 1). □

Proposition 2 *The class of c.c.p. functions with M-bounded gradients $\mathcal{C}_M$ with $0 \leq M \leq \infty$ is contraction preserving.*

Proof Let $S = \{(x_i, g_i, f_i)\}_{i \in I}$ be a $\mathcal{C}_M$-interpolable set and suppose $\hat{S} = \{(\hat{x}_i, g_i, f_i)\}_{i \in I}$ with $\hat{x}_i$ satisfying (4). Since the interpolation conditions (3) hold for S, it immediately follows from the equality relations in (4) that these interpolation conditions also hold for $\hat{S}$, concluding the proof. $\qquad\square$

3 Analysis of a greedy first-order method

The goal of this section is to introduce a framework for studying the worst-case performance of the following subspace-search based greedy method, reminiscent of conjugate gradient methods.

Greedy first-order method (GFOM)

Input: $f \in \mathcal{F}(\mathbb{R}^d)$, initial guess $x_0 \in \mathbb{R}^d$, $N \in \mathbb{N}$.

For $i = 1, 2, \ldots, N$:

$$\text{Set } x_i \in \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ f(x) : x \in x_0 + \operatorname{span}\{f'(x_0), \ldots, f'(x_{i-1})\} \right\}, \tag{5}$$

$$\text{Choose } f'(x_i) \in \partial f(x_i) \text{ such that } \langle f'(x_i), f'(x_j) \rangle = 0 \quad \text{for all } 0 \leq j < i, \tag{6}$$

Output: $\text{GFOM}_N(f, x_0) := x_N$.

GFOM clearly becomes intractable after the first few iterations; nevertheless, a few fundamental theoretical properties render its analysis interesting independently of the focus of the following sections. One such property is that GFOM attains the best possible behavior that can be achieved by a first-order method on functions that have a form similar to the "worst function in the world" introduced by Nesterov [36] (as a way of establishing lower-complexity bounds). GFOM is therefore a natural candidate when looking for "the best algorithm in the world". Additionally, GFOM can be seen as a generalization of the Conjugate Gradient method [18] which remains a very fruitful field of study to this day.

Note that the iteration rule (5) is not well-defined when the function f does not attain its infimum on the provided subspace. In such cases, GFOM cannot proceed and we say that it does not yield an output. For cases where (5) is well-defined, note that by the first-order optimality conditions, a choice for $f'(x_i)$ that satisfies the (Conjugate Gradient like) conditions (6) necessarily exists, and in particular, when f is differentiable at x_i these requirements are fulfilled by the gradient of f at x_i.

3.1 Estimating the worst-case performance of GFOM

The analysis below is based on the performance estimation methodology which was first introduced in [13] and has been successfully applied to analyze methods in a wide range of settings, including smooth and nonsmooth minimization [14,22,23,50], proximal gradient methods [48,51], saddle-point problems [11] and more recently to operator splitting methods [43]. Here we build upon an approach developed for the analysis of line-searching methods [6], and improve it by providing a tightness proof under some mild conditions.

   Clearly, a meaningful analysis can only be attained by making some assumptions on the structure of the problem: namely, that f belongs to some given class of functions $\mathcal{F}$ and that the initial point x_0 satisfies some conditions. In the sequel we restrict our attention to the standard initial condition on $\|x_0 - x_*\|$, which we assume to be bounded by some constant $R_x > 0$ (see Definition 1). In addition, we evaluate the performance of a method in terms of its worst-case absolute inaccuracy $f(x_N) - f_*$. In other words, we are looking for worst-case guarantees of type

$$f(x_N) - f_* \leq \tau \|x_0 - x_*\|^2,$$

with $\tau \geq 0$ being as small as possible. Note that the presented analysis allows for more general initial conditions and performance measures (see discussions in [31, Section 4], and more specifically in [51, Section 4] in the context of performance estimation), however, for the sake of simplicity we do not pursue this direction in this work.

   We start the analysis of GFOM with the observation that, under the assumptions discussed above, the worst-case performance of GFOM is *by definition* the optimal value to the following *performance estimation problem* (PEP):

$$\sup_{f, x_0} f(\text{GFOM}_N(f, x_0)) - f_* \tag{PEP}$$

$$\text{subject to: } (f, x_0) \text{ is an } (\mathcal{F}, R_x)\text{-input.}$$

   As an immediate consequence of the definition of (PEP), if (f, x_0) is such that $f \in \mathcal{F}$ and $\|x_0 - x_*\| \leq R_x$ (for some $x_* \in \text{argmin}_x f(x)$) then the following bound hold:

$$f(\text{GFOM}_N(f, x_0)) - f_* \leq \text{val}(\text{PEP}). \tag{7}$$

   Although the form (PEP) appears at first to be a purely theoretical notational reformulation, it provides a convenient framework for manipulating the problem in a way that will eventually yield a tractable bound on the worst-case performance of GFOM.

   As a first step for obtaining tractable bounds we formulate (PEP) as a finite-dimensional optimization problem.

Lemma 1 *Let $N \in \mathbb{N}$, $R_x \geq 0$ and let $\mathcal{F}$ be a class of c.c.p. functions. Then for any $(\mathcal{F}, R_x)$-input (f, x_0),*

$$f(\text{GFOM}_N(f, x_0)) - f_* \leq \text{val}(\text{PEP-GFOM}) \tag{8}$$

holds with

$$\sup_{\{(x_i, g_i, f_i)\}_{i \in I_N^*}} f_N - f_* \qquad \text{(PEP-GFOM)}$$

subject to: $\{(x_i, g_i, f_i)\}_{i \in I_N^*}$ *is* $\mathcal{F}$-*interpolable,*

$$\|x_0 - x_*\| \leq R_x,$$
$$g_* = 0,$$
$$\langle g_i, g_j \rangle = 0, \quad \text{for all } 0 \leq j < i = 1, \ldots N,$$
$$\langle g_i, x_j - x_0 \rangle = 0, \quad \text{for all } 1 \leq j \leq i = 1, \ldots N.$$

Furthermore, if $\mathcal{F}$ *is contraction-preserving (see Definition* 3*) then the bound* (8) *is tight, i.e., for any* $\varepsilon > 0$ *there exists an* $(\mathcal{F}, R_x)$-*input* $(\hat{f}, \hat{x}_0)$*, such that*

$$\hat{f}(\text{GFOM}_N(\hat{f}, \hat{x}_0)) - \hat{f}_* \geq \text{val}\,(\text{PEP-GFOM}) - \varepsilon.$$

Problem (PEP-GFOM) can be seen as a discretized version of (PEP), where the variable f_i acts as the value of the function at x_i and the variable g_i acts as the gradient of the function at x_i. In order to ensure that x_* corresponds to an optimal value of f, the constraint $g_* = 0$ is included. In view of this interpretation, all the constraints in (PEP-GFOM) are clearly necessary and follow directly from the properties of the problem and from the definition of GFOM, making (PEP-GFOM) a relaxation of (PEP) and therefore an upper bound on its value. The main issue in establishing the tightness claim is to show that the variables x_i fall in the span of the previous gradients [i.e., $x_i \in x_0 + \text{span}\{g_0, \ldots, g_{i-1}\}$ as required by (5)]; here the contraction-preserving assumption is used, allowing to "project" the variable x_i on the span of the previous gradients without affecting the constraints or objective, thus showing that an optimal solution can be assumed to satisfy the required property (5). We postpone the formal proof of Lemma 1 to Appendix 1.

In surprisingly many situations, the constraint '$\{(x_i, g_i, f_i)\}_{i \in I_N^*}$ is $\mathcal{F}$-interpolable' in (PEP-GFOM) can be expressed as a finite set of inequalities that depend on the $\{(x_i, g_i, f_i)\}_{i \in I_N^*}$ variables via linear combinations of $\{f_i\}_{i \in I_N^*}$ and inner products of the vectors $\{(x_i, g_i)\}_{i \in I_N^*}$ [for example, see (2) and (3) for the $\mathcal{F}_{\mu,L}$ and $\mathcal{C}_M$ classes, respectively]. In these cases, (PEP-GFOM) becomes a quadratically constrained quadratic program (QCQP), which has efficient SDP relaxations [2,3]. We now consider such cases, and provide sufficient conditions under which these relaxations are exact.

3.2 A tractable bound on the worst-case performance of GFOM

We begin by introducing the following notations: given a set of triplets $S = \{(x_i, g_i, f_i)\}_{i \in I_N^*}$ with $g_* = 0$, let $P \in \mathbb{R}^{d \times (2N+2)}$ and $F \in \mathbb{R}^{N+2}$ be defined as containing the information collected after N iterations as follows:

$$\begin{aligned} P &= [\, g_0 \, g_1 \, \ldots \, g_N \mid x_1 - x_0 \, \ldots \, x_N - x_0 \mid x_* - x_0 \,], \\ F &= [\, f_0 \, f_1 \, \ldots \, f_N \, f_* \,]^\top. \end{aligned} \qquad (9)$$

Further denote by $G \in \mathbb{R}^{(2N+2)\times(2N+2)}$ the corresponding positive semidefinite Gram matrix

$$G = P^\top P \succeq 0,$$

and by $\mathbf{x}_i, \mathbf{g}_i \in \mathbb{R}^{2N+2}$ and $\mathbf{f}_i \in \mathbb{R}^{N+2}$ the following zero and unit vectors

$$
\begin{aligned}
\mathbf{x}_0 &:= 0, \\
\mathbf{x}_i &:= e_{N+1+i}, & i &= 1, \ldots, N, \\
\mathbf{x}_* &:= e_{2N+2}, \\
\mathbf{g}_* &:= 0, \\
\mathbf{g}_i &:= e_{i+1}, & i &= 0, \ldots, N, \\
\mathbf{f}_i &:= e_{i+1}, & i &= 0, \ldots, N, \\
\mathbf{f}_* &:= e_{N+2},
\end{aligned}
$$

which are defined such that $x_i - x_0 = P(\mathbf{x}_i - \mathbf{x}_0)$, $g_i = P\mathbf{g}_i$ and $f_i = F\mathbf{f}_i$. Using these notations together with the notations defined in Sect. 1.3, the following reformulations hold:

$$
\begin{aligned}
\langle g_i, g_j \rangle &= \langle \mathbf{g}_i, \mathbf{g}_j \rangle_G = 0, \\
\langle g_i, x_j - x_0 \rangle &= \langle \mathbf{g}_i, \mathbf{x}_j - \mathbf{x}_0 \rangle_G = 0, \\
\|x_0 - x_*\|^2 &= \|\mathbf{x}_0 - \mathbf{x}_*\|_G^2.
\end{aligned}
\tag{10}
$$

These notations allow us to encode the equality and inequality constraints in (PEP-GFOM) within an SDP. In order to encode the interpolation conditions, we further require the class $\mathcal{F}(\mathbb{R}^d)$ to have interpolation conditions that can be expressed as a set of affine constraints in the entries of the matrices G and F defined above.

Assumption 1 The constraint '$\{(x_i, g_i, f_i)\}_{i \in I_N^*}$ is $\mathcal{F}(\mathbb{R}^d)$-interpolable' can be encoded within an SDP. I.e., there exists an index set K_N and an appropriate choice of matrices $A_k^{ic} \in \mathbb{R}^{(2N+2)\times(2N+2)}$, vectors $a_k^{ic} \in \mathbb{R}^{N+2}$ and scalars $b_k^{ic} \in \mathbb{R}$ for all $k \in K_N$, such that for any set of triples $S = \{(x_i, g_i, f_i)\}_{i \in I_N^*}$, S is $\mathcal{F}(\mathbb{R}^d)$-interpolable if and only if

$$\mathrm{Tr}\left(A_k^{ic} G\right) + (a_k^{ic})^\top F + b_k^{ic} \leq 0, \quad \text{for all } k \in K_N. \tag{11}$$

(The notation *ic* above is an abbreviation of *interpolation conditions*.)

