An optimal gradient method for smooth strongly convex minimization

Adrien Taylor · Yoel Drori

Date of current version: May 3, 2021

Abstract We present an optimal gradient method for smooth strongly convex optimization. The method is optimal in the sense that its worst-case bound on the distance to an optimal point *exactly* matches the lower bound on the oracle complexity for the class of problems, meaning that no black-box first-order method can have a better worst-case guarantee without further assumptions on the class of problems at hand. In addition, we provide a constructive recipe for obtaining the algorithmic parameters of the method and illustrate that it can be used for deriving methods for other optimality criteria as well.

1 Introduction

Consider the unconstrained minimization problem

$$\min_{x\in\mathbb{R}^d} f(x), \tag{1}$$

where f is a smooth convex function strongly convex. For solving such problems, one can rely on black-box first-order methods, which iteratively acquire information about f by evaluating its gradient at a sequence of iterates. In this context, the question of designing first-order methods with good worst-case guarantees occupy an important place.

In this work, we provide a black-box first-order method, the Information-Theoretic Exact Method (ITEM), designed for minimizing smooth strongly convex functions. This method attains the lower bound on the oracle complexity, sometimes referred to as information-theoretic complexity [Nemirovskii, 1992], of smooth strongly convex minimization when optimality is measured by the distance of the method's output to an optimal solution.

Given an L-smooth μ-strongly convex function f with $0 < \mu < L$, the method can be concisely written as

$$\begin{aligned} y_k &= (1-\beta_k)z_k + \beta_k\left(y_{k-1} - \frac{1}{L}\nabla f(y_{k-1})\right)\\ z_{k+1} &= (1-q\delta_k)z_k + q\delta_k\left(y_k - \frac{1}{\mu}\nabla f(y_k)\right), \end{aligned} \tag{2}$$

A. Taylor acknowledges support from the European Research Council (grant SEQUOIA 724063).This work was funded in part by the french government under management of Agence Nationale de la recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

Adrien Taylor
INRIA, Département d'informatique de l'ENS, École normale supérieure, CNRS, PSL Research University, Paris, France
Email: adrien.taylor@inria.fr
Yoel Drori
Google Research Israel. Email: dyoel@google.com

where $q = \mu/L$ denotes the inverse condition number (note that as $\mu \to 0$ the alternate formulation below should be preferred for obvious numerical reasons). Both sequences $\{\beta_k\}$ and $\{\delta_k\}$ are parametrized by a sequence A_k, incorporating the dependency on the current iteration number, as follows

$$\beta_k = \frac{A_k}{(1-q)A_{k+1}}, \text{ and } \delta_k = \frac{1}{2}\frac{(1-q)^2 A_{k+1} - (1+q)A_k}{1+q+qA_k},$$

with $A_0 = 0$ and

$$A_{k+1} = \frac{(1+q)A_k + 2\left(1+\sqrt{(1+A_k)(1+qA_k)}\right)}{(1-q)^2}, \quad k \geq 0.$$

As shown in the following, this sequence allows to describe the worst-case performance of (2) as

$$\|z_N - x_\star\|^2 \leq \frac{1}{1+qA_N}\|x_0 - x_\star\|^2,$$

which is the exact lower bound for smooth strongly convex minimization, as obtained in [Drori and Taylor, 2021]. Therefore, no black-box first-order method can further improve this guarantee, and ITEM achieves the lower bound on the oracle (or information-theoretic) complexity for smooth strongly convex minimization. In addition, as $A_N \geq (1-\sqrt{q})^{-2N}$, this bound provides a guarantee that z_k strictly improves over z_0 with a worst-case convergence rate $(1-\sqrt{q})^2$.

ITEM is also closely related to other methods. In particular, when $\mu > 0$ and as $k \to \infty$, the method's parameters β_k and δ_k tends to those of the Triple Momentum Method (TMM) by Van Scoy et al. [2018], and in the case $\mu = 0$ the parameters correspond to those of the Optimized Gradient Method (OGM) of Kim and Fessler [2016], which exactly achieves a lower complexity bound for minimizing function values as established in Drori [2017]. Details on those relationships are provided in Section 2.2.

1.1 Related works

Lower bounds and accelerated methods. The method presented in this work is closely related to the celebrated fast gradient methods (FGMs) by Nesterov [1983, 2004]. Lyapunov and potential function-based analyses of FGMs were presented in many works, including in the original [Nesterov, 1983]. The analyses are usually tailored for the smooth convex minimization setting [Nesterov, 1983, Beck and Teboulle, 2009], for the smooth strongly convex one [Wilson et al., 2016, Bansal and Gupta, 2019], and sometimes deal with both simultaneously [Nesterov, 2004, Gasnikov and Nesterov, 2018]. In the large-scale quadratic smooth (possibly strongly) convex minimization setting, optimal worst-case accuracies are achieved by Chebyshev and conjugate gradient methods [Nemirovskii, 1992, Nemirovski, 1999].

Performance estimation problems. The idea of computing worst-case accuracy of a given method through semidefinite programming dates back to Drori and Teboulle [2014]. It was refined using the concept of convex interpolation in [Taylor et al., 2017b], which allows guaranteeing that worst-case accuracies provided by the semidefinite programs are tight (i.e., the worst-case guarantee corresponds to a matching example in the problem class). The approach was taken further in different directions for analyzing and designing numerical methods in different contexts. A very related line of works, initiated by [Lessard et al., 2016], presents such analyses from a control theoretic perspective, and corresponds to the problem of looking for Lyapunov functions. Those works hence rather target *asymptotic* properties of time-invariant numerical methods, which allows using smaller sized SDPs.

Optimized gradient methods. The method presented in this work was first obtained as a solution to a convex optimization problem, through an approach closely related to that taken by Drori and Teboulle [2014] and Kim and Fessler [2016] for obtaining the Optimized Gradient Method (OGM). The OGM for smooth convex minimization ($\mu = 0$), obtained by Kim and Fessler [2016], was obtained by explicitly choosing the step sizes of a method for minimizing an upper bound on the worst-case inaccuracy criterion. The resulting method was later proved to achieve the lower bound in [Drori, 2017]. When $\mu = 0$, optimal methods for

optimizing function value accuracy $f(x_N) - f_\star$ include the OGM [Kim and Fessler, 2016, Drori, 2017], and the conjugate gradient method [Drori and Taylor, 2020]. It is also worth mentioning that optimized methods can be developed for other criteria as well. In particular, optimized methods for gradient norms $\|\nabla f(x_N)\|^2$ are studied by Kim and Fessler [2020], in the smooth convex setting.

The *Triple Momentum Method* (TMM) [Van Scoy et al., 2018] was designed as an optimized gradient method through Lyapunov arguments, using an idea similar to that of the OGM, but for time-independent methods (i.e., whose coefficients do not depend on the iteration counter), for when $\mu > 0$. The method was originally obtained using the integral quadratic framework by Lessard et al. [2016]; see [Van Scoy et al., 2018] and [Lessard and Seiler, 2020]. The problem of devising optimized methods for smooth strongly convex minimization (with $\mu > 0$) is also addressed in [Zhou et al., 2020, Gramlich et al., 2020], which also recovers the TMM as a particular case in their analyses.

So far, it remained unclear how to conciliate both *optimal* methods, as the OGM is clearly not optimal anymore when $\mu > 0$ (its worst-case guarantees remain unchanged in the presence of strong convexity [Kim and Fessler, 2017]), and as the TTM is not defined when $\mu = 0$.

1.2 Organization

A worst-case analysis of the Information-Theoretic Exact Method is provided in Section 2. In Section 3, we describe a constructive approach that leads to the method and illustrate that it can be used for developing optimized methods for other performance criteria. We draw some conclusions in Section 4.

1.3 Preliminaries and notations

We use the standard notation $\langle \cdot ; \cdot \rangle : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ to denote the Euclidean inner product, and the corresponding induced Euclidean norm $\|\cdot\|$. Furthermore, we denote by $x_\star$ some optimal solution to (1) (which is unique if $\mu > 0$), and by $f_\star$ its optimal value. The class of L-smooth μ-strongly convex functions is standard and can be defined as follows.

Definition 1 Let $f : \mathbb{R}^d \to \mathbb{R}$ be a proper, closed, and convex function, and consider two constant $0 \leq \mu < L < \infty$. We say that f is L-smooth and μ-strongly convex, denoted $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$, if

- (L-smooth) for all $x, y \in \mathbb{R}^d$, it holds that $f(x) \leq f(y) + \langle \nabla f(y); x - y \rangle + \frac{L}{2}\|x - y\|^2$,
- (μ-strongly convex) for all $x, y \in \mathbb{R}^d$, it holds that $f(x) \geq f(y) + \langle \nabla f(y); x - y \rangle + \frac{\mu}{2}\|x - y\|^2$.