Example 1 **(Interpolation conditions for $\mathcal{F}_{\mu,L}$)** For the class of L-smooth and μ-strongly convex functions, the index set K_N can be defined by $K_N = \{(i, j) : i, j \in I_N^*\}$, and $\{(A_k^{ic}, a_k^{ic}, b_k^{ic})\}_{k \in K_N}$ can be defined by:

$$A_{(i,j)}^{\text{ic}} = \mathbf{g}_j \odot (\mathbf{x}_i - \mathbf{x}_j) + \frac{1}{2(1 - \mu/L)} \left(\frac{1}{L} (\mathbf{g}_i - \mathbf{g}_j) \odot (\mathbf{g}_i - \mathbf{g}_j) \right.$$

$$\left. + \mu(\mathbf{x}_i - \mathbf{x}_j) \odot (\mathbf{x}_i - \mathbf{x}_j) - 2\frac{\mu}{L}(\mathbf{x}_j - \mathbf{x}_i) \odot (\mathbf{g}_j - \mathbf{g}_i) \right),$$

$$a_{(i,j)}^{\text{ic}} = \mathbf{f}_j - \mathbf{f}_i,$$

$$b_{(i,j)}^{\text{ic}} = 0,$$

for all $(i, j) \in K_N$ (see Theorem 1).

Example 2 **(Interpolation conditions for $\mathcal{C}_M$)** Consider the class of Lipschitz c.c.p. functions whose gradients are bounded in norm by M for some $0 \leq M \leq \infty$. In this setting, there are two types of constraints encoding the interpolation conditions (see Theorem 2): constraints bounding the gradients, and constraints ensuring convexity. We therefore set $K_N = I_N^* \cup \{(i, j) \in I_N^* \times I_N^* : i \neq j\}$ and define $\{(A_k^{\text{ic}}, a_k^{\text{ic}}, b_k^{\text{ic}})\}_{k \in K_N}$ as follows:

$$
\begin{aligned}
A_i^{\text{ic}} &= \mathbf{g}_i \odot \mathbf{g}_i, & i &\in I_N^*, \\
a_i^{\text{ic}} &= 0, & i &\in I_N^*, \\
b_i^{\text{ic}} &= -M^2, & i &\in I_N^*, \\
A_{(i,j)}^{\text{ic}} &= \mathbf{g}_j \odot (\mathbf{x}_i - \mathbf{x}_j), & i &\neq j \in I_N^*, \\
a_{(i,j)}^{\text{ic}} &= \mathbf{f}_j - \mathbf{f}_i, & i &\neq j \in I_N^*, \\
b_{(i,j)}^{\text{ic}} &= 0, & i &\neq j \in I_N^*.
\end{aligned}
$$

We can now formulate a tractable performance estimation problem for GFOM.

Lemma 2 *Let (f, x_0) be an $(\mathcal{F}(\mathbb{R}^d), R_x)$-input, $N \in \mathbb{N}$, and suppose $\{(A_k^{\text{ic}}, a_k^{\text{ic}}, b_k^{\text{ic}})\}_{k \in K_N}$ encodes the interpolation conditions for $\mathcal{F}(\mathbb{R}^d)$ (see Assumption 1). Then*

$$\text{val}\,(\text{PEP-GFOM}) \leq \text{val}\,(\text{sdp-PEP-GFOM}), \tag{12}$$

where

$$\sup_{F \in \mathbb{R}^{N+1}, G \in \mathbb{R}^{2N+2 \times 2N+2}} F^\top \mathbf{f}_N - F^\top \mathbf{f}_* \qquad \text{(sdp-PEP-GFOM)}$$

subject to: $\ \mathrm{Tr}\left(A_k^{\text{ic}} G\right) + (a_k^{\text{ic}})^\top F + b_k^{\text{ic}} \leq 0, \quad \text{for all } k \in K_N,$

$$\langle \mathbf{g}_i, \mathbf{g}_j \rangle_G = 0, \quad \text{for all } 0 \leq j < i = 1, \dots N,$$

$$\langle \mathbf{g}_i, \mathbf{x}_j - \mathbf{x}_0 \rangle_G = 0, \quad \text{for all } 1 \leq j \leq i = 1, \dots N,$$

$$\|\mathbf{x}_0 - \mathbf{x}_*\|_G^2 - R_x^2 \leq 0,$$

$$G \succeq 0.$$

Furthermore, if $d \geq 2N + 2$ then (12) holds with equality.

Proof Since any feasible solution $\{(x_i, g_i, f_i)\}_{i \in I_N^*}$ to (PEP-GFOM) can be transformed to a feasible solution to (sdp-PEP-GFOM), by setting $G = P^\top P$, where P is defined as in (9), then (12) immediately follows.

Now, suppose $d \geq 2N+2$ and let (F, G) be feasible to (sdp-PEP-GFOM). Since G is a $(2N+2) \times (2N+2)$ positive-semidefinite matrix, there exits some $d \times (2N+2)$ matrix P such that $G = P^\top P$, and thus (F, G) can be transformed to a feasible solution for (PEP-GFOM) by assigning values for $\{(x_i, g_i, f_i)\}_{i \in I_N^*}$ according to (9). We have obtained

$$\text{val} \, (\text{sdp-PEP-GFOM}) \leq \text{val} \, (\text{PEP-GFOM}),$$

which completes the proof. $\square$

We now describe the final form of the bound on the performance of GFOM, which is the standard Lagrangian dual of (sdp-PEP-GFOM). This bound provides the basic building block for SSEP.

Theorem 3 *Let $N \in \mathbb{N}$, $R_x \geq 0$, $\mathcal{F}(\mathbb{R}^d)$ a class of c.c.p. functions and suppose $\{(A_k^{\mathrm{ic}}, a_k^{\mathrm{ic}}, b_k^{\mathrm{ic}})\}_{k \in K_N}$ encodes the interpolation conditions for $\mathcal{F}$ (see Assumption 1). Then for any $(\mathcal{F}(\mathbb{R}^d), R_x)$-input (f, x_0),*

$$f(\text{GFOM}_N(f, x_0)) - f_* \leq \text{val} \, (\text{dual-PEP-GFOM}) \tag{13}$$

holds with

$$\inf_{\{\alpha_k\}, \{\beta_{i,j}\}, \{\gamma_{i,j}\}, \tau_x} \tau_x R_x^2 - \sum_{k \in K_N} \alpha_k b_k^{\mathrm{ic}} \tag{dual-PEP-GFOM}$$

subject to: $\alpha_k \geq 0, \ \tau_x \geq 0,$

$$\sum_{k \in K_N} \alpha_k A_k^{\mathrm{ic}} + \sum_{i=1}^{N} \mathbf{g}_i \odot \left[\sum_{j=0}^{i-1} \beta_{i,j} \mathbf{g}_j + \sum_{j=1}^{i} \gamma_{i,j} (\mathbf{x}_j - \mathbf{x}_0) \right]$$
$$+ \tau_x [(\mathbf{x}_0 - \mathbf{x}_*) \odot (\mathbf{x}_0 - \mathbf{x}_*)] \succeq 0,$$
$$\mathbf{f}_N - \mathbf{f}_* - \sum_{k \in K_N} \alpha_k a_k^{\mathrm{ic}} = 0.$$

Furthermore, if $\mathcal{F}(\mathbb{R}^d)$ is contraction-preserving, $d \geq 2N+2$ and there exists some $(\mathcal{F}(\mathbb{R}^d), R_x)$-input $(\hat{f}, \hat{x}_0)$ such that $\text{GFOM}_{2N+1}(\hat{f}, \hat{x}_0)$ is not optimal for $\hat{f}$, then the bound (13) is tight.

Proof The first part of the claim follows directly by establishing weak duality between (dual-PEP-GFOM) and (sdp-PEP-GFOM). Indeed, one can use the following association between the constraints and dual variables along with the definition of Lagrange duality:

$$\text{Tr}\left(A_k^{\text{ic}}G\right) + (a_k^{\text{ic}})^\top F + b_k^{\text{ic}} \leq 0, \quad \text{for all } k \in K_N \qquad\qquad : \alpha_k,$$
$$\langle \mathbf{g}_i, \mathbf{g}_j \rangle_G = 0, \quad \text{for all } 0 \leq j < i = 1, \ldots N : \beta_{i,j},$$
$$\langle \mathbf{g}_i, \mathbf{x}_j - \mathbf{x}_0 \rangle_G = 0, \quad \text{for all } 1 \leq j \leq i = 1, \ldots N : \gamma_{i,j},$$
$$\|\mathbf{x}_0 - \mathbf{x}_*\|_G^2 - R_x^2 \leq 0, \qquad\qquad\qquad\qquad : \tau_x.$$

The proof for the tightness claim is presented in Appendix 2. □

Remark 1 Since (dual-PEP-GFOM) is an infimum problem, it has the useful property that any feasible solution to this problem corresponds to an upper bound on the worst-case accuracy of GFOM. We take advantage of this property in the following, where bounds on the performance of GFOM for different classes of problems are established by providing a (dual-PEP-GFOM)-feasible solution.

Remark 2 As an example of a function in $\mathcal{F}_{\mu,\text{L}}(\mathbb{R}^d)$ that cannot be minimized by GFOM within $k < d$ iterations (as required by the tightness claim in the previous theorem) recall that for quadratic functions, the iterates of GFOM coincide with the iterates of the Conjugate Gradient method, hence by a well-known result, any quadratic form with d distinct eigenvalues requires d iterations to minimize (for a general starting point). In the non-smooth case, $\mathcal{C}_M(\mathbb{R}^d)$, one can take, for example,

$$\hat{f}(x) = M \max(\langle x, e_1 \rangle, \ldots, \langle x, e_d \rangle, \|x\| - 1),$$

with $x_0 = 0$ (where e_i are the canonical unit vectors). For more details, see [14, Appendix A].

To conclude this section, we note that although the analysis above was performed under Assumption 1, for classes of functions for which a set of interpolation conditions is either unknown or complex, the analysis can still proceed using a set of *necessary* conditions for interpolability, with the only change being that tightness claims no longer apply.

4 The subspace-search elimination procedure

In this section, we introduce a technique for constructing first-order methods with a worst-case absolute inaccuracy that is guaranteed to be not worse than that of GFOM. We begin by stating the main technical result, we then outline the SSEP technique, and finally demonstrate the application of the technique on several cases.

Theorem 4 *Let $N \in \mathbb{N}$, $R_x \geq 0$, $\mathcal{F}$ a class of c.c.p. functions for which Assumption 1 holds, and let $(\{\tilde{\alpha}_k\}, \{\tilde{\beta}_{i,j}\}, \{\tilde{\gamma}_{i,j}\}, \tilde{\tau}_x)$ be a feasible solution to (dual-PEP-GFOM) that attains the objective value $\tilde{\omega}$. For any $(\mathcal{F}, R_x)$-input (f, x_0), if $\{x_i\}$ is a sequence that satisfies*

$$\left\langle f'(x_i), \sum_{j=0}^{i-1} \tilde{\beta}_{i,j} f'(x_j) + \sum_{j=1}^{i} \tilde{\gamma}_{i,j}(x_j - x_0) \right\rangle = 0, \quad i = 1, \ldots, N, \qquad (14)$$

then the bound $f(x_N) - f_* \leq \tilde{\omega}$ holds for any choice of $f'(x_i) \in \partial f(x_i)$.

The proof, presented in Appendix 3, is based on the observation that by carefully aggregating the constraints in (dual-PEP-GFOM), it is possible to reach a PEP for methods satisfying (14) without adversely affecting the optimal value of the PEP.

Based on this result, the design procedure can be summarized as follows.

Subspace-search elimination procedure (SSEP)

1. Choose $N \geq 0$, $R_x > 0$ and a set of interpolation conditions $\{(A_k^{\mathrm{ic}}, a_k^{\mathrm{ic}}, b_k^{\mathrm{ic}})\}_{k \in K_N}$ for $\mathcal{F}$.
2. Find a feasible solution $(\{\tilde{\alpha}_k\}, \{\tilde{\beta}_{i,j}\}, \{\tilde{\gamma}_{i,j}\}, \tilde{\tau}_x)$ to (dual-PEP-GFOM), and denote the objective value of the solution by $\tilde{\omega}$.
3. Find a method satisfying (14).
$\rightarrow$ The worst-case absolute inaccuracy at its Nth iterate is guaranteed to be at most $\tilde{\omega}$.

Remark 3 Theorem 4 states that any point feasible to (dual-PEP-GFOM) can be used as a basis for constructing new methods with worst-case performances that are bounded by the value of the objective at that feasible point. In particular, this applies to an optimal solution of (dual-PEP-GFOM), hence in cases where (dual-PEP-GFOM) attains the worst-case performance of GFOM (e.g., under the tightness conditions in Theorem 3), the worst-case performance of any method constructed according to (14) from an optimal solution is guaranteed to be equal to or better than the worst-case performance of GFOM.

Equality (14) can be enforced in different ways. Perhaps the most straightforward one by an appropriate fixed-step size policy as detailed below.

Corollary 1 *Let $N \in \mathbb{N}$, $R_x \geq 0$, $\mathcal{F}$ a class of c.c.p. functions for which Assumption 1 holds, and let $(\{\tilde{\alpha}_k\}, \{\tilde{\beta}_{i,j}\}, \{\tilde{\gamma}_{i,j}\}, \tilde{\tau}_x)$ be a feasible solution to (dual-PEP-GFOM) that attains the objective value $\tilde{\omega}$ and satisfies $\tilde{\gamma}_{i,i} \neq 0$ for $i = 1, \ldots, N$. For any $(\mathcal{F}, R_x)$-input (f, x_0), if $\{x_i\}$ is defined by*

$$x_i = x_0 - \sum_{j=1}^{i-1} \frac{\tilde{\gamma}_{i,j}}{\tilde{\gamma}_{i,i}} (x_j - x_0) - \sum_{j=0}^{i-1} \frac{\tilde{\beta}_{i,j}}{\tilde{\gamma}_{i,i}} f'(x_j), \quad i = 1, \ldots, N, \qquad (15)$$

then the bound $f(x_N) - f_* \leq \tilde{\omega}$ holds for any choice of $f'(x_i) \in \partial f(x_i)$.

Proof Under the assumption that $\tilde{\gamma}_{i,i} \neq 0$, any sequence satisfying (15) also satisfies (14) and hence Theorem 4 directly applies.   □

Remark 4 Since the optimization variables $\gamma_{i,i}$ in (sdp-PEP-GFOM) are the dual variables to the constraints "$\langle g_i, x_i - x_0 \rangle = 0$" that define how GFOM takes advantage of the first-order information at x_i, the condition $\tilde{\gamma}_{i,i} \neq 0$ in the previous corollary appears to be a natural requirement in this setting, and indeed, this condition is fulfilled in all cases treated in the sequel. See also [13, Theorem 3] for a similar condition that arises in a related context.

In addition to the fixed-step method defined in Corollary 1, there are additional strategies for enforcing (14), where, in particular, the iterates of GFOM satisfy these equalities. As described in the examples below, this flexibility allows the construction of methods that have additional properties, such as independence on problem parameters (e.g., Lipschitz constants or initial distance to optimality R_x).

4.1 Example: non-smooth convex minimization

Consider the problem of minimizing a non-smooth convex function

$$\min_{x \in \mathbb{R}^d} f(x),$$

with $f \in \mathcal{C}_M$ (i.e., f is convex with $\|f'(x)\| \leq M$ for all $x \in \mathbb{R}^d$ and for all $f'(x) \in \partial f(x)$) and under the assumption that the distance between the initial point and an optimal point $\|x_0 - x_*\|$ is bounded by a constant R_x.

We start the SSEP by establishing a worst-case bound on GFOM in this case. As discussed above, for this purpose it is sufficient to find a dual feasible solution to (PEP-GFOM) when the interpolation conditions used matches the class of functions under consideration.

Lemma 3 *The following assignment is feasible to* (dual-PEP-GFOM) *under the interpolation conditions for $\mathcal{C}_M$ defined in Example 2, and attain the objective value of $\frac{MR_x}{\sqrt{N+1}}$:*

$$
\begin{aligned}
&\alpha_i = \frac{R_x}{2M(N+1)^{3/2}}, \quad i = 0, \ldots, N, \\
&\alpha_{i-1,i} = \frac{i}{N+1}, \quad i = 1, \ldots, N; \quad \alpha_{*,i} = \frac{1}{N+1}, \quad i = 0, \ldots, N, \\
&\gamma_{i,i} = \frac{i+1}{N+1}, \quad i = 1, \ldots, N; \quad \gamma_{i,i-1} = \frac{-i}{N+1}, \quad i = 2, \ldots, N, \quad (16) \\
&\beta_{i,j} = \frac{R_x}{M(N+1)^{3/2}}, \quad i = 1, \ldots, N, \; j = 0, \ldots, i-1, \\
&\tau_x = \frac{M}{2R_x\sqrt{N+1}},
\end{aligned}
$$

where all the other optimization variables in (dual-PEP-GFOM) *are set to zero.*

For the sake of coherence, the following proof relies on the SDP formalism developed above. One can note, though, that it is possible to reformulate the proof below using equivalent sum-of-squares arguments.

Proof Since the inequality constraint in (dual-PEP-GFOM) clearly holds for (16), it is enough to verify the positive-semidefinite constraint and the equality constraint, i.e.:

$$\sum_{i=0}^{N} \alpha_i A_i^{\text{ic}} + \sum_{i \neq j \in I_N^*} \alpha_{i,j} A_{i,j}^{\text{ic}} + \sum_{i=1}^{N} \mathbf{g}_i \odot \left[\sum_{j=0}^{i-1} \beta_{i,j} \mathbf{g}_j + \sum_{j=1}^{i} \gamma_{i,j} (\mathbf{x}_j - \mathbf{x}_0) \right]$$
$$+ \tau_x [(\mathbf{x}_0 - \mathbf{x}_*) \odot (\mathbf{x}_0 - \mathbf{x}_*)] \succeq 0,$$

$$\mathbf{f}_N - \sum_{i=0}^{N} \alpha_i a_i^{\text{ic}} - \sum_{i \neq j \in I_N^*} \alpha_{i,j} a_{i,j}^{\text{ic}} = 0.$$

Substituting with (16) and the definition of A^{ic} and a^{ic}, we reach

$$\frac{R_x}{2M(N+1)^{3/2}} \sum_{i=0}^{N} \mathbf{g}_i \odot \mathbf{g}_i + \sum_{i=1}^{N} \frac{i}{N+1} \mathbf{g}_i \odot (\mathbf{x}_{i-1} - \mathbf{x}_i) + \frac{1}{N+1} \sum_{i=0}^{N} \mathbf{g}_i \odot (\mathbf{x}_* - \mathbf{x}_i)$$

$$+ \sum_{i=1}^{N} \mathbf{g}_i \odot \left[\frac{R_x}{M(N+1)^{3/2}} \sum_{j=0}^{i-1} \mathbf{g}_j + \frac{-i}{N+1} (\mathbf{x}_{i-1} - \mathbf{x}_0) + \frac{i+1}{N+1} (\mathbf{x}_i - \mathbf{x}_0) \right]$$

$$+ \frac{M}{2R_x\sqrt{N+1}} [(\mathbf{x}_0 - \mathbf{x}_*) \odot (\mathbf{x}_0 - \mathbf{x}_*)] \succeq 0,$$

$$\mathbf{f}_N - \frac{1}{N+1} \sum_{i=0}^{N} (\mathbf{f}_i - \mathbf{f}_*) - \sum_{i=1}^{N} \frac{i}{N+1} (\mathbf{f}_i - \mathbf{f}_{i-1}) = 0.$$

The first condition can be verified by showing that it is equal to the following positive-semidefinite rank-one matrix (this may require some work to obtain directly, but can easily be verified by developing both expressions):

$$\frac{M}{2R_x\sqrt{N+1}} \left(\mathbf{x}_0 - \mathbf{x}_* - \frac{1}{\sqrt{N+1}} \frac{R_x}{M} \sum_{i=0}^{N} \mathbf{g}_i \right) \odot \left(\mathbf{x}_0 - \mathbf{x}_* - \frac{1}{\sqrt{N+1}} \frac{R_x}{M} \sum_{i=0}^{N} \mathbf{g}_i \right).$$

The equality is straightforward to verify. Finally, the objective value is given by

$$\tau_x R_x^2 - \sum_{k \in K_N} \alpha_k b_k^{\text{ic}} = \frac{M}{2R_x\sqrt{N+1}} R_x^2 + \sum_{i=0}^{N} \frac{R_x}{2M(N+1)^{3/2}} M^2 = \frac{MR_x}{\sqrt{N+1}}.$$

□

The following is now immediate from Theorem 3, Proposition 2 and Theorem 4.

Corollary 2 *Let* (f, x_0) *be a* $(\mathcal{C}_M, R_x)$-*input for some* $M, R_x \geq 0$.

1. *For any* $N \in \mathbb{N}$

$$f(\text{GFOM}_N(f, x_0)) - f_* \leq \frac{MR_x}{\sqrt{N+1}}.$$

Furthermore, this bound is tight when $d \geq 2N + 2$.

2. *For any sequence $x_1, \ldots, x_N$ that satisfies*

$$\left\langle f'(x_i), x_i - \left[\frac{i}{i+1}x_{i-1} + \frac{1}{i+1}x_0 - \frac{1}{i+1}\frac{R_x}{M\sqrt{N+1}} \sum_{j=0}^{i-1} f'(x_j) \right] \right\rangle = 0,$$
$$i = 1, \ldots, N, \tag{17}$$

we have

$$f(x_N) - f_* \leq \frac{MR_x}{\sqrt{N+1}}.$$

As noted above, sequences satisfying (17) can be generated in several ways. We start by describing an efficient implementation of the fixed-step scheme described in Corollary 1.

SSEP-based subgradient method

Input: $f \in \mathcal{C}_M(\mathbb{R}^d)$, initial guess $x_0 \in \mathbb{R}^d$ such that $\|x_0 - x_*\| \leq R_x$, number of iterations N.

For $i = 1, \ldots, N$:

$$y_i = \frac{i}{i+1}x_{i-1} + \frac{1}{i+1}x_0$$

$$d_i = \frac{1}{i+1} \sum_{j=0}^{i-1} f'(x_j)$$

$$x_i = y_i - \frac{R_x}{M\sqrt{N+1}}d_i$$

Output: x_N.

Corollary 3 *Let (f, x_0) be a $(\mathcal{C}_M, R_x)$-input for some $M, R_x \geq 0$. If x_N is an output of the SSEP-based subgradient method given f, x_0 and $N \in \mathbb{N}$, then*

$$f(x_N) - f_* \leq \frac{MR_x}{\sqrt{N+1}}.$$

Proof The claim then follows directly from the second part of Corollary 2. $\qquad\square$

Note though that the SSEP-based subgradient method has a guarantee on its *last* iterate, whereas the guarantees on standard subgradient methods are usually either on the averaged iterate $f\left(\frac{1}{N+1}\sum_{i=0}^{N} x_i\right) - f_*$ or in terms of the best iterate $\min_{0 \leq i \leq N} f(x_i) - f_*$. Interestingly, the SSEP-based subgradient method appears to be similar to the so-called quasi-monotone subgradient methods [39].