We simply denote $f \in \mathcal{F}_{\mu,L}$ when the dimension is either clear from the context or unspecified. In addition, we use $q := \mu/L$ the (inverse) condition number of the class (hence $0 \leq q < 1$), and do not explicitly treat the trivial cases $L = \mu$ for readability purposes.

Smooth strongly convex functions satisfy many inequalities, see e.g., [Nesterov, 2004, Theorem 2.1.5]. For the developments below, we need only one specific inequality characterizing functions in $\mathcal{F}_{\mu,L}$.

Theorem 1 *Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$. For all $x, y \in \mathbb{R}^d$, it holds that*

$$f(y) \geq f(x) + \langle \nabla f(x); y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 + \frac{\mu}{2(1-\mu/L)}\|x - y - \tfrac{1}{L}(\nabla f(x) - \nabla f(y))\|^2.$$

This inequality turns out to be key in proving worst-case guarantees for first-order methods applied on smooth strongly convex problems, due to the following result [Taylor et al., 2017b, Theorem 4].

Theorem 2 ($\mathcal{F}_{\mu,L}$-interpolation) *Let I be an index set and $S = \{(x_i, g_i, f_i)\}_{i \in I} \subseteq \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ be a set of triplets. There exists $f \in \mathcal{F}_{\mu,L}$ satisfying $f(x_i) = f_i$ and $g_i \in \partial f(x_i)$ for all $i \in I$ if and only if*

$$f_i \geq f_j + \langle g_j; x_i - x_j \rangle + \frac{1}{2L}\|g_i - g_j\|^2 + \frac{\mu}{2(1-\mu/L)}\|x_i - x_j - \tfrac{1}{L}(g_i - g_j)\|^2$$

holds for all $i, j \in I$.

2 An optimal gradient method

For our purposes, probably the most convenient formulation of ITEM, allowing a unified treatment for the case $\mu = 0$, is as presented in Algorithm 1.

Algorithm 1 Information-Theoretic Exact Method (ITEM)

Input: $f \in \mathcal{F}_{\mu,L}$ with $0 \leq \mu < L < \infty$, initial guess $x_0 \in \mathbb{R}^d$
Initialization: $y_{-1} = z_0 = x_0$, $A_0 = 0$, $q = \mu/L$
For $k = 0, 1, \ldots$

$$\begin{aligned}
\text{Set } A_{k+1} &= \frac{(1+q)A_k + 2\left(1 + \sqrt{(1+A_k)(1+qA_k)}\right)}{(1-q)^2} \\
\beta_k &= \frac{A_k}{(1-q)A_{k+1}}, \text{ and } \delta_k = \frac{1}{2}\frac{(1-q)^2 A_{k+1} - (1+q)A_k}{1+q+qA_k} \\
y_k &= (1-\beta_k)z_k + \beta_k x_k \\
x_{k+1} &= y_k - \frac{1}{L}\nabla f(y_k) \\
z_{k+1} &= (1-q\delta_k)z_k + q\delta_k y_k - \frac{\delta_k}{L}\nabla f(y_k).
\end{aligned} \tag{3}$$

The following theorem states the main results concerning Algorithm 1: firstly, a bound on $\|z_N - x_\star\|^2$, and secondly, a bound involving function values, which is more relevant as $\mu \to 0$. A proof for this theorem is provided in the next section.

Theorem 3 *Let $f \in \mathcal{F}_{\mu,L}$ and denote $q = \mu/L$. For any $x_0 = z_0 \in \mathbb{R}^d$ and $N \in \mathbb{N}$ with $N \geq 1$, the iterates of (3) satisfy*

$$\begin{aligned}
\|z_N - x_\star\|^2 &\leq \frac{1}{1+qA_N}\|z_0 - x_\star\|^2 \leq \tfrac{(1-\sqrt{q})^{2N}}{(1-\sqrt{q})^{2N}+q}\|z_0 - x_\star\|^2, \\
\psi_N &\leq \frac{L}{(1-q)A_{N+1}}\|z_0 - x_\star\|^2 \leq \min\left\{(1-\sqrt{q})^{2(N+1)}, \frac{1}{(N+1)^2}\right\}\frac{L}{(1-q)}\|z_0 - x_\star\|^2,
\end{aligned}$$

with $\psi_k = f(y_k) - f_\star - \frac{1}{2L}\|\nabla f(y_k)\|^2 - \frac{\mu}{2(1-\mu/L)}\|y_k - \frac{1}{L}\nabla f(y_k) - x_\star\|^2 \geq 0$.

The quantity ψ_N defined above is related to a potential (or Lyapunov) function that turns out to be key to the analysis of the method as provided in the next section. Although ψ_N might appear as slightly unnatural, it can be interpreted, in light of Theorem 1 with $x = x_\star$ and $y = y_N$, as a measure of uncertainty on the value of $f_\star$ given $x_\star$. In the special case when $\mu = 0$, ψ_N corresponds to $f(y_N) - f_\star - \frac{1}{2L}\|\nabla f(y_N)\|^2$, thus applying the standard *descent lemma*, stating that $f(x_{k+1}) \leq f(y_k) - \frac{1}{2L}\|\nabla f(y_k)\|^2$ we end up with a classical guarantee of type $f(x_{N+1}) - f_\star \leq \psi_N \leq \frac{L}{(N+1)^2}\|z_0 - x_\star\|^2$.

Note that as a result of Theorem 3, the three sequences generated by Algorithm 1 are all valid approximations of $x_\star$ in the following senses: (i) z_N converges to the optimal solution in terms of distance to the solution, (ii) y_N has a guarantee of having a small corresponding ψ_N, and (iii) x_N is obtained from a gradient step from y_{N-1}, corresponds to having a guarantee on $f(x_N) - f_\star$ when $\mu = 0$.

2.1 Worst-case analysis

For performing the analysis, we use a potential function argument (see e.g., the nice review by Bansal and Gupta [2019]) similar to those used for standard accelerated methods [Nesterov, 1983, Beck and Teboulle, 2009].

We show that for all $y_{k-1}, z_k \in \mathbb{R}^d$ and $A_k \geq 0$, the function

$$\begin{aligned}\phi_k =&(1-q)A_k\psi_{k-1} + (L+\mu A_k)\|z_k - x_\star\|^2 \\ =&(1-q)A_k\Big[f(y_{k-1}) - f_\star - \tfrac{1}{2L}\|\nabla f(y_{k-1})\|^2 - \tfrac{\mu}{2(1-\mu/L)}\|y_{k-1} - \tfrac{1}{L}\nabla f(y_{k-1}) - x_\star\|^2\Big] \\ &+ (L+\mu A_k)\|z_k - x_\star\|^2\end{aligned} \tag{4}$$

satisfies $\phi_{k+1} \leq \phi_k$ when z_{k+1}, y_k and A_{k+1} are generated according to (3).

Lemma 1 *Let $f \in \mathcal{F}_{\mu,L}$ and $k \geq 0$. For any $y_{k-1}, z_k \in \mathbb{R}^d$ and $A_k \geq 0$, two consecutive iterations of (3) satisfy*

$$\phi_{k+1} \leq \phi_k$$

with A_{k+1} being defined as in (3).