As noted above, there are additional ways of enforcing equality (14), allowing the introduction of methods with different properties. As an example, one can subsume prior knowledge of the constants R_x, M and N by using an exact line search procedure, as demonstrated by the following optimal subgradient method.

SSEP-based subgradient method with an exact line search

Input: $f \in \mathcal{C}_M(\mathbb{R}^d)$, initial guess $x_0 \in \mathbb{R}^d$.
For $i = 1, 2, \ldots$:

$$y_i = \frac{i}{i+1} x_{i-1} + \frac{1}{i+1} x_0$$

$$d_i = \frac{1}{i+1} \sum_{j=0}^{i-1} f'(x_j)$$

$$\alpha \in \operatorname*{argmin}_{\alpha \in \mathfrak{R}} f(y_i - \alpha d_i)$$

$$x_i = y_i - \alpha d_i$$

Choose $f'(x_i) \in \partial f(x_i)$ such that $\langle f'(x_i), d_i \rangle = 0$

Corollary 4 *Let* (f, x_0) *be an* $(\mathcal{C}_M, R_x)$-*input for some* $M, R_x \geq 0$. *For any sequence* $\{x_i\}$ *generated by SSEP-based subgradient method with an exact line search given* f *and* x_0

$$f(x_N) - f_* \leq \frac{M R_x}{\sqrt{N+1}}, \quad \forall N \in \mathbb{N}.$$

Proof First, note that from the first-order optimality conditions on the exact line search procedure, for all i there exist $f'(x_i) \in \partial f(x_i)$ that satisfies the requirement $\langle f'(x_i), d_i \rangle = 0$. Now, from definition of x_i and y_i, we have

$$\left\langle f'(x_i), x_i - \left[\frac{i}{i+1} x_{i-1} + \frac{1}{i+1} x_0 - \frac{1}{i+1} \frac{R_x}{M\sqrt{N+1}} \sum_{j=0}^{i-1} f'(x_j) \right] \right\rangle$$

$$= \left\langle f'(x_i), x_i - y_i + \frac{R_x}{M\sqrt{N+1}} d_i \right\rangle = \left\langle f'(x_i), -\alpha d_i + \frac{R_x}{M\sqrt{N+1}} d_i \right\rangle = 0,$$

which concludes the proof, since this establishes (17) as required by Corollary 2. □

Remark 5 In the setting considered in this section, it is known that no first-order method can have a worst-case absolute inaccuracy behavior that better than

$$f(x_N) - f_* \geq \frac{M R_x}{\sqrt{N+1}}, \tag{18}$$

after N iterations when $d \geq N + 2$ [14, Theorem 2], hence the subgradient algorithms described above are optimal and the corresponding bounds are tight.

Note that the methods developed above are not the first subgradient schemes that attains the optimal worst-case complexity for non-smooth minimization, as it is achieved, for example by the optimal step-size policy $\frac{R_x}{M\sqrt{N+1}}$ proposed in [36, Section 3.2.3] and by the Kelley-like cutting plane method developed in [14].

4.2 Example: smooth convex minimization

In this section we demonstrate the application of SSEP on the problem of minimizing a smooth convex function. We show that GFOM attains the best possible worst-case bound on this problem, and that the resulting fixed-step method is the optimized gradient method (OGM) developed in [13,22]. Finally, we construct a method that has the same worst-case performance, but does not require prior knowledge on the problem parameters, at the cost of performing a one-dimensional line search at each iteration.

We use the following notations:

$$\theta_{0,N} := 1,$$

$$\theta_{i+1,N} := \frac{1 + \sqrt{4\theta_{i,N}^2 + 1}}{2}, \quad i = 0, 1, \ldots, N - 1,$$

$$\theta_{N,N} := \frac{1 + \sqrt{8\theta_{N-1,N}^2 + 1}}{2}.$$

Lemma 4 *The following assignment is feasible to* (dual-PEP-GFOM) *under the interpolation conditions for $\mathcal{F}_{0,L}$ defined in Example 1, and attains the objective value of $\frac{LR_x^2}{2\theta_{N,N}^2}$:*

$$\alpha_{i-1,i} = \frac{2\theta_{i-1,N}^2}{\theta_{N,N}^2}, \quad i = 1, \ldots, N; \quad \alpha_{*,i} = \frac{2\theta_{i,N}}{\theta_{N,N}^2}, \quad i = 0, \ldots, N-1; \quad \alpha_{*,N} = \frac{1}{\theta_{N,N}},$$

$$\gamma_{i,i-1} = \frac{-2\theta_{i-1,N}^2}{\theta_{N,N}^2}, \quad i = 2, \ldots, N; \quad \gamma_{i,i} = \frac{2\theta_{i,N}^2}{\theta_{N,N}^2}, \quad i = 1, \ldots, N-1; \quad \gamma_{N,N} = 1,$$

$$\tau_x = \frac{L}{2\theta_{N,N}^2},$$

and

$$\beta_{i,j} = \frac{1}{L} \frac{4\theta_{i,N}\theta_{j,N}}{\theta_{N,N}^2}, \qquad\qquad i = 2, \ldots, N-1, \ j = 0, \ldots, i-2,$$

$$\beta_{i,i-1} = \frac{1}{L} \frac{2\theta_{i-1,N}^2 + 4\theta_{i-1,N}\theta_{i,N}}{\theta_{N,N}^2}, \qquad i = 1, \ldots, N-1,$$

$$\beta_{N,i} = \frac{1}{L}\frac{2\theta_{i,N}}{\theta_{N,N}}, \qquad\qquad i = 0,\ldots,N-2,$$

$$\beta_{N,N-1} = \frac{1}{L}\left(\frac{2\theta_{N-1,N}}{\theta_{N,N}} + \frac{2\theta_{N-1,N}^2}{\theta_{N,N}^2}\right),$$

where all the other optimization variables in (dual-PEP-GFOM) *are set to zero.*

Proof As in the non-smooth case, it is enough to show that the following constraints hold:

$$\sum_{i\neq j\in I_N^*} \alpha_{i,j} A_{i,j}^{\mathrm{ic}} + \sum_{i=1}^{N} \mathbf{g}_i \odot \left[\sum_{j=0}^{i-1}\beta_{i,j}\mathbf{g}_j + \sum_{j=1}^{i}\gamma_{i,j}(\mathbf{x}_j - \mathbf{x}_0)\right]$$
$$+ \tau_x[(\mathbf{x}_0 - \mathbf{x}_*)\odot(\mathbf{x}_0 - \mathbf{x}_*)] \succeq 0,$$
$$\mathbf{f}_N - \sum_{i\neq j\in I_N^*}\alpha_{i,j}a_{i,j}^{\mathrm{ic}} = 0.$$

With some effort, one can show that the first expression above can be written as:

$$\frac{L}{2\theta_{N,N}^2}\left(\mathbf{x}_0 - \mathbf{x}_* - \frac{\theta_{N,N}}{L}\mathbf{g}_N - \frac{2}{L}\sum_{i=0}^{N-1}\theta_{i,N}\mathbf{g}_i\right)$$
$$\odot\left(\mathbf{x}_0 - \mathbf{x}_* - \frac{\theta_{N,N}}{L}\mathbf{g}_N - \frac{2}{L}\sum_{i=0}^{N-1}\theta_{i,N}\mathbf{g}_i\right),$$

which is clearly a PSD matrix. As the second equality above also holds, it follows that the selected values define a feasible point. Finally, the objective value that corresponds to that point is given by

$$\tau_x R_x^2 = \frac{L R_x^2}{2\theta_{N,N}^2}.$$

$\square$

From Theorem 3 Proposition 1 and Theorem 4 we get the following result.

Corollary 5 *Let* (f, x_0) *be an* $(\mathcal{F}_{0,L}, R_x)$-*input for some* $L, R_x \geq 0$.

1. *For any* $N \in \mathbb{N}$

$$f(\mathrm{GFOM}_N(f, x_0)) - f_* \leq \frac{L R_x^2}{2\theta_{N,N}^2}.$$

Furthermore, this bound is tight when $d \geq 2N + 2$.

2. *For any sequence $x_1, \ldots, x_N$ that satisfies*

$$\left\langle g_i, x_i - \left[\left(1 - \frac{1}{\theta_{i,N}}\right) \left(x_{i-1} - \frac{1}{L} f'(x_{i-1})\right) + \frac{1}{\theta_{i,N}} \left(x_0 - \frac{2}{L} \sum_{j=0}^{i-1} \theta_{j,N} f'(x_j)\right) \right] \right\rangle$$

$$= 0, \quad i = 1, \ldots, N,$$

(19)

we have

$$f(x_N) - f_* \leq \frac{L R_x^2}{2\theta_{N,N}^2}.$$

As an immediate application of Corollary 1, we recover the optimized gradient method (OGM), which was developed in [13,22].

Optimized gradient method (OGM) [22]

Input: $f \in \mathcal{F}_{0,L}(\mathbb{R}^d)$, initial guess $x_0 \in \mathbb{R}^d$, number of iterations N.
For $i = 1, \ldots, N$:

$$y_i = \left(1 - \frac{1}{\theta_{i,N}}\right) x_{i-1} + \frac{1}{\theta_{i,N}} x_0$$

$$d_i = \left(1 - \frac{1}{\theta_{i,N}}\right) f'(x_{i-1}) + \frac{2}{\theta_{i,N}} \sum_{j=0}^{i-1} \theta_{j,N} f'(x_j)$$

$$x_i = y_i - \frac{1}{L} d_i$$

Output: x_N.

As in the non-smooth case, one can trade the knowledge of the problem parameters with an exact line search procedure, resulting in an optimized gradient method with exact line search.

Optimized gradient method with exact line search (OGM-LS)

Input: $f \in \mathcal{F}_{0,L}(\mathbb{R}^d)$, initial guess $x_0 \in \mathbb{R}^d$, number of iterations N.
For $i = 1, \ldots, N$:

$$y_i = \left(1 - \frac{1}{\theta_{i,N}}\right) x_{i-1} + \frac{1}{\theta_{i,N}} x_0$$

$$d_i = \left(1 - \frac{1}{\theta_{i,N}}\right) f'(x_{i-1}) + \frac{2}{\theta_{i,N}} \sum_{j=0}^{i-1} \theta_{j,N} f'(x_j)$$

$$\alpha \in \operatorname*{argmin}_{\alpha \in \Re} f(y_i - \alpha d_i)$$

$$x_i = y_i - \alpha d_i$$

Output: x_N.

We omit the rate of convergence proofs, as they are identical to the proofs presented in the previous section. As in the non-smooth case, tightness of the worst-case bounds can be established by observing that they coincide with the lower complexity bound for the problem [12, Corollary 4] and therefore they cannot be improved in the large-scale setting ($d \geq N + 2$).

Remark 6 Before proceeding, let us note that the results above are reminiscent to the historical developments premising accelerated methods. Indeed, acceleration was first discovered in the work of Nemirovski [30] and Nemirovski and Yudin [33], who needed two or three-dimensional subspace-searches for obtaining the optimal convergence rate for smooth convex minimization. This rather strong requirement was removed later on in the work of Nesterov, resulting in the first version of the celebrated fast gradient method [35] not requiring any line (or space) search.

4.3 Example: a universal method for non-smooth and smooth convex minimization

In this short section, we build upon the results of Sects. 4.1 and 4.2 to develop a *universal* method for both smooth and non-smooth minimization, i.e., a method that does not require any knowledge on which of the two classes the function belongs to, nor does it require knowledge on the specific parameters of the classes. The knowledge is replaced by the capability of performing exact three-dimensional subspace minimizations.

A universal method for non-smooth and smooth minimization (UM)

Input: $f \in \mathcal{F}_{0,L}(\mathbb{R}^d)$ or $f \in \mathcal{C}_M(\mathbb{R}^d)$, initial guess $x_0 \in \mathbb{R}^d$, number of iterations N.

For $i = 1, \ldots, N$:

$$y_i^{(1)} = \frac{i}{i+1} x_{i-1} + \frac{1}{i+1} x_0$$

$$d_i^{(1)} = \frac{1}{i+1} \sum_{j=0}^{i-1} f'(x_j)$$

$$y_i^{(2)} = \left(1 - \frac{1}{\theta_{i,N}}\right) x_{i-1} + \frac{1}{\theta_{i,N}} x_0$$

$$d_i^{(2)} = \left(1 - \frac{1}{\theta_{i,N}}\right) f'(x_{i-1}) + \frac{2}{\theta_{i,N}} \sum_{j=0}^{i-1} \theta_{j,N} f'(x_j)$$

$$\alpha, \beta, \gamma \in \operatorname*{argmin}_{\alpha,\beta,\gamma \in \Re} f(\alpha y_i^{(1)} + (1-\alpha) y_i^{(2)} - \beta d_i^{(1)} - \gamma d_i^{(2)})$$

$$x_i = \alpha y_i^{(1)} + (1-\alpha) y_i^{(2)} - \beta d_i^{(1)} - \gamma d_i^{(2)}$$

Choose $f'(x_i) \in \partial f(x_i)$ such that $\langle f'(x_i), d_i \rangle = 0$

Output: x_N.

Corollary 6 *Let x_N be an output generated by UM given f, x_0 and $N \in \mathbb{N}$.*

1. *If (f, x_0) is an $(\mathcal{C}_M, R_x)$-input for some $M \geq 0$, then*

$$f(x_N) - f_* \leq \frac{M R_x}{\sqrt{N+1}},$$

2. *If (f, x_0) is an $(\mathcal{F}_{0,L}, R_x)$-input for some $L \geq 0$, then*

$$f(x_N) - f_* \leq \frac{L R_x^2}{2\theta_{N,N}^2}.$$

Proof For the first part of the claim, it is enough to establish that the sequence generated by UM satisfies (17). Indeed, for $i = 1, \ldots, N$

$$
\left\langle f'(x_i), x_i - \left[\frac{i}{i+1} x_{i-1} + \frac{1}{i+1} x_0 - \frac{1}{i+1} \frac{R_x}{M\sqrt{N+1}} \sum_{j=0}^{i-1} f'(x_j) \right] \right\rangle
$$
$$
= \left\langle f'(x_i), x_i - y_i^{(1)} + \frac{R_x}{M\sqrt{N+1}} d_i^{(1)} \right\rangle
$$
$$
= \left\langle f'(x_i), (\alpha - 1)y_i^{(1)} + (1 - \alpha)y_i^{(2)} - \beta d_i^{(1)} - \gamma d_i^{(2)} + \frac{R_x}{M\sqrt{N+1}} d_i^{(1)} \right\rangle
$$
$$
= 0,
$$

where the last equality follows from the optimality conditions of the exact subspace minimization step:

$$
\left\langle f'(x_i), y_i^{(1)} - y_i^{(2)} \right\rangle = 0,
$$
$$
\left\langle f'(x_i), d_i^{(1)} \right\rangle = 0,
$$
$$
\left\langle f'(x_i), d_i^{(2)} \right\rangle = 0.
$$

The second part of the claim follows in an analogous way using (19).   □

5 Numerical construction of efficient methods

In this section we focus our attention to situations where either (dual-PEP-GFOM) does not have a known analytical solution, or the analytical solution is too complex for practical purposes. In particular, we examine the performance of the methods generated by SSEP in the case where $f \in \mathcal{F}_{\mu,L}$ (an L-smooth, μ-strongly convex function). Note that since $f \in \mathcal{F}_{\mu,L}$ implies $\frac{1}{L} f \in \mathcal{F}_{\frac{\mu}{L},1}$, the worst-case analyses can be limited to the case $L = 1$, where simple homogeneity arguments allow extending the results to the case of arbitrary $L > 0$. For a short discussion on this topic, we refer to [47, Section 4.2.5].

In this setting, finding analytical solutions to (dual-PEP-GFOM) appears to be significantly more involved than in the previous examples, nevertheless the problem can be efficiently approximated numerically using standard SDP solvers and SSEP can proceed as described above. A problem with this approach, however, is that the computation complexity of a naïve implementation of Corollary 1 would require storing all subgradients encountered throughout the computations and performing $O(i)$ vector operations at the ith iteration, making such an implementation undesirable in practice. In other words, SSEP generally allows recovering a set of coefficients $\{\tilde{\beta}_{i,j}, \tilde{\gamma}_{i,j}\}$ of a

first-order method

$$x_i = x_0 - \sum_{j=1}^{i-1} \frac{\tilde{\gamma}_{i,j}}{\tilde{\gamma}_{i,i}}(x_j - x_0) - \sum_{j=0}^{i-1} \frac{\tilde{\beta}_{i,j}}{\tilde{\gamma}_{i,i}} f'(x_j), \quad i = 1, \ldots, N, \qquad (20)$$

that enjoys the same worst-case guarantees as GFOM; however, using such a method requires, in general, keeping track of all coefficients $\{\tilde{\beta}_{i,j}, \tilde{\gamma}_{i,j}\}$ and all previous iterates and gradients $\{x_j, f'(x_j)\}$.

An approach for overcoming those drawbacks originates from the observation that in practically all situations we encountered, the coefficients $\{\tilde{\beta}_{i,j}, \tilde{\gamma}_{i,j}\}$ resulting from an optimal solution of (dual-PEP-GFOM) enjoyed advantageous structural properties allowing to *factor* the parameters $\{\tilde{\beta}_{i,j}, \tilde{\gamma}_{i,j}\}$ in a way that (a) does not require storing all coefficients (those coefficients are essentially *separable* in i and j), and (b) does not require storing all previous iterates and gradients in memory.

Below we present one such algorithm structure, and demonstrate how its parameters can be extracted from the numerical solution of (dual-PEP-GFOM).

Factorization of the SSEP-based method for strongly convex minimization   As discussed above, one can observe that optimal solutions to (dual-PEP-GFOM) often enjoys certain advantageous structural properties. In particular, in the case of smooth strongly convex optimization, instances of the SSEP method fit the form (21) below within high accuracy levels. Note that the form (21) is a straightforward generalization of OGM, which is the SSEP method derived for the non-strongly convex case, $\mu = 0$, established in Sect. 4.2 (see e.g. [22, Section 7.1]).

SSEP-based gradient method for smooth strongly convex minimization

Input: $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$, initial guess $y_0 = x_0 \in \mathbb{R}^d$, maximum number of iterations N.

For $i = 1, \ldots, N$:

$$\begin{aligned} y_i &= x_{i-1} - \frac{1}{L} f'(x_{i-1}) \\ x_i &= y_i + \zeta_i(y_i - y_{i-1}) + \eta_i(y_i - x_{i-1}) \end{aligned} \qquad (21)$$

Output: x_N.

Let us quickly describe a procedure for deriving candidate values for $\{(\eta_i, \zeta_i)\}$ to be used in (21) given a set $\{\tilde{\beta}_{i,j}, \tilde{\gamma}_{i,j}\}$ of numerically-derived coefficients for (20). We start the procedure by recursively eliminating all instances of x_j with $j \geq 1$, from the right-hand-side of algorithm (20) using their definition, reaching a canonical form

$$x_i = x_0 - \sum_{j=0}^{i-1} h_{i,j} f'(x_j).$$

This representation is important as all fixed-step first-order methods have a unique representation using $\{h_{i,j}\}$, but generally no unique representation in terms of $\{\tilde{\beta}_{i,j}, \tilde{\gamma}_{i,j}\}$. Then, one can write algorithm (21) in the same format with step-sizes $\{h'_{i,j}\}$, which in this case should satisfy

$$
\begin{aligned}
&h'_{-1,j} = h'_{0,j} = 0 \text{ for all } j \\
&h'_{i,j} = \left\{
\begin{array}{ll}
\frac{1}{L}(1 + \zeta_i + \eta_i) & \text{if } j = i-1 \\
h'_{i-1,j}(1 + \zeta_i) - \frac{1}{L}\zeta_i & \text{if } j = i-2 \\
h'_{i-1,j} + \zeta_i(h'_{i-1,j} - h'_{i-2,j}) & \text{otherwise.}
\end{array}
\right\} \text{ for } i \geq 1.
\end{aligned}
\tag{22}
$$

Now, assuming that both implementations (20) and (21) describe the same algorithm, we have, in particular, $h_{i,i-1} = h'_{i,i-1}$ and $h_{i,i-2} = h'_{i,i-2}$. As a result, one can identify candidate values for η_i and ζ_i as follows:

$$
\begin{aligned}
\zeta_i &= \frac{h_{i,i-2} - h_{i-1,i-2}}{h_{i-1,i-2} - 1/L} \\
\eta_i &= L\, h_{i,i-1} - 1 - \zeta_i,
\end{aligned}
$$

where we also arbitrarily set $\zeta_0 = 0$. Finally, one can numerically verify that $h_{i,j} \approx h'_{i,j}$ indeed holds for all i, j by generating the full set $\{h'_{i,j}\}$ using (22) and the candidate values $\{(\eta_i, \zeta_i)\}$.

The numerical values of $\{(\eta_i, \zeta_i)\}$ derived in the case of $N = 10$ for a few values of κ's are reported in Table 1. Since the approach is limited by our capability to accurately solve SDPs, it is important to note that PEPs can be used again for validating the performances of the final method (shown in the one before last row in Table 1).

Comparison to other methods Figure 1 presents a comparison of the worst-case bounds in the case $\kappa = 100$ for GFOM, for an SSEP-based method performing no line searches, the celebrated fast (or accelerated) gradient method (FGM) for smooth strongly convex minimization [36, Theorem 2.1.12], and the very recent triple momentum method (TMM) [52]. These worst-case bounds were derived numerically by solving the corresponding PEPs using the interpolation conditions presented in Example 1 (see [13,50] for details on the derivation of PEPs for fixed-step methods). Note that the bound for the SSEP method was generated for the form (20); bounds for the efficient form (21) behave almost exactly like the bounds for the form (20)—the difference could not be observed from this plot—and were therefore omitted from the comparison.

Finally, the numerical results presented above support the conjectures that the subspace-searching GFOM enjoys an $O((1 - \sqrt{2}\sqrt{\kappa^{-1}})^N R_x^2)$ rate of convergence while the SSEP-based method that does not perform any line searches, converges at the faster rate $O((1 - 2\sqrt{\kappa^{-1}})^N R_x^2)$ (matching the rate of convergence bounds on TMM [52, Theorem 1]). This is not in contraction with the theory, as noted in Remark 3, however, we are currently unable to find an intuitive explanation to this phenomenon besides the algebraic observation that the PEPs corresponding to the methods (20) and (21) have less degrees of freedom than the PEP for GFOM.

6 Conclusion

The main goal of this work is to provide a systematic and efficient approach for designing first-order optimization algorithms. The contribution is essentially threefold: first, we extend the performance estimation framework for obtaining worst-case guarantees for a greedy method that performs arbitrary subspace searches, and show that the generated guarantees are tight in the large-scale setting under some weak assumptions. Then, we describe a methodology for systematically designing fixed-step methods that share the same worst-case guarantees as the subspace searching greedy method. Finally, based on the new methodology, we derive optimal methods for non-smooth and smooth convex minimization and versions of these methods that do not require prior knowledge of the problem parameters.