Proof We perform a weighted sum of two inequalities due to Theorem 1:

– smoothness and strong convexity of f between $x_\star$ and y_k with weight $\lambda_1 = (1-q)(A_{k+1} - A_k)$

$$f_\star \geq f(y_k) + \langle \nabla f(y_k); x_\star - y_k\rangle + \tfrac{1}{2L}\|\nabla f(y_k)\| + \tfrac{\mu}{2(1-q)}\|y_k - x_\star - \tfrac{1}{L}\nabla f(y_k)\|^2,$$

– smoothness and strong convexity between y_{k-1} and y_k with weight $\lambda_2 = (1-q)A_k$

$$\begin{aligned}f(y_{k-1}) \geq& f(y_k) + \langle \nabla f(y_k); y_{k-1} - y_k\rangle + \tfrac{1}{2L}\|\nabla f(y_k) - \nabla f(y_{k-1})\|^2 \\ &+\tfrac{\mu}{2(1-q)}\|y_k - y_{k-1} - \tfrac{1}{L}(\nabla f(y_k) - \nabla f(y_{k-1}))\|^2.\end{aligned}$$

Summing up and reorganizing those two inequalities (without substituting A_{k+1} by its definition for now), we arrive to the following inequality

$$\begin{aligned}0 \geq& \lambda_1\Big[f(y_k) - f_\star + \langle \nabla f(y_k); x_\star - y_k\rangle + \tfrac{1}{2L}\|\nabla f(y_k)\| + \tfrac{\mu}{2(1-q)}\|y_k - x_\star - \tfrac{1}{L}\nabla f(y_k)\|^2\Big] \\ &+\lambda_2\Big[f(y_k) + \langle \nabla f(y_k); y_{k-1} - y_k\rangle + \tfrac{1}{2L}\|\nabla f(y_k) - \nabla f(y_{k-1})\|^2 \\ &+ \tfrac{\mu}{2(1-q)}\|y_k - y_{k-1} - \tfrac{1}{L}(\nabla f(y_k) - \nabla f(y_{k-1}))\|^2\Big].\end{aligned}$$

Substituting

$$\begin{aligned}y_k &= (1-\beta_k)z_k + \beta_k\left(y_{k-1} - \frac{1}{L}\nabla f(y_{k-1})\right) \\ z_{k+1} &= (1-q\delta_k)z_k + q\delta_k y_k - \frac{\delta_k}{L}\nabla f(y_k),\end{aligned}$$

(note that this substitution is also valid when $k = 0$ as $\beta_k = 0$ in this case, and hence $y_0 = z_0$) the weighted sum can be reformulated exactly as (this can be verified by expanding both expressions and matching them on a term by term basis[1])

$$\begin{aligned}\phi_{k+1} \leq& \phi_k - LK_1P(A_{k+1}, A_k)\|z_k - x_\star\|^2 \\ &+ \tfrac{1}{4L}K_2P(A_{k+1}, A_k)\|(1-q)A_{k+1}\nabla f(y_k) - \mu A_k\left(y_{k-1} - x_\star - \tfrac{1}{L}\nabla f(y_{k-1})\right) + K_3\mu(z_k - x_\star)\|^2\end{aligned}$$

with three constants (well defined given that $0 \leq \mu < L < \infty$ and $A_k, A_{k+1} \geq 0$)

$$\begin{aligned}K_1 &= \frac{q^2}{(1+q)^2 + (1-q)^2qA_{k+1}} \\ K_2 &= \frac{(1+q)^2 + (1-q)^2qA_{k+1}}{(1-q)^2\left(1+q+qA_k\right)^2A_{k+1}^2} \\ K_3 &= (1+q)\frac{(1+q)A_k - (1-q)(2+qA_k)A_{k+1}}{(1+q)^2 + (1-q)^2qA_{k+1}},\end{aligned}$$

[1] The puzzled reader can verify this using basic symbolic computations. We provide a notebook for verifying the equivalence of the expressions in Section 4.

as well as

$$P(x,y) = (y-(1-q)x)^2 - 4x(1+qy).$$

For obtaining the desired potential inequality, it remains to remark that A_{k+1} corresponds to the largest solution to $P(x, A_k) = 0$, reaching the claim $\phi_{k+1} \leq \phi_k$. □

We are now armed for proving our main result, presented in Theorem 3.

Proof (Theorem 3) From Lemma 1, we get

$$\phi_N \leq \phi_{N-1} \leq \ldots \leq \phi_0 = L\|z_0 - x_\star\|^2.$$

From Theorem 1 (evaluated at $x \leftarrow x_\star$, and $y \leftarrow y_{k-1}$), we have that $(L+\mu A_N)\|z_N - x_\star\|^2 \leq \phi_N$, reaching

$$\|z_N - x_\star\|^2 \leq \frac{\phi_0}{(L+\mu A_N)} = \frac{1}{1+qA_N}\|z_0 - x_\star\|^2.$$

Similarly, we have that $(1-q)A_{N+1}\psi_N \leq \phi_{N+1} \leq \phi_N$ and hence

$$\psi_N \leq \frac{\phi_0}{(1-q)A_{N+1}} = \frac{1}{(1-q)A_{N+1}}\|z_0 - x_\star\|^2.$$

For reaching the claims, it is therefore sufficient to characterize the growth rate of $\{A_k\}$. Because finding a closed-form expression for $\{A_k\}$ appears to be out of reach, we consider the classical two scenarios for bounding its growth rate. First, when $\mu = 0$,

$$A_{k+1} = 2 + A_k + 2\sqrt{1+A_k} \geq 2 + A_k + 2\sqrt{A_k} \geq (1+\sqrt{A_k})^2,$$

reaching $\sqrt{A_{k+1}} \geq 1 + \sqrt{A_k}$ and hence $\sqrt{A_k} \geq k$ and $A_k \geq k^2$. Second, when $\mu > 0$, one also has

$$A_{k+1} = \frac{(1+q)A_k + 2\left(1+\sqrt{(1+A_k)(1+qA_k)}\right)}{(1-q)^2} \geq \frac{(1+q)A_k + 2\sqrt{qA_k^2}}{(1-q)^2} = \frac{A_k}{(1-\sqrt{q})^2}.$$

This last bound, together with $A_1 = \frac{4}{(1-q)^2} = \frac{4}{(1+\sqrt{q})^2(1-\sqrt{q})^2} \geq (1-\sqrt{q})^{-2}$, allows reaching the target $A_N \geq (1-\sqrt{q})^{-2N}$, thereby concluding the proof. □

2.2 Limit cases

In this section, we inspect two limit cases of ITEM. First, when $\mu = 0$, ITEM can be compared to the Optimized Gradient Method of Kim and Fessler [2016]. In their notations, we denote by $\theta_k^2 = \frac{A_{k+1}}{4}$, a sequence that can alternatively be defined recursively as $\theta_0 = 1$ and $\theta_{k+1} = \frac{1+\sqrt{4\theta_k^2+1}}{2}$. In this setting, the parameters correspond to $\beta_k = \frac{A_k}{A_{k+1}}$, and $\delta_k = \frac{A_{k+1}-A_k}{2}$, and we recover, using Kim and Fessler [2016]'s notations (using the identity $\theta_k^2 = \theta_{k-1}^2 + \theta_k$)

$$\begin{aligned} y_k &= \tfrac{\theta_k - 1}{\theta_k} x_k + \tfrac{1}{\theta_k} z_k \\ x_{k+1} &= y_k - \tfrac{1}{L}\nabla f(y_k) \\ z_{k+1} &= z_k - \tfrac{2}{L}\theta_k \nabla f(y_k). \end{aligned}$$

Note though that Kim and Fessler [2016] uses a "last iteration adjustment" by setting $\theta_N = \frac{1+\sqrt{8\theta_{N-1}^2+1}}{2}$. This adjustment is not needed for the purpose of obtaining the optimal bound on $\|z_N - x_\star\|$, and a detailed treatment can be found in [d'Aspremont et al., 2021, Section 4.3.1].

Second, when $\mu > 0$ and $k \to \infty$, one can explicitly compute the limits of the algorithmic parameters

$$\lim_{k\to\infty} \frac{A_k}{A_{k+1}} = \lim_{A_k\to\infty} \frac{(1-q)^2 A_k}{(1+q)A_k + 2 + 2\sqrt{1+(1+q)A_k + qA_k^2}} = \frac{(1-q)^2}{(1+\sqrt{q})^2} = (1-\sqrt{q})^2$$
$$\lim_{k\to\infty} \beta_k = \lim_{k\to\infty} \frac{A_k}{(1-q)A_{k+1}} = \frac{1-\sqrt{q}}{1+\sqrt{q}}$$
$$\lim_{k\to\infty} \delta_k = \lim_{k\to\infty} \frac{1}{2}\frac{(1-q)^2 A_{k+1} - (1+q)A_k}{1+q+qA_k} = \frac{1}{2}\frac{(1-q)^2 - (1+q)(1-\sqrt{q})^2}{q(1-\sqrt{q})^2} = \sqrt{\frac{1}{q}},$$

reaching

$$y_k = \frac{1-\sqrt{q}}{1+\sqrt{q}}\big(y_{k-1} - \tfrac{1}{L}\nabla f(y_{k-1})\big) + \left(1 - \frac{1-\sqrt{q}}{1+\sqrt{q}}\right) z_k$$
$$z_{k+1} = \sqrt{q}\big(y_k - \tfrac{1}{\mu}\nabla f(y_k)\big) + (1-\sqrt{q})z_k,$$

which is the Triple Momentum Method [Van Scoy et al., 2018] and its convergence rate $(1-\sqrt{q})^2$.

For those two limit cases, the analysis from Section 2.1 can be simplified accordingly. For the OGM, this leads to the same potential as that provided in e.g., [Taylor and Bach, 2019, Theorem 11]), or [Park et al., 2021, Section 2]. For the TMM, this allows recovering the known Lyapunov function from e.g., [Cyrus et al., 2018, Inequality (10)].

2.3 Lower bound and matching examples

In this section, we show the correspondence with the lower bound from [Drori and Taylor, 2021] and provide two very simple one-dimensional examples on which the method achieves its worst-case.

First, the lower bound from [Drori and Taylor, 2021, Corollary 4] states that for any black-box first-order, there exists $f \in \mathcal{F}_{\mu,L}$ such that

$$\|x_N - x_\star\|^2 \geq \frac{\lambda_N^2}{q}\|x_0 - x_\star\|^2,$$

where $x_\star = \operatorname{argmin}_x f(x)$, x_N is the output of the black-box first-order method under consideration, and where the sequence $\{\lambda_i\}$ is defined recursively as $\lambda_0 = \sqrt{q}$ and

$$\lambda_{k+1} = \frac{1-\sqrt{q-(1-q)\lambda_k^2}}{1+\lambda_k^2}\lambda_k.$$

Let us show that it matches the upper bound provided by Theorem 3. One can verify the identity

$$\lambda_k^2 = \frac{q}{1+qA_k},$$

by verifying that it is compatible with $A_0 = 0$ and through a recurrence argument. That is, assuming $\lambda_k = \sqrt{\frac{q}{1+qA_k}}$, it is relatively simple to establish that

$$\lambda_{k+1} = \frac{1-\sqrt{q-(1-q)\left(\frac{q}{1+qA_k}\right)}}{1+\left(\frac{q}{1+qA_k}\right)}\sqrt{\frac{q}{1+qA_k}} = \sqrt{\frac{q}{1+qA_{k+1}}}.$$

Because the lower bound from [Drori and Taylor, 2021] and the upper bound from Theorem 3 match, it is clear that the worst-case guarantee of ITEM cannot be improved.

The lower bound proof from [Drori and Taylor, 2021] is constructive in the sense that it exhibits a "worst function in the world" on which any first-order method cannot attain a worst-case guarantee better than the one stated above. Clearly, such a function would naturally attain the worst-case behavior of ITEM, however, this function is rather complex and it is the purpose of the following paragraphs to show that the worst-case

behavior of ITEM is also attained on very simple functions. In particular, the worst-case is achieved on the two base quadratic functions

$$f_L(x) = \frac{L}{2}|x|^2, \quad f_\mu = \frac{\mu}{2}|x|^2,$$

i.e., the guarantee $\|z_N - x_\star\|^2 \leq \frac{\|x_0 - x_\star\|^2}{1+qA_N}$ holds with equality on both $f_L(.)$ and $f_\mu(.)$.

Lemma 2 *Let $0 < \mu < L < \infty$, and $f_L, f_\mu \in \mathcal{F}_{\mu,L}(\mathbb{R})$ with $f_\mu(x) = \frac{\mu}{2}x^2$ and $f_L(x) = \frac{L}{2}x^2$. The iterates of ITEM* (3) *satisfy*

$$z_k^2 = \frac{z_0^2}{1 + qA_k}$$

when applied to either f_μ or f_L.

Proof We proceed by recurrence. It is clear that $z_0^2 = \frac{z_0^2}{1+qA_0}$ (recall $A_0 = 0$), which establishes the base recurrence case.

(i) Let us start with f_L. It is clear from explicit computations that for all $y_k \in \mathbb{R}$, $x_{k+1} = y_k - \frac{1}{L}\nabla f_L(y_k) = x_\star = 0$. Therefore, we have $y_k = \frac{1}{L}\nabla f_L(y_k)$ along with

$$y_k = (1 - \beta_k)z_k$$

(this also trivially holds for $k = 0$, as in this case $\beta_k = 0$ and $y_0 = z_0 = x_0$), and therefore

$$z_{k+1} = z_k + q\delta_k(y_k - z_k) - \delta_k y_k = ((1-q)\beta_k\delta_k - \delta_k + 1)z_k.$$

Substituting the expressions of β_k, δ_k, A_{k+1}, and $z_k^2 = \frac{z_0^2}{1+qA_k}$ in this equality (squared) leads to

$$z_{k+1}^2 = \frac{\left(1 + qA_k - q\sqrt{(1+A_k)(1+qA_k)}\right)^2}{(1+qA_k)(1+q+qA_k)^2} z_0^2 = \frac{z_0^2}{1+qA_{k+1}},$$

where the last equality can be verified by basic algebra.

(ii) We proceed with f_μ. In this case, for all $y_k \in \mathbb{R}$ we have $y_k - \frac{1}{\mu}\nabla f_\mu(y_k) = x_\star = 0$. Therefore,

$$z_{k+1} = (1 - q\delta_k)z_k.$$

Substituting the expression of δ_k and the recurrence hypothesis $z_k^2 = \frac{z_0^2}{1+qA_k}$ we arrive to the same expression as before

$$z_{k+1}^2 = \frac{\left(1 + qA_k - q\sqrt{(1+A_k)(1+qA_k)}\right)^2}{(1+qA_k)(1+q+qA_k)^2} z_0^2 = \frac{z_0^2}{1+qA_{k+1}},$$

reaching the desired claim. □

3 A constructive approach to ITEM

The intent of this section is to provide a constructive procedure for obtaining the Information-Theoretic Exact Method, as well as other similar methods designed based on alternate optimality criteria. We would like to emphasize that although ITEM was discovered using the technique described below, its proof, as provided above, is independent of the following.

As a starting point, consider the class of black-box first-order methods gathering information about the objective function f only by evaluating an *oracle* $\mathcal{O}_f(x) = (f(x), \nabla f(x))$. We describe such a black-box method M as a set of rules $\{M_1, M_2, \ldots, M_N\}$ for forming its iterates, which we denote by w_k for avoiding confusions with any of the sequences defined by ITEM, as

$$\begin{aligned}
w_1 &= M_1(x_0, \mathcal{O}_f(w_0)) \\
w_2 &= M_2(x_0, \mathcal{O}_f(w_0), \mathcal{O}_f(w_1)) \\
&\vdots \\
w_N &= M_N(x_0, \mathcal{O}_f(w_0), \mathcal{O}_f(w_1), \ldots, \mathcal{O}_f(w_{N-1})),
\end{aligned}$$

and we denote by $\mathcal{M}_N$ the set of black-box first-order methods that perform N gradient evaluations. Furthermore, we call the *efficiency estimate* of a method M the following quantity

$$W_{\mu,L}(M) = \sup_{f\in\mathcal{F}_{\mu,L}} \left\{ \frac{\|w_N - w_\star\|^2}{\|w_0 - w_\star\|^2} : \text{for any sequence } w_1,\dots,w_N \text{ generated by } M \text{ on } f, \right. \\ \left. \text{initiated at some } w_0, \text{ and } w_\star \in \operatorname{argmin}_w f(w) \right\}, \tag{5}$$

which correspond to the worst-case performance of M on the class $\mathcal{F}_{\mu,L}$ for the criterion $\frac{\|w_N-w_\star\|^2}{\|w_0-w_\star\|^2}$. A direct consequence of Theorem 3 and the lower complexity bound discussed in Section 2.3 is that ITEM belong to the class of black-box first-order methods with optimal performances with respect to $W_{\mu,L}(M)$ with $M \in \mathcal{M}_N$. ITEM is therefore a solution to

$$\min_{M\in\mathcal{M}_N} W_{\mu,L}(M). \tag{6}$$

Although this minimax problem appears to be hard to solve directly, we illustrate below that it can be approached using semidefinite programming.

In a nutshell, we consider two simplified upper bounds to this minimax problem. First, we consider a subclass of black-box first-order methods, referred to as *fixed-step first-order methods*. Those are first-order methods that are described by a set of fixed coefficients $\{h_{i,j}\}$, and whose formal description is provided below. Second, given a fixed-step first-order method M, the idea is to develop a tractable upper bound on the efficiency estimate of M, written $\mathrm{UB}_{\mu,L}(M)$ and such that $\mathrm{UB}_{\mu,L}(M) \geq W_{\mu,L}(M)$. After that, we show that minimizing this upper bound over M is also tractable. That is, we can solve $\min_{\{h_{i,j}\}} \mathrm{UB}_{\mu,L}(M)$ to obtain the Information-Theoretic Exact Method as a solution.