As illustrated in Sect. 5, the methodology can also serve in cases where numerical results cannot easily be converted to practical analytical optimization schemes. For example, for real-time embedded optimization [10], where it is acceptable to spend some time performing pre-computations in order to more efficiently perform simple repetitive routines.

Direct extensions of the approach include considering additional families of functions and oracles, such as composite functions involving proximal terms and projections [4,38], inexactness [6,8,9,44], stochastic oracles arising in finite sums [7,20,24], or block-coordinate descent [37,54]. Even further extensions include considering additional variants of the greedy method, e.g., using alternative optimality criteria, as the distance to an optimal point.

Implementation All the worst-case performance analyses presented above were numerically validated using the PESTO toolbox [49]. The code for reproducing the worst-case guarantees is available at

https://github.com/AdrienTaylor/GreedyMethods

Numerical experiments were produced using CVX and YALMIP [17,27] along with MOSEK [28].

Table 1 Parameters for the implementation (21) of the SSEP-based gradient methods for smooth strongly convex minimization with $N = 10$ and $\kappa \in \{\infty, 1000, 100, 50\}$ recovered using MOSEK [28], and the associated worst-case guarantees on $f(x_{10}) - f(x_*)$. The worst-case objective function accuracy for GFOM and for the vanilla (unfactored) SSEP method (20) are also given for validation purposes. Finally, we provide the absolute inaccuracy observed between the coefficients $h_{i,j}$ of the vanilla SSEP method (20) versus the corresponding coefficients $h'_{i,j}$ obtained for the factored method (21)

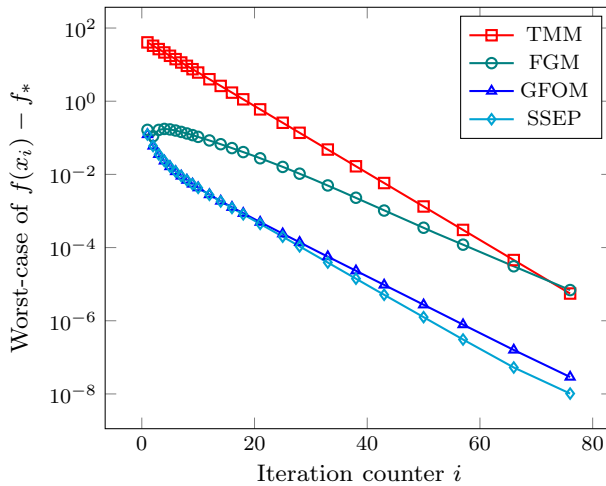| i	$\kappa = \infty$		$\kappa = 1000$		$\kappa = 100$		$\kappa = 50$	
	ζ_i	η_i	ζ_i	η_i	ζ_i	η_i	ζ_i	η_i
1	0	0.6180	0	0.6173	0	0.6110	0	0.6039
2	0.2818	0.7376	0.2810	0.7365	0.2744	0.7259	0.2671	0.7142
3	0.4340	0.7977	0.4325	0.7960	0.4184	0.7812	0.4030	0.7650
4	0.5311	0.8346	0.5286	0.8324	0.5068	0.8132	0.4835	0.7929
5	0.5988	0.8597	0.5954	0.8570	0.5661	0.8339	0.5352	0.8099
6	0.6489	0.8780	0.6448	0.8750	0.6085	0.8485	0.5708	0.8216
7	0.6876	0.8920	0.6829	0.8888	0.6413	0.8607	0.5981	0.8321
8	0.7185	0.9030	0.7136	0.9001	0.6701	0.8738	0.6240	0.8462
9	0.7437	0.9120	0.7392	0.9097	0.6985	0.8892	0.6539	0.8663
10	0.5542	0.6663	0.5514	0.6652	0.5265	0.6553	0.4988	0.6441
$f(\mathrm{GFOM}_{10}(f, x_0)) - f_* \leq$	$\frac{L\|x_0 - x_*\|^2}{159.07}$		$\frac{L\|x_0 - x_*\|^2}{164.95}$		$\frac{L\|x_0 - x_*\|^2}{230.87}$		$\frac{L\|x_0 - x_*\|^2}{340.41}$	
Using (20), $f(x_{10}) - f_* \leq$	$\frac{L\|x_0 - x_*\|^2}{159.07}$		$\frac{L\|x_0 - x_*\|^2}{165.04}$		$\frac{L\|x_0 - x_*\|^2}{232.86}$		$\frac{L\|x_0 - x_*\|^2}{347.88}$	
Using (21), $f(x_{10}) - f_* \leq$	$\frac{L\|x_0 - x_*\|^2}{159.07}$		$\frac{L\|x_0 - x_*\|^2}{165.04}$		$\frac{L\|x_0 - x_*\|^2}{232.86}$		$\frac{L\|x_0 - x_*\|^2}{347.88}$	
$\max_{i,j} \lvert h_{i,j} - h'_{i,j} \rvert \approx$	8.2×10^{-9}		2.1×10^{-6}		1.8×10^{-6}		9.1×10^{-6}	

Fig. 1 Numerical approximation of the worst-case performance of the FGM, TMM [52], GFOM and the SSEP method with no line-search for minimizing a 1-smooth 0.01-strongly convex function for instances with $\|x_0 - x_*\| \leq 1$. For the fast gradient method, we used the inertial parameter $(1 - \sqrt{\mu/L})/(1 + \sqrt{\mu/L})$

Appendix A: Proof of Lemma 1

We start the proof of Lemma 1 with the following a technical lemma.

Lemma 5 *Let $\mathcal{F}$ be a class of contraction-preserving c.c.p. functions (see Definition 3), and let $S = \{(x_i, g_i, f_i)\}_{i \in I_N^*}$ be an $\mathcal{F}$-interpolable set satisfying*

$$\langle g_i, g_j \rangle = 0, \quad \text{for all } 0 \leq j < i = 1, \ldots, N, \tag{23}$$

$$\langle g_i, x_j - x_0 \rangle = 0, \quad \text{for all } 1 \leq j \leq i = 1, \ldots, N, \tag{24}$$

then there exists $\{\hat{x}_i\}_{i \in I_N^*} \subset \mathbb{R}^d$ such that the set $\hat{S} = \{(\hat{x}_i, g_i, f_i)\}_{i \in I_N^*}$ is $\mathcal{F}$-interpolable, and

$$\|\hat{x}_0 - \hat{x}_*\| \leq \|x_0 - x_*\|, \tag{25}$$

$$\hat{x}_i \in \hat{x}_0 + \text{span}\{g_0, \ldots, g_{i-1}\}, \quad i = 0, \ldots, N. \tag{26}$$

Proof By the orthogonal decomposition theorem there exists $\{h_{i,j}\}_{0 \leq j < i \leq N} \subset \mathbb{R}$ and $\{v_i\}_{0 \leq i \leq N} \subset \mathbb{R}^d$ with $\langle g_k, v_i \rangle = 0$ for all $0 \leq k < i \leq N$ such that

$$x_i = x_0 - \sum_{j=0}^{i-1} h_{i,j} g_j + v_i, \quad i = 0, \ldots, N,$$

furthermore, there exist $r_* \in \mathbb{R}^d$ satisfying $\langle r_*, v_j \rangle = 0$ for all $0 \le j \le N$ and some $\{v_j\}_{0 \le j \le N} \subset \mathbb{R}$, such that

$$x_* = x_0 + \sum_{j=0}^{N} v_j v_j + r_*.$$

By (23) and (24) it then follows that for all $k \ge i$

$$\langle g_k, v_i \rangle = \left\langle g_k, x_i - x_0 + \sum_{j=0}^{i-1} h_{i,j} g_j \right\rangle = 0,$$

hence, together with the definition of v_i, we get

$$\langle g_k, v_i \rangle = 0, \quad i, k = 0, \ldots, N. \tag{27}$$

Let us now choose $\{\hat{x}_i\}_{i \in I_N^*}$ as follows:

$$\hat{x}_0 := x_0,$$
$$\hat{x}_i := x_0 - \sum_{j=0}^{i-1} h_{i,j} g_j, \quad i = 0, \ldots, N,$$
$$\hat{x}_* := x_0 + r_*.$$

It follows immediately from this definition that (26) holds, it thus remains to show that $\hat{S}$ is $\mathcal{F}$-interpolable and that (25) holds.

In order to establish that $\hat{S}$ is $\mathcal{F}$-interpolable, from Definition 3 it is enough to show that the conditions in (4) are satisfied. This is indeed the case, as $\langle g_j, \hat{x}_i - \hat{x}_0 \rangle = \langle g_j, x_i - x_0 \rangle$ follows directly from definition of $\{\hat{x}_i\}$ and (27), whereas $\|\hat{x}_i - \hat{x}_j\| \le \|x_i - x_j\|$ in the case $i, j \ne *$ follows from

$$\begin{aligned}
\|x_i - x_j\|^2 &= \left\| x_0 - \sum_{k=0}^{i-1} h_{i,k} g_k + v_i - x_0 + \sum_{k=0}^{j-1} h_{j,k} g_k - v_j \right\|^2 \\
&= \|\hat{x}_i - \hat{x}_j\|^2 + \|v_i - v_j\|^2 \\
&\ge \|\hat{x}_i - \hat{x}_j\|^2, \quad i, j = 0, \ldots, N,
\end{aligned}$$

and in the case $j = *$, follows from

$$\|x_i - x_*\|^2 = \left\| x_0 - \sum_{k=0}^{i-1} h_{i,k} g_k + v_i - x_0 - \sum_{j=0}^{N} v_j v_j - r_* \right\|^2$$

$$= \left\| \hat{x}_i - \hat{x}_* \right\|^2 + \left\| v_i - \sum_{j=0}^{N} v_j v_j \right\|^2$$

$$\geq \left\| \hat{x}_i - \hat{x}_* \right\|^2, \quad i = 0, \dots, N,$$

where for the second equality we used $\langle v_i, r_* \rangle = 0$. The last inequality also establishes (25), which completes the proof. □

Proof of Lemma 1 By the first-order necessary and sufficient optimality conditions (see e.g., [42, Theorem 3.5]), the definitions of x_i and $f'(x_i)$ in (5) and (6) can be equivalently defined as a solution to the problem of finding $x_i \in \mathbb{R}^d$ and $f'(x_i) \in \partial f(x_i)$ ($0 \leq i \leq N$), that satisfy:

$$\langle f'(x_i), f'(x_j) \rangle = 0, \quad \text{for all } 0 \leq j < i = 1, \dots, N,$$
$$x_i \in x_0 + \text{span}\{f'(x_0), \dots, f'(x_{i-1})\}, \quad \text{for all } i = 1, \dots, N,$$

hence the problem (PEP) can be equivalently expressed as follows:

$$\sup_{f, \{x_i\}_{i \in I_N^*}, \{f'(x_i)\}_{i \in I_N^*}} f(x_N) - f_*$$

$$\text{subject to: } f \in \mathcal{F}(\mathbb{R}^d), \ x_* \text{ is a minimizer of } f,$$
$$f'(x_i) \in \partial f(x_i), \quad \text{for all } i \in I_N^*,$$
$$\|x_0 - x_*\| \leq R_x,$$
$$\langle f'(x_i), f'(x_j) \rangle = 0, \quad \text{for all } 0 \leq j < i = 1, \dots, N,$$
$$x_i \in x_0 + \text{span}\{f'(x_0), \dots, f'(x_{i-1})\}, \quad \text{for all } i = 1, \dots, N.$$
$$\tag{28}$$

Now, since all constraints in (28) depend only on the first-order information of f at $\{x_i\}_{i \in I_N^*}$, by taking advantage of Definition 2 we can denote $f_i := f(x_i)$ and $g_i := f'(x_i)$ and treat these and as optimization variables, thereby reaching the following *equivalent* formulation

$$\sup_{\{(x_i, g_i, f_i)\}_{i \in I_N^*}} f_N - f_*$$

subject to: $\{(x_i, g_i, f_i)\}_{i \in I_N^*}$ is $\mathcal{F}(\mathbb{R}^d)$-interpolable,

$$\|x_0 - x_*\| \le R_x,$$
$$g_* = 0,$$
$$\langle g_i, g_j \rangle = 0, \text{ for all } 0 \le j < i = 1, \ldots N,$$
$$x_i \in x_0 + \text{span}\{g_0, \ldots, g_{i-1}\}, \quad \text{for all } i = 1, \ldots, N. \tag{29}$$

Since (PEP-GFOM) is a relaxation of (29), we get

$$f(x_N) - f_* \le \text{val (PEP)} \le \text{val (PEP-GFOM)},$$

which establishes the bound (13).