As a comparison, let us mention that the Optimized Gradient Method [Drori and Teboulle, 2014, Kim and Fessler, 2016] was obtained through similar steps for the objective $(f(w_N) - f_\star)/\|w_0 - w_\star\|^2$ when $\mu = 0$. Note, however, that a straightforward application of the technique presented in [Drori and Teboulle, 2014, Kim and Fessler, 2016] does not yield tractable problems in the strongly convex case.

More precisely, we proceed as follows:

- In Section 3.1, we describe the class of fixed-step first-order methods. This class of methods is somewhat natural and contains classical numerical methods such as gradient, heavy-ball, and accelerated gradient methods, but excludes adaptive methods. For this class of methods, it is known that $W_{\mu,L}(M)$ can be formulated as a convex semidefinite program (see e.g., [Taylor et al., 2017b, Theorem 6]). However, when it comes to optimizing over step size parameters, this formulation leads to a bilinear/quadratic problem which we do not know how to solve directly.
- In Section 3.3 and 3.4, we provide an equivalent reparametrization of the class of fixed-step first-order methods, allowing to reach an alternate semidefinite formulation for $W_{\mu,L}(M)$ with simpler structure. We further detail a tractable upper bound $\mathrm{UB}_{\mu,L}(M)$ which is more convenient for optimizing over the method's parameters.
- In Section 3.5, we show how to render $\min_{\{h_{i,j}\}} \mathrm{UB}_{\mu,L}(M)$ tractable, yielding the Information-Theoretic Exact Method as a solution.

We complement those developments with numerical examples of the design procedure, as well as applications to alternate design criterion that include $(f(w_N) - f_\star)/\|w_N - w_\star\|^2$. For doing that, the developments of this section have to be slightly adapted (see Appendix D). Corresponding numerical examples are provided in Appendix E, and source code for reproducing the results is provided in Section 4.

3.1 Fixed-step first-order methods

In this section, we introduce a subclass of black-box first-order methods described by a set of fixed coefficients. This parametric subset of $\mathcal{M}_N$ allows for more convenient formulations of optimization problems over the class of methods, such as the minimax problem (6).

We start with the following "natural" description of the class of methods of interest, then introduce an alternate parametrization which is more convenient for the step size optimization procedure of the following sections.

Definition 2 A black box first-order method is called a *fixed-step first-order method* if there exists a set $\{h_{i,j}\} \subset \mathbb{R}$ such that the method admits the following description

$$\begin{aligned} w_1 &= w_0 - \tfrac{h_{1,0}}{L}\nabla f(w_0) \\ w_2 &= w_1 - \tfrac{h_{2,0}}{L}\nabla f(w_0) - \tfrac{h_{2,1}}{L}\nabla f(w_1) \\ &\vdots \\ w_N &= w_{N-1} - \sum_{i=0}^{N-1} \tfrac{h_{N,i}}{L}\nabla f(w_i), \end{aligned} \tag{7}$$

for any function f.

For fixed-step first-order methods M described by a set of normalized coefficients $\{h_{i,j}\}$, it is shown in [Taylor et al., 2017b, Theorem 6] that $W_{\mu,L}(M)$ can be formulated as a convex semidefinite program (SDP). Given such an SDP formulation, our goal is to solve

$$\min_{\{h_{i,j}\}} W_{\mu,L}(M),$$

is a bilinear/quadratic problem, due to the structure of the SDP formulation of $W_{\mu,L}(M)$ in [Taylor et al., 2017b, Theorem 6]. Such problems are nonconvex and NP-hard in general [Toker and Ozbay, 1995], nevertheless, by performing reparametrization, followed by relaxation and linearization steps, as shown in the following sections, it is possible to attain a tractable relaxation of the problem.

3.2 A reparametrization of fixed-step first-order methods

In what follows, we restrict ourselves to these fixed-step first-order methods, which we will reparameterize in a slightly different, but equivalent, fashion. Informally, the alternate parameterization allows formulating the maximization problem arising in the efficiency estimate $W_{\mu,L}(M)$ (see (5)) in a more convenient way than that of [Taylor et al., 2017b, Theorem 6] for our purposes. Indeed, the new formulation presented in the next sections allows obtaining a problem that is "only" bilinear in terms of the method parameters and of some multipliers $\lambda_{i,j}$'s. Those problems are still NP-hard in general [Toker and Ozbay, 1995], however, in this case this simplification will enable us to optimize over the method parameters, a simplification that appears to be hard to reach with previous formulations.

In order to proceed, we express first-order methods for minimizing f as acting instead on a function $\tilde{f}$, using

$$\tilde{f}(x) := f(x) - \tfrac{\mu}{2}\|x - w_\star\|^2,$$

where $w_\star$ is a minimizer of both f and $\tilde{f}$. It is known (see e.g. [Nesterov, 2004]) that $f \in \mathcal{F}_{\mu,L}$ if and only if $\tilde{f} \in \mathcal{F}_{0,L-\mu}$. Then, one can express (7) in terms of evaluations of the gradient of $\tilde{f}$, instead of that of f. Concretely, we reformulate (7) in terms of some coefficients $\{\alpha_{i,j}\}$ as follows

$$\begin{aligned} w_1 - w_\star &= (w_0 - w_\star)(1 - \tfrac{\mu}{L}\alpha_{1,0}) - \tfrac{\alpha_{1,0}}{L}\nabla \tilde{f}(w_0) \\ w_2 - w_\star &= (w_0 - w_\star)(1 - \tfrac{\mu}{L}(\alpha_{2,0} + \alpha_{2,1})) - \tfrac{\alpha_{2,0}}{L}\nabla \tilde{f}(w_0) - \tfrac{\alpha_{2,1}}{L}\nabla \tilde{f}(w_1) \\ &\vdots \\ w_N - w_\star &= (w_0 - w_\star)\left(1 - \tfrac{\mu}{L}\sum_{i=0}^{N-1}\alpha_{N,i}\right) - \sum_{i=0}^{N-1}\tfrac{\alpha_{N,i}}{L}\nabla \tilde{f}(w_i). \end{aligned} \tag{8}$$

One can show that there is a bijection between representation (7) and (8). Therefore, the problem of designing an optimal method in the form (7) is equivalent to that of devising an optimal method in the form (8). This is formalized by the following lemma.

Lemma 3 *Let $N \in \mathbb{N}$ and a first-order method $M \in \mathcal{M}_N$. The following statements are equivalent.*

- *There exists a set* $\{h_{i,j}\}_{i=1,\dots,N;j=0,\dots,i-1}$ *such that for any* $f \in \mathcal{F}_{\mu,L}$ *and* $w_0 \in \mathbb{R}^d$ *the sequence* $\{w_k\}_{k=0,\dots,N} \subset \mathbb{R}^d$ *generated by* M *satisfies* (7) *(i.e.,* M *is a fixed-step first-order method).*
- *There exists a set* $\{\alpha_{i,j}\}_{i=1,\dots,N;j=0,\dots,i-1}$ *such that for any* $f \in \mathcal{F}_{\mu,L}$ *and* $w_0 \in \mathbb{R}^d$ *the sequence* $\{w_k\}_{k=0,\dots,N} \subset \mathbb{R}^d$ *generated by* M *satisfies* (8)*.*

Proof The proof follows from a short recurrence argument (provided in Appendix A) for showing that the two representation are isomorphic, and that they are linked through the following triangular system of equations

$$\alpha_{k+1,i} = \begin{cases} h_{k+1,k} & \text{if } i = k \\ h_{k+1,i} + \alpha_{k,i} - \frac{\mu}{L}\sum_{j=i+1}^{k} h_{k+1,j}\alpha_{j,i} & \text{if } 0 \leq i < k. \end{cases} \tag{9}$$

Therefore, although we use (8) in the following sections, any method formulated in terms of $\{\alpha_{i,j}\}$ can be converted to the more natural $\{h_{i,j}\}$ notation, and reciprocally. □

In the next section, we develop an upper bound on $W_{\mu,L}(M)$ of a form similar to that of [Taylor et al., 2017b, Theorem 6], but which is linear in $\{\alpha_{i,j}\}$, instead of quadratic in $\{h_{i,j}\}$.

3.3 A performance estimation problem and its relaxation

The goal of this section is to construct an upper bound on (5) that can be computed efficiently. The reformulation and relaxation techniques used for obtaining the upper bound are not new and rely on the same steps as those taken in [Taylor et al., 2017b] (so readers familiar with such procedures can safely fly over the section). We provide details which allows optimizing over the step sizes afterwards. Let us start by rephrasing (5) as

$$\begin{aligned} W_{\mu,L}(M) = \sup_{\substack{\tilde{f},d\in\mathbb{N} \\ \{w_i\}_{i\in I}\subset\mathbb{R}^d}} \quad & \frac{\|w_N - w_\star\|^2}{\|w_0 - w_\star\|^2} \\ \text{s.t. } & w_k \text{ generated by (8) applied on } \tilde{f}, \text{ and initiated at some } w_0 \\ & \tilde{f} \in \mathcal{F}_{0,L-\mu}(\mathbb{R}^d) \\ & w_\star \in \operatorname{argmin}_w \tilde{f}(w), \end{aligned}$$

where we used an index set $I = \{\star, 0, \dots, N\}$. Note the maximization over d, which aims at obtaining dimension-independent guarantees. For such problems, it is known that the supremum is attained (see e.g., [Taylor et al., 2017b, Proposition 1]), and we therefore use "max" instead of "sup" in what follows.

As a first step towards an "efficient" upper bound, we reformulate (5) using an *extension* (or interpolation) argument. That is, the previous maximization problem can be restated using an existence argument for replacing the function by a finite set of samples. In other words, we optimize over the oracle's responses while keeping the responses consistent with assumptions on f

$$\begin{aligned} \max_{\substack{\{(w_i,g_i,f_i)\}_{i\in I}\subset\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R} \\ d\in\mathbb{N}}} \quad & \frac{\|w_N - w_\star\|^2}{\|w_0 - w_\star\|^2} \\ \text{s.t. } & w_k \text{ generated by (8) for } k = 1, \dots, N \\ & \exists \tilde{f} \in \mathcal{F}_{0,L-\mu}(\mathbb{R}^d) : g_i \in \partial\tilde{f}(w_i),\ f_i = \tilde{f}(w_i)\, \forall i \in I, \\ & g_\star = 0. \end{aligned}$$

Using an homogeneity argument, one can reformulate this problem without the fractional objective. More precisely, for any feasible point $S = \{(x_i, g_i, f_i)\}_{i\in I}$ and any $\alpha > 0$, the point $S' = \{(\alpha x_i, \alpha g_i, \alpha^2 f_i)\}_{i\in I}$ is also feasible while reaching the same objective value (this can be verified using the definition of $\mathcal{F}_{\mu,L}$). We

can therefore arbitrarily fix the scale of the problem to $\|w_0 - w_\star\| = 1$, reaching the following problem with the same optimal value

$$\begin{aligned}\max_{\substack{\{(w_i,g_i,f_i)\}_{i\in I}\subset\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}\\ d\in\mathbb{N}}} & \|w_N - w_\star\|^2 \\ \text{s.t. } & \|w_0 - w_\star\|^2 = 1, \\ & w_k \text{ generated by (8) for } k = 1,\dots,N \\ & \exists \tilde{f} \in \mathcal{F}_{0,L-\mu}(\mathbb{R}^d) : g_i \in \partial\tilde{f}(w_i),\ f_i = \tilde{f}(w_i)\ \forall i \in I, \\ & g_\star = 0.\end{aligned}$$

It follows from Theorem 2 that the previous problem can be reformulated exactly as

$$\begin{aligned}\max_{\substack{\{(w_i,g_i,f_i)\}_{i\in I}\subset\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}\\ d\in\mathbb{N}}} & \|w_N - w_\star\|^2 \\ \text{s.t. } & \|w_0 - w_\star\|^2 = 1, \\ & w_k \text{ generated by (8) for } k = 1,\dots,N \\ & f_i \geq f_j + \langle g_j; x_i - x_j\rangle + \tfrac{1}{2(L-\mu)}\|g_i - g_j\|^2\ \ \forall i,j \in I, \\ & g_\star = 0.\end{aligned} \tag{10}$$

Whereas equivalence between the two previous problems might be regarded as technical, the fact (10) produces upper bounds on (5) is quite direct. Indeed, any $\tilde{f} \in \mathcal{F}_{0,L-\mu}$ satisfies the above inequalities, and hence any feasible point to (5) can be converted to a feasible point to (10) by sampling $\tilde{f}$.

In what follows, we use the following relaxation of (10), by incorporating only a specific subset of the previous quadratic inequalities, therefore forming an upper bound on the original problem. Many inequalities were removed because they introduce undesirable nonlinearities in the steps taken in the next sections. Perhaps luckily, this relaxation will turn out to be tight for evaluating $W_{\mu,L}(M)$ of ITEM.

$$\begin{aligned}\max_{\substack{\{(w_i,g_i,f_i)\}_{i\in I}\subset\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}\\ d\in\mathbb{N}}} & \|w_N - w_\star\|^2 \\ \text{s.t. } & \|w_0 - w_\star\|^2 = 1,\ g_\star = 0 && \\ & w_k \text{ generated by (8)} && \text{for } k = 1,\dots,N \\ & f_i \geq f_{i+1} + \langle g_{i+1}; w_i - w_{i+1}\rangle + \tfrac{1}{2(L-\mu)}\|g_i - g_{i+1}\|^2 && \text{for } i = 0,\dots,N-2 \\ & f_\star \geq f_i + \langle g_i, w_\star - w_i\rangle + \tfrac{1}{2(L-\mu)}\|g_i\|^2 && \text{for } i = 0,\dots,N-1 \\ & f_{N-1} \geq f_\star + \tfrac{1}{2(L-\mu)}\|g_{N-1}\|^2. &&\end{aligned} \tag{R}$$

As shown in the next section, this problem is semidefinite-representable and we can thus use standard packages for approximating its solution numerically. Looking at the structure of these numerical solutions helped us to choose this particular relaxation.

The following lemma summarizes what we have obtained so far, that is, $W_{\mu,L}(M) \leq \mathrm{val(R)}$, where val(R) denotes the optimal value of (R).

Lemma 4 *Let $N \in \mathbb{N}$, $0 \leq \mu < L < \infty$, and $M \in \mathcal{M}_N$ be a fixed-step first-order method* (8) *performing N gradient evaluations and described by a set of coefficients $\{\alpha_{i,j}\}_{i,j}$. For any $d \in \mathbb{N}$, $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$, $w_\star \in \operatorname{argmin}_w f(w)$, initial guess $w_0 \in \mathbb{R}^d$, and $w_N = M(w_0, f)$, it holds that*

$$\|w_N - w_\star\|^2 \leq \mathrm{val(R)}\,\|w_0 - w_\star\|^2,$$

where val(R) *denotes the optimal value of* (R).

3.4 Tractable upper bounds using semidefinite programming (SDP)

We now show how to reach a standard SDP formulation for (R). One can reformulate the maximization problem (R) in terms of the variables (G, F) (after substituting w_k's by their expressions) defined by

$$G = \begin{pmatrix} \|w_0 - w_\star\|^2 & \langle g_0; w_0 - w_\star\rangle & \langle g_1; w_0 - w_\star\rangle & \dots & \langle g_{N-1}; w_0 - w_\star\rangle \\ \langle g_0; w_0 - w_\star\rangle & \|g_0\|^2 & \langle g_1; g_0\rangle & \dots & \langle g_{N-1}; g_0\rangle \\ \langle g_1; w_0 - w_\star\rangle & \langle g_1; g_0\rangle & \|g_1\|^2 & \dots & \langle g_{N-1}; g_1\rangle \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \langle g_{N-1}; w_0 - w_\star\rangle & \langle g_{N-1}; g_0\rangle & \langle g_{N-1}; g_1\rangle & \dots & \|g_{N-1}\|^2 \end{pmatrix} \succeq 0, \tag{11}$$
$$F = \begin{pmatrix} f_0 - f_\star \\ f_1 - f_\star \\ \vdots \\ f_{N-1} - f_\star \end{pmatrix},$$

(this representation is equivalent to that in terms of triplets $\{(x_i, g_i, f_i)\}$, as one can construct such a pair (G, F) from a set of such triplets, and vice-versa via a Cholesky factorization of G). Formally, let us introduce the following notations for picking elements in G and F and conveniently formulating the SDPs

$$\mathbf{w}_0 = e_1 \in \mathbb{R}^{N+1}, \quad \mathbf{g}_i = e_{i+2} \in \mathbb{R}^{N+1}, \quad \mathbf{f}_i = e_{i+1} \in \mathbb{R}^N,$$

with $i = 0, \dots, N-1$ and e_i being the unit vector whose ith component is equal to 1. In addition, we also denote by

$$\mathbf{w}_k = \mathbf{w}_0 \left(1 - \tfrac{\mu}{L} \sum_{i=0}^{k-1} \alpha_{k,i}\right) - \sum_{i=0}^{k-1} \tfrac{\alpha_{k,i}}{L} \mathbf{g}_i,$$

for $i = 0, \dots, N$ (note that $\mathbf{w}_k$ is therefore linearly parameterized by $\{\alpha_{k,i}\}$). Those notations allow to express the objective and constraints of (R) directly in terms of G and F using the following identities

$$\begin{aligned} &f_i - f_\star = F\mathbf{f}_i && i = 0, 1, \dots, N-1 \\ &\|g_i\|^2 = \mathbf{g}_i^\top G \mathbf{g}_i && i = 0, 1, \dots, N-1 \\ &\|w_i - w_\star\|^2 = \mathbf{w}_i^\top G \mathbf{w}_i && i = 0, 1, \dots, N \\ &\langle g_i; w_j - w_\star\rangle = \mathbf{g}_i^\top G \mathbf{w}_j && i = 0, 1, \dots, N-1,\ j = 0, 1, \dots, N. \end{aligned}$$