In order to establish the second part of the claim, let $\varepsilon > 0$. We will proceed to show that there exists some valid input for GFOM (f, x_0), such that $f(\text{GFOM}_N(f, x_0)) - f_* \ge \text{val (PEP-GFOM)} - \varepsilon$.

Indeed, by the definition of (PEP-GFOM), there exists a set $S = \{(x_i, g_i, f_i)\}_{i \in I_N^*}$ that satisfies the constraints in (PEP-GFOM) and reaches an objective value $f_N - f_* \ge \text{val (PEP-GFOM)} - \varepsilon$. Since S satisfies the requirements of Lemma 5 [as these requirements are constraints in (PEP-GFOM)], there exists a set of vectors $\{\hat{x}_i\}_{i \in I_N^*}$ for which

$$\|\hat{x}_0 - \hat{x}_*\| \le R_x,$$
$$\hat{x}_i \in \hat{x}_0 + \text{span}\{g_0, \ldots, g_{i-1}\}, \quad i = 0, \ldots, N,$$

hold, and in addition, $\hat{S} := \{(\hat{x}_i, g_i, f_i)\}_{i \in I_N^*}$ is $\mathcal{F}(\mathbb{R}^d)$-interpolable. By definition of an $\mathcal{F}(\mathbb{R}^d)$-interpolable set, it follows that there exists a function $\hat{f} \in \mathcal{F}(\mathbb{R}^d)$ such that $\hat{f}(\hat{x}_i) = f_i$, $g_i \in \partial \hat{f}(\hat{x}_i)$, hence satisfying

$$\left\langle \hat{f}'(\hat{x}_i), \hat{f}'(\hat{x}_j) \right\rangle = 0, \quad \text{for all } 0 \le j < i = 1, \ldots, N,$$
$$\hat{x}_i \in \hat{x}_0 + \text{span}\{\hat{f}'(x_0), \ldots, \hat{f}'(\hat{x}_{i-1})\}, \quad \text{for all } i = 1, \ldots, N.$$

Furthermore, since $g_* = 0$ we have that $\hat{x}_*$ is an optimal solution of $\hat{f}$.

We conclude that the sequence $\hat{x}_0, \ldots, \hat{x}_N$ forms a valid execution of GFOM on the input $(\hat{f}, \hat{x}_0)$, that the requirement $\|\hat{x}_0 - \hat{x}_*\| \le R_x$ is satisfied, and that the output of the method, $\hat{x}_N$, attains the absolute inaccuracy value of $\hat{f}(\hat{x}_N) - \hat{f}(\hat{x}_*) = f_N - f_* \ge \text{val (PEP-GFOM)} - \varepsilon$. □

Appendix B: Proof of Theorem 3

Lemma 6 *Suppose there exists a pair (f, x_0) such that $f \in \mathcal{F}$, $\|x_0 - x_*\| \le R_x$ and $\text{GFOM}_{2N+1}(f, x_0)$ is not optimal for f, then (sdp-PEP-GFOM) satisfies*

Slater's condition. In particular, no duality gap occurs between the primal-dual pair (sdp-PEP-GFOM)*,* (dual-PEP-GFOM)*, and the dual optimal value is attained.*

Proof Let (f, x_0) be a pair satisfying the premise of the lemma and denote by $\{x_i\}_{i \geq 0}$ the sequence generated according to GFOM and by $\{f'(x_i)\}_{i \geq 0}$ the subgradients chosen at each iteration of the method, respectively. By the assumption that the optimal value is not obtained after $2N + 1$ iterations, we have $f(x_{2N+1}) > f_*$.

We show that the set $\{(\tilde{x}_i, \tilde{g}_i, \tilde{f}_i)\}_{i \in I_N^*}$ with

$$
\begin{aligned}
\tilde{x}_i &:= x_{2i}, \quad i = 0, \ldots, N, \\
\tilde{x}_* &:= x_*, \\
\tilde{g}_i &:= f'(x_{2i}), \quad i = 0, \ldots, N, \\
\tilde{g}_* &:= 0, \\
\tilde{f}_i &:= f(x_{2i}), \quad i = 0, \ldots, N, \\
\tilde{f}_* &:= f(x_*),
\end{aligned}
$$

corresponds to a Slater point for (sdp-PEP-GFOM).

In order to proceed, we consider the Gram matrix $\tilde{G}$ and the vector $\tilde{F}$ constructed from the set $\{(\tilde{x}_i, \tilde{g}_i, \tilde{f}_i)\}_{i \in I_N^*}$ as in Sect. 3.2. We then continue in two steps:

(i) we show that $(\tilde{G}, \tilde{F})$ is feasible for (sdp-PEP-GFOM),
(ii) we show that $\tilde{G} \succ 0$.

The proofs follow.

(i) First, we note that the set $\{(\tilde{x}_i, \tilde{g}_i, \tilde{f}_i)\}_{i \in I_N^*}$ satisfies the interpolation conditions for $\mathcal{F}$, as it was obtained by taking the values and gradients of a function in $\mathcal{F}$. Furthermore, since $\tilde{x}_0 = x_0$ and $\tilde{x}_* = x_*$ we also get that the initial condition $\|\tilde{x}_0 - \tilde{x}_*\| \leq R_x$ is respected, and since $\{x_i\}$ correspond to the iterates of GFOM, we also have by Lemma 5 that

$$
\begin{aligned}
\langle \tilde{g}_i, \tilde{g}_j \rangle &= 0, \quad \text{for all } 0 \leq j < i = 1, \ldots N, \\
\langle \tilde{g}_i, \tilde{x}_j - \tilde{x}_0 \rangle &= 0, \quad \text{for all } 1 \leq j \leq i = 1, \ldots N.
\end{aligned}
$$

It then follows from the construction of $\tilde{G}$ and $\tilde{F}$ and by (10) that $\tilde{G}$ and $\tilde{F}$ satisfies the constrains of (sdp-PEP-GFOM).

(ii) In order to establish that $\tilde{G} \succ 0$ it suffices to show that the vectors

$$
\{\tilde{g}_0, \ldots, \tilde{g}_N; \tilde{x}_1 - \tilde{x}_0, \ldots, \tilde{x}_N - \tilde{x}_0; \tilde{x}_* - \tilde{x}_0\}
$$

are linearly independent. Indeed, this follows from Lemma 5, since these vectors are all non-zero, and since $\tilde{x}_*$ does not fall in the linear space spanned by $\tilde{g}_0, \ldots, \tilde{g}_N; \tilde{x}_1 - \tilde{x}_0, \ldots, \tilde{x}_N - \tilde{x}_0$ (as otherwise x_{2N+1} would be an optimal solution).

We conclude that $(\tilde{G}, \tilde{F})$ forms a Slater point for (sdp-PEP-GFOM). □

Proof of Theorem 3 The bound follows directly from

$$f(\text{GFOM}_N(f, x_0)) - f_* \leq \text{val}(\text{PEP-GFOM}) \leq \text{val}(\text{sdp-PEP-GFOM}),$$

established by Lemmas 1 and 2. The tightness claim follows from the tightness claims of Lemmas 1, 2 and 6. □

Appendix C: Proof of Theorem 4

We begin the proof of Theorem 4 by recalling a well-known lemma on constraint aggregation, showing that it is possible to aggregate the constraints of a minimization problem while keeping the optimal value of the resulting program bounded from below.

Lemma 7 *Consider the problem*

$$w := \min\{f(x) : h(x) = 0, g(x) \leq 0\}, \tag{P}$$

where $f : \mathbb{R}^d \to \mathbb{R}$, $h : \mathbb{R}^d \to \mathbb{R}^n$, $g : \mathbb{R}^d \to \mathbb{R}^m$ are some (not necessarily convex) functions, and suppose $(\tilde{\alpha}, \tilde{\beta}) \in \mathbb{R}^n \times \mathbb{R}_+^m$ is a feasible point for the Lagrangian dual of (P) *that attains the value $\tilde{\omega}$. Let $k \in \mathbb{N}$, and let $M \in \mathbb{R}^{n \times k}$ be a linear map such that $\tilde{\alpha} \in \text{range}(M)$, then*

$$w' := \min\{f(x) : M^\top h(x) = 0, g(x) \leq 0\} \tag{P'}$$

is bounded from below by $\tilde{\omega}$.

Proof Let

$$L(x, \alpha, \beta) = f(x) + \alpha^\top h(x) + \beta^\top g(x)$$

be the Lagrangian for the problem (P), then by the assumption on $(\tilde{\alpha}, \tilde{\beta})$ we have $\min_x L(x, \tilde{\alpha}, \tilde{\beta}) = \tilde{\omega}$. Now, let $u \in \mathbb{R}^k$ be some vector such that $Mu = \tilde{\alpha}$, then for every x in the domain of (P')

$$\tilde{\alpha}^\top h(x) = u^\top M^\top h(x) = 0,$$
$$\tilde{\beta}^\top g(x) \leq 0,$$

where that last inequality follows from nonnegativity of $\tilde{\beta}$. We get

$$f(x) \geq f(x) + \tilde{\alpha}^\top h(x) + \tilde{\beta}^\top g(x) = L(x, \tilde{\alpha}, \tilde{\beta}) \geq \tilde{\omega}, \quad \forall x : M^\top h(x) = 0, g(x) \leq 0,$$

and thus the desired result $w' \geq \tilde{\omega}$ holds. □

Before proceeding with the proof of the main results, let us first formulate a performance estimation problem for the class of methods described by (14).

Lemma 8 *Let $R_x \geq 0$ and let $\{\beta_{i,j}\}_{1 \leq i \leq N, 0 \leq j \leq i-1}$, $\{\gamma_{i,j}\}_{1 \leq i \leq N, 1 \leq j \leq i}$ be some given sets of real numbers, then for any pair (f, x_0) such that $f \in \mathcal{F}(\mathbb{R}^d)$ and $\|x_0 - x_*\| \leq R_x$ (where $x_* \in \arg\min_x f(x)$). Then for any sequence $\{x_i\}_{1 \leq i \leq N}$ that satisfies*

$$\left\langle f'(x_i), \sum_{j=0}^{i-1} \beta_{i,j} f'(x_j) + \sum_{j=1}^{i} \gamma_{i,j}(x_j - x_0) \right\rangle = 0, \quad i = 1, \dots, N \qquad (30)$$

for some $f'(x_i) \in \partial f(x_i)$, the following bound holds:

$$f(x_N) - f_* \leq \sup_{F \in \mathbb{R}^{N+1}, G \in \mathbb{R}^{2N+2 \times 2N+2}} F^\top \mathbf{f}_N - F^\top \mathbf{f}_*$$

$$\text{subject to:} \quad \mathrm{Tr}\left(A_k^{\mathrm{ic}} G\right) + (a_k^{\mathrm{ic}})^\top F + b_k^{\mathrm{ic}} \leq 0, \quad \text{for all } k \in K_N,$$

$$\left\langle \mathbf{g}_i, \sum_{j=0}^{i-1} \beta_{i,j} \mathbf{g}_j + \sum_{j=1}^{i} \gamma_{i,j}(\mathbf{x}_j - \mathbf{x}_0) \right\rangle_G = 0, \quad \text{for all } i = 1, \dots N,$$

$$\|\mathbf{x}_0 - \mathbf{x}_*\|_G^2 - R_x^2 \leq 0,$$

$$G \succeq 0.$$

We omit the proof since it follows the exact same lines as for (sdp-PEP-GFOM) (c.f. the derivations in [13,50]).