Using those notations, any feasible point to (R) can be transformed to a feasible point to the following (SDP-R), using the Gram matrix representation. Hence, the optimal value to the following problem is an upper bound on that of (R)

$$\begin{aligned} \max_{\substack{G \in \mathbb{S}^{N+1} \\ F \in \mathbb{R}^N \\ d \in \mathbb{N}}} \ & \mathbf{w}_N^\top G \mathbf{w}_N \\ \text{s.t. } & G \succeq 0 \qquad \text{(SDP-R)} \\ & \mathbf{w}_0^\top G \mathbf{w}_0 = 1, \\ & 0 \geq (\mathbf{f}_{i+1} - \mathbf{f}_i)^\top F + \mathbf{g}_{i+1}^\top G(\mathbf{w}_i - \mathbf{w}_{i+1}) + \tfrac{1}{2(L-\mu)} (\mathbf{g}_i - \mathbf{g}_{i+1})^\top G (\mathbf{g}_i - \mathbf{g}_{i+1}) \quad \text{for } i = 0, \dots, N-2 \\ & 0 \geq \mathbf{f}_i^\top F - \mathbf{g}_i^\top G \mathbf{w}_i + \tfrac{1}{2(L-\mu)} \mathbf{g}_i^\top G \mathbf{g}_i \quad \text{for } i = 0, \dots, N-1 \\ & 0 \geq -\mathbf{f}_{N-1}^\top F + \tfrac{1}{2(L-\mu)} \mathbf{g}_{N-1}^\top G \mathbf{g}_{N-1} \\ & \operatorname{rank}(G) \leq d. \end{aligned}$$

After getting rid of the variable d and the rank constraint (which is void due to maximization over d), this problem is a linear SDP, parametrized by $L > \mu \geq 0$, and $\{\alpha_{i,j}\}$.

For transforming the minimax problem to a bilinear minimization problem, the next key step in our procedure is to express the Lagrangian dual of (SDP-R), substituting the inner maximization problem by a minimization, hence replacing the minimax by a minimization problem. Note that we do not assume strong duality, as weak duality suffices for obtaining an upper bound on the original problem. That is, we perform the following primal-dual associations

$$\begin{aligned}
&\|w_0 - w_\star\|^2 = 1 && && : \tau\\
&f_i \geq f_{i+1} + \langle g_{i+1}; w_i - w_{i+1}\rangle + \tfrac{1}{2(L-\mu)}\|g_i - g_{i+1}\|^2 && \text{for } i = 0,\dots,N-2 && : \lambda_{i,i+1}\\
&f_\star \geq f_i + \langle g_i, w_\star - w_i\rangle + \tfrac{1}{2(L-\mu)}\|g_i\|^2 && \text{for } i = 0,\dots,N-1 && : \lambda_{\star,i}\\
&f_{N-1} \geq f_\star + \tfrac{1}{2(L-\mu)}\|g_{N-1}\|^2 && && : \lambda_{N-1,\star}
\end{aligned}$$

and arrive to the following dual formulation of (SDP-R), whose optimal value is denoted by $\mathrm{UB}_{\mu,L}(M)$

$$\begin{aligned}
\mathrm{UB}_{\mu,L}(\{\alpha_{i,j}\}) := \min_{\tau,\lambda_{i,j}\geq 0} \ & \tau,\\
\text{s.t. } & S(\tau,\{\lambda_{i,j}\},\{\alpha_{i,j}\}) \succeq 0, \qquad \text{(dual-SDP-R)}\\
& \sum_{i=0}^{N-2} \lambda_{i,i+1}(\mathbf{f}_{i+1} - \mathbf{f}_i) + \sum_{i=0}^{N-1} \lambda_{\star,i}\mathbf{f}_i - \lambda_{N-1,\star}\mathbf{f}_{N-1} = 0,
\end{aligned}$$

with (note the dependence on $\{\alpha_{i,j}\}$ via $\mathbf{w}_i$'s)

$$\begin{aligned}
S(\tau,\{\lambda_{i,j}\},\{\alpha_{i,j}\}) =& \tau\,\mathbf{w}_0\mathbf{w}_0^\top - \mathbf{w}_N\mathbf{w}_N^\top + \frac{\lambda_{N-1,\star}}{2(L-\mu)}\mathbf{g}_{N-1}\mathbf{g}_{N-1}^\top + \sum_{i=0}^{N-1}\frac{\lambda_{\star,i}}{2}\left(-\mathbf{g}_i\mathbf{w}_i^\top - \mathbf{w}_i\mathbf{g}_i^\top + \tfrac{1}{L-\mu}\mathbf{g}_i\mathbf{g}_i^\top\right)\\
&+ \sum_{i=0}^{N-2}\frac{\lambda_{i,i+1}}{2}\left(\mathbf{g}_{i+1}(\mathbf{w}_i - \mathbf{w}_{i+1})^\top + (\mathbf{w}_i - \mathbf{w}_{i+1})\mathbf{g}_{i+1}^\top + \tfrac{1}{L-\mu}(\mathbf{g}_i - \mathbf{g}_{i+1})(\mathbf{g}_i - \mathbf{g}_{i+1})^\top\right).
\end{aligned}$$

Note that in the case $N > 1$, the equality constraint in (dual-SDP-R) can be written as

$$\begin{aligned}
&\lambda_{\star,0} - \lambda_{0,1} = 0\\
&\lambda_{i-1,i} + \lambda_{\star,i} - \lambda_{i,i+1} = 0 \qquad \text{for } i = 1,\dots,N-2\\
&\lambda_{N-2,N-1} + \lambda_{\star,N-1} - \lambda_{N-1,\star} = 0.
\end{aligned} \tag{12}$$

When $N = 1$, it reduces to $\lambda_{\star,0} - \lambda_{0,\star} = 0$. The following lemma recaps the current situation.

Lemma 5 *Let $N \in \mathbb{N}$, $0 \leq \mu < L < \infty$, and $M \in \mathcal{M}_N$ be a black-box first-order method* (8) *performing N gradient evaluations and described by a set of coefficients $\{\alpha_{i,j}\}_{i,j}$. For any $d \in \mathbb{N}$, $w_0 \in \mathbb{R}^d$, $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$, $w_\star \in \operatorname{argmin}_w f(w)$, and $w_N = M(w_0, f)$, it holds that*

$$\|w_N - w_\star\|^2 \leq \mathrm{UB}_{\mu,L}(\{\alpha_{i,j}\})\,\|w_0 - w_\star\|^2.$$

Proof The result follows Lemma 4. More precisely, any feasible point to (R) can be translated to a feasible point to (SDP-R) using the Gram matrix representation (11), hence val(R) $\leq$ val(SDP-R). Furthermore, weak duality implies val(SDP-R) $\leq$ val(dual-SDP-R) $= \mathrm{UB}_{\mu,L}(\{\alpha_{i,j}\})$. Therefore

$$\mathrm{val(R)} \leq \mathrm{val(SDP\text{-}R)} \leq \mathrm{UB}_{\mu,L}(\{\alpha_{i,j}\}),$$

and it follows that

$$\|w_N - w_\star\|^2 \leq \mathrm{val(R)}\|w_0 - w_\star\|^2 \leq \mathrm{UB}_{\mu,L}(\{\alpha_{i,j}\})\,\|w_0 - w_\star\|^2,$$

where the first inequality is due to Lemma 4. □

The last remaining difficulty is that $S(.)$ appearing in (dual-SDP-R) is bilinear in terms of the algorithmic parameters $\{\alpha_{i,j}\}$ (the vectors $\mathbf{w}_i$ depend linearly on those parameters) and the dual variables $\{\lambda_{i,j}\}$. Therefore, it might be unclear how to efficiently solve

$$\begin{aligned}\min_{\{\alpha_{i,j}\}} \mathrm{UB}_{\mu,L}(\{\alpha_{i,j}\}) \equiv \min_{\{\alpha_{i,j}\}} \min_{\tau,\{\lambda_{i,j}\}\geq 0} & \tau, \\ \text{s.t. } & S(\tau,\{\lambda_{i,j}\},\{\alpha_{i,j}\}) \succeq 0, \\ & \sum_{i=0}^{N-2} \lambda_{i,i+1}(\mathbf{f}_{i+1}-\mathbf{f}_i) + \sum_{i=0}^{N-1} \lambda_{\star,i}\mathbf{f}_i - \lambda_{N-1,\star}\mathbf{f}_{N-1} = 0,\end{aligned}$$

as problems involving such bilinear matrix inequalities are NP-hard in general. In the next section, we introduce a *linearization* trick which allows tackling this specific problem.