Proof of Theorem 4 The key observation underlying the proof is that by taking the PEP for GFOM (sdp-PEP-GFOM) and aggregating the constraints that define its iterates, we can reach a PEP for the class of methods (14). Furthermore, by Lemma 7, this aggregation can be done in a way that maintains the optimal value of the program, thereby reaching a specific method in this class whose corresponding PEP attains an optimal value that is at least as good as that of the PEP for GFOM.

We perform the aggregation of the constraints as follows: for all $i = 1, \dots, N$ we aggregate the constraints which correspond to $\{\beta_{i,j}\}_{0 \leq j < i}$, $\{\gamma_{i,j}\}_{1 \leq j \leq i}$ (weighted by $\{\tilde{\beta}_{i,j}\}_{0 \leq j < i}$, $\{\tilde{\gamma}_{i,j}\}_{1 \leq j \leq i}$, respectively) into a single constraint, reaching

$$w'(N, \mathcal{F}(\mathbb{R}^d), R_f, R_x, R_g) := \sup_{F \in \mathbb{R}^{N+1}, G \in \mathbb{R}^{2N+2 \times 2N+2}} F^\top \mathbf{f}_N - F^\top \mathbf{f}_*$$

$$\text{(PEP-SSEP}(\mathcal{F}(\mathbb{R}^d)))$$

$$\text{subject to:} \quad \mathrm{Tr}\left(A_k^{\mathrm{ic}} G\right) + (a_k^{\mathrm{ic}})^\top F + b_k^{\mathrm{ic}} \leq 0, \quad \text{for all } k \in K_N,$$

$$\left\langle \mathbf{g}_i, \sum_{j=0}^{i-1} \tilde{\beta}_{i,j} \mathbf{g}_j + \sum_{j=1}^{i} \tilde{\gamma}_{i,j}(\mathbf{x}_j - \mathbf{x}_0) \right\rangle_G = 0, \quad \text{for all } i = 1, \dots N,$$

$$\|\mathbf{x}_0 - \mathbf{x}_*\|_G^2 - R_x^2 \leq 0,$$

$$G \succeq 0.$$

By Lemma 7 and the choice of weights $\{\tilde{\beta}_{i,j}\}_{0 \leq j < i}$, $\{\tilde{\gamma}_{i,j}\}_{1 \leq j \leq i}$ it follows that

$$w'(N, \mathcal{F}(\mathbb{R}^d), R_x) \leq \tilde{\omega}.$$

Finally, by Lemma 8, we conclude that $w'(N, \mathcal{F}(\mathbb{R}^d), R_x)$ forms an upper bound on the performance of the method (14), i.e., for any valid pair (f, x_0) and any $\{x_i\}_{i \geq 0}$ that satisfies (14) we have

$$f(x_N) - f_* \leq w'(N, \mathcal{F}(\mathbb{R}^d), R_x) \leq \tilde{\omega}.$$

$\square$

References

1. Arjevani, Y., Shalev-Shwartz, S., Shamir, O.: On lower and upper bounds in smooth and strongly convex optimization. J. Mach. Learn. Res. **17**(126), 1–51 (2016)
2. Beck, A.: Quadratic matrix programming. SIAM J. Optim. **17**(4), 1224–1238 (2007)
3. Beck, A., Drori, Y., Teboulle, M.: A new semidefinite programming relaxation scheme for a class of quadratic matrix problems. Oper. Res. Lett. **40**(4), 298–302 (2012)
4. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
5. Bubeck, S., Lee, Y.T., Singh, M.: A geometric alternative to Nesterov's accelerated gradient descent (2015). arXiv preprint arXiv:1506.08187
6. De Klerk, E., Glineur, F., Taylor, A.B.: On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. Optim. Lett. **11**(7), 1185–1199 (2017)
7. Defazio, A., Bach, F., Lacoste-Julien, S.: SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. In: Advances in Neural Information Processing Systems (NIPS), pp. 1646–1654 (2014)
8. Devolder, O., Glineur, F., Nesterov, Y.: Intermediate gradient methods for smooth convex problems with inexact oracle. Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), Technical report (2013)
9. Devolder, O., Glineur, F., Nesterov, Y.: First-order methods of smooth convex optimization with inexact oracle. Math. Program. **146**(1–2), 37–75 (2014)
10. Diehl, M., Ferreau, H.J., Haverbeke, N.: Efficient numerical methods for nonlinear MPC and moving horizon estimation. Nonlinear Model Predict. Control **384**, 391–417 (2009)
11. Drori, Y.: Contributions to the complexity analysis of optimization algorithms. Ph.D. thesis, Tel-Aviv University (2014)
12. Drori, Y.: The exact information-based complexity of smooth convex minimization. J. Complex. **39**, 1–16 (2017)
13. Drori, Y., Teboulle, M.: Performance of first-order methods for smooth convex minimization: a novel approach. Math. Program. **145**(1–2), 451–482 (2014)
14. Drori, Y., Teboulle, M.: An optimal variant of Kelley's cutting-plane method. Math. Program. **160**(1–2), 321–351 (2016)
15. Drusvyatskiy, D., Fazel, M., Roy, S.: An optimal first order method based on optimal quadratic averaging. SIAM J. Optim. **28**(1), 251–271 (2018)
16. Fazlyab, M., Ribeiro, A., Morari, M., Preciado, V.M.: Analysis of optimization algorithms via integral quadratic constraints: nonstrongly convex problems. SIAM J. Optim. **28**(3), 2654–2689 (2018)
17. Grant, M., Boyd, S.: CVX: Matlab software for disciplined convex programming. version 2.0 beta. http://cvxr.com/cvx (2013)
18. Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Res. Natl. Bureau Stand. **49**(6), 409–436 (1952)
19. Hu, B., Lessard, L.: Dissipativity theory for Nesterov's accelerated method. In: International Conference on Machine Learning (ICML), pp. 1549–1557 (2017)
20. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: Advances in Neural Information Processing Systems (NIPS), pp. 315–323 (2013)
21. Karimi, S., Vavasis, S.A.: A unified convergence bound for conjugate gradient and accelerated gradient. (2016). arXiv preprint arXiv:1605.00320

22. Kim, D., Fessler, J.A.: Optimized first-order methods for smooth convex minimization. Math. Program. **159**(1–2), 81–107 (2016)
23. Kim, D., Fessler, J.A.: On the convergence analysis of the optimized gradient method. J. Optim. Theory Appl. **172**(1), 187–205 (2017)
24. Le Roux, N., Schmidt, M., Bach, F.: A stochastic gradient method with an exponential convergence rate for finite training sets. In: Advances in Neural Information Processing Systems (NIPS), pp. 2663–2671 (2012)
25. Lemaréchal, C., Sagastizábal, C.: Variable metric bundle methods: from conceptual to implementable forms. Math. Program. **76**(3), 393–410 (1997)
26. Lessard, L., Recht, B., Packard, A.: Analysis and design of optimization algorithms via integral quadratic constraints. SIAM J. Optim. **26**(1), 57–95 (2016)
27. Löfberg, J.: YALMIP: a toolbox for modeling and optimization in MATLAB. In: Proceedings of the CACSD Conference (2004)
28. Mosek, A.: The MOSEK Optimization Software, vol. 54 (2010). http://www.mosek.com
29. Narkiss, G., Zibulevsky, M.: Sequential subspace optimization method for large-scale unconstrained problems. In: Technion-IIT, Department of Electrical Engineering (2005)
30. Nemirovski, A.: Orth-method for smooth convex optimization. Izvestia AN SSSR **2**, 937–947 (1982). (in Russian)
31. Nemirovski, A.: Information-based complexity of linear operator equations. J. Complex. **8**(2), 153–175 (1992)
32. Nemirovski, A.: Prox-method with rate of convergence O(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. SIAM J. Optim. **15**(1), 229–251 (2004)
33. Nemirovski, A., Yudin, D.: Information-based complexity of mathematical programming. Izvestia AN SSSR, Ser. Tekhnicheskaya Kibernetika **1** (1983) **(in Russian)**
34. Nemirovski, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Willey-Interscience, New York (1983)
35. Nesterov, Y.: A method of solving a convex programming problem with convergence rate O($1/k^2$)). Soviet Mathematics Doklady **27**, 372–376 (1983)
36. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer, London (2004)
37. Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J. Optim. **22**(2), 341–362 (2012)
38. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
39. Nesterov, Y., Shikhman, V.: Quasi-monotone subgradient methods for nonsmooth convex minimization. J. Optim. Theory Appl. **165**(3), 917–940 (2015)
40. Polyak, B.T.: Some methods of speeding up the convergence of iteration methods. USSR Comput. Math. Math. Phys. **4**(5), 1–17 (1964)
41. Polyak, B.T.: Introduction to Optimization. Optimization Software, New York (1987)
42. Ruszczyński, A.P.: Nonlinear Optimization, vol. 13. Princeton University Press, Princeton (2006)
43. Ryu, E.K., Taylor, A.B., Bergeling, C., Giselsson, P.: Operator splitting performance estimation: tight contraction factors and optimal parameter selection (2018). arXiv preprint arXiv:1812.00146
44. Schmidt, M., Le Roux, N., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: Advances in Neural Information Processing Systems (NIPS), pp. 1458–1466 (2011)
45. Scieur, D., Roulet, V., Bach, F., d'Aspremont, A.: Integration methods and optimization algorithms. In: Advances in Neural Information Processing Systems (NIPS), pp. 1109–1118 (2017)
46. Su, W., Boyd, S., Candes, E.: A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. In: Advances in Neural Information Processing Systems (NIPS), pp. 2510–2518 (2014)
47. Taylor, A.: Convex interpolation and performance estimation of first-order methods for convex optimization. Ph.D. thesis, Université catholique de Louvain (2017)
48. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Exact worst-case performance of first-order methods for composite convex optimization. SIAM J. Optim. **27**(3), 1283–1313 (2017)
49. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Performance estimation toolbox (PESTO): automated worst-case analysis of first-order optimization methods. In: IEEE 56th Annual Conference on Decision and Control (CDC), pp. 1278–1283 (2017)

50. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Smooth strongly convex interpolation and exact worst-case performance of first-order methods. Math. Program. **161**(1–2), 307–345 (2017)
51. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Exact worst-case convergence rates of the proximal gradient method for composite convex minimization. J. Optim. Theory Appl. **178**(2), 455–476 (2018)
52. Van Scoy, B., Freeman, R.A., Lynch, K.M.: The fastest known globally convergent first-order method for minimizing strongly convex functions. IEEE Control Syst. Lett. **2**(1), 49–54 (2018)
53. Wilson, A.C., Recht, B., Jordan, M.I.: A Lyapunov analysis of momentum methods in optimization. (2016). arXiv preprint arXiv:1611.02635
54. Wright, S.: Coordinate descent algorithms. Math. Program. **151**(1), 3–34 (2015)
55. Wright, S., Nocedal, J.: Numerical optimization. Science **35**, 67–68 (1999)