3.5 An approximate minimax and its semidefinite representation

In this section, we proceed with the last stage of our construction, by showing how to solve

$$\min_{\{\alpha_{i,j}\}} \mathrm{UB}_{\mu,L}(\{\alpha_{i,j}\}), \tag{13}$$

which is a minimization problem jointly on τ, $\alpha_{i,j}$'s and $\lambda_{i,j}$'s. As it is, the problem features a *bilinear matrix inequality*. A few algebraic manipulations on the matrix S allows rewriting the bilinear matrix inequality in terms of a matrix S', in a slightly more explicit and convenient way (those manipulations are provided in Appendix B, where S' is defined). This structure reveals that the following change of variables allows linearizing the bilinear matrix inequality

$$\beta_{i,j} = \begin{cases} \lambda_{i,i+1}\alpha_{i,j} & \text{if } 0 \leq i < N-1 \\ \lambda_{N-1,\star}\alpha_{N-1,j} & \text{if } i = N-1 \\ \alpha_{N,j} & \text{if } i = N. \end{cases} \tag{14}$$

As provided in Lemma 6 below, this change of variables is invertible for the problem under consideration. In other words, for any $N > 1$ and $0 \leq \mu < L$, one can solve (13) via its reformulation (intermediate computations, involving a matrix S' are provided in Appendix B; the important thing to see about those formulation is how variables $\{\alpha_{i,j}\}$ and $\{\lambda_{i,j}\}$ interact with each others), as

$$\begin{aligned}\min_{\substack{\tau,\{\lambda_{i,j}\}\geq 0 \\ \{\beta_{i,j}\}}} & \tau \\ \text{s.t. } & \begin{pmatrix} S''(\tau,\{\lambda_{i,j}\},\{\beta_{i,j}\}) & \mathbf{w}_N \\ \mathbf{w}_N^\top & 1 \end{pmatrix} \succeq 0, \\ & \sum_{i=0}^{N-2} \lambda_{i,i+1}(\mathbf{f}_{i+1}-\mathbf{f}_i) + \sum_{i=0}^{N-1} \lambda_{\star,i}\mathbf{f}_i - \lambda_{N-1,\star}\mathbf{f}_{N-1} = 0,\end{aligned} \tag{Minimax-R}$$

which is a standard linear semidefinite program, with the following definitions (note the linear dependencies in all parameters $\tau, \{\lambda_{i,j}\}, \{\beta_{i,j}\}$)

$$
\begin{aligned}
&S''(\tau,\{\lambda_{i,j}\},\{\beta_{i,j}\}) \\
&\quad= \tfrac{1}{2(L-\mu)}\left(\lambda_{N-1,\star}\,\mathbf{g}_{N-1}\mathbf{g}_{N-1}^\top + \sum_{i=0}^{N-1}\lambda_{\star,i}\mathbf{g}_i\mathbf{g}_i^\top + \sum_{i=0}^{N-2}\lambda_{i,i+1}(\mathbf{g}_i-\mathbf{g}_{i+1})(\mathbf{g}_i-\mathbf{g}_{i+1})^\top\right) \\
&\qquad - \tfrac{1}{2}\lambda_{\star,0}(\mathbf{g}_0\mathbf{w}_0^\top + \mathbf{w}_0\mathbf{g}_0^\top) - \sum_{i=1}^{N-2}\left(\lambda_{i,i+1} - \tfrac{\mu}{L}\sum_{j=0}^{i-1}\beta_{i,j}\right)\tfrac{1}{2}(\mathbf{g}_i\mathbf{w}_0^\top + \mathbf{w}_0\mathbf{g}_i^\top) \\
&\qquad + \sum_{i=1}^{N-2}\sum_{j=0}^{i-1}\tfrac{\beta_{i,j}}{L}\tfrac{1}{2}(\mathbf{g}_i\mathbf{g}_j^\top + \mathbf{g}_j\mathbf{g}_i^\top) - \left(\lambda_{N-1,\star} - \tfrac{\mu}{L}\sum_{j=0}^{N-2}\beta_{N-1,j}\right)\tfrac{1}{2}(\mathbf{g}_{N-1}\mathbf{w}_0^\top + \mathbf{w}_0\mathbf{g}_{N-1}^\top) \\
&\qquad + \sum_{j=0}^{N-2}\tfrac{\beta_{N-1,j}}{L}\tfrac{1}{2}(\mathbf{g}_{N-1}\mathbf{g}_j^\top + \mathbf{g}_j\mathbf{g}_{N-1}^\top) + \sum_{i=0}^{N-2}\left(\lambda_{i,i+1} - \tfrac{\mu}{L}\sum_{j=0}^{i-1}\beta_{i,j}\right)\tfrac{1}{2}(\mathbf{g}_{i+1}\mathbf{w}_0^\top + \mathbf{w}_0\mathbf{g}_{i+1}^\top) \\
&\qquad - \sum_{i=0}^{N-2}\sum_{j=0}^{i-1}\tfrac{\beta_{i,j}}{L}\tfrac{1}{2}(\mathbf{g}_{i+1}\mathbf{g}_j^\top + \mathbf{g}_j\mathbf{g}_{i+1}^\top) + \tau\mathbf{w}_0\mathbf{w}_0^\top,
\end{aligned}
$$

and $\mathbf{w}_N = \mathbf{w}_0\left(1-\frac{\mu}{L}\sum_{i=0}^{N-1}\alpha_{N,i}\right) - \sum_{i=0}^{N-1}\frac{\alpha_{N,i}}{L}\mathbf{g}_i$. From the solution to (Minimax-R), one can recover a fixed-step first-order method whose worst-case performance satisfies

$$\|x_N - x_\star\|^2 \leq \text{val(Minimax-R)}\,\|x_0 - x_\star\|^2,$$

as formalized by the next lemma.

Lemma 6 *Let $N \in \mathbb{N}$ with $N > 1$, and $0 \leq \mu < L < \infty$. Furthermore, let $(\tau, \{\beta_{i,j}\}, \{\lambda_{i,j}\})$ be a solution to* (Minimax-R)*. The following statements hold.*

(i) If $\lambda_{i,i+1} = 0$, then $\beta_{i,j} = 0$.
(ii) If $\lambda_{N-1,\star} = 0$, then $\beta_{N-1,j} = 0$.
(iii) Let $\{\alpha_{i,j}\}$ be defined as

$$\alpha_{i,j} = \begin{cases} 0 & \text{if } \beta_{i,j} = 0, \\ \beta_{i,j}/\lambda_{i,i+1} & \text{if } 0 \leq i < N-1, \\ \beta_{N-1,j}/\lambda_{N-1,\star} & \text{if } i = N-1, \\ \beta_{N,j} & \text{if } i = N. \end{cases} \tag{15}$$

The output of the corresponding method of the form (8) *satisfies*

$$\|w_N - w_\star\|^2 \leq \tau\|w_0 - w_\star\|^2$$

for any $d \in \mathbb{N}$, $w_0 \in \mathbb{R}^d$, and $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$.

Proof (i) assume $\lambda_{k,k+1} = 0$ for some $k > 0$; it follows from $\lambda_{i,j} \geq 0$ and (12) that $\lambda_{\star,i} = 0$ and $\lambda_{i-1,i} = 0$ for all $1 \leq i \leq k$. From the expression of S'', it means that there are no diagonal entries corresponding to the entries $e_2e_2^\top$, ..., $e_{k+2}e_{k+2}^\top$ (corresponding to $\mathbf{g}_0\mathbf{g}_0^\top$, ..., $\mathbf{g}_k\mathbf{g}_k^\top$). Therefore, the constraint $S'' \succeq 0$ imposes the corresponding off-diagonal elements to be equal to zero as well (i.e., all entries corresponding to $\mathbf{w}_0\mathbf{g}_0^\top, \ldots, \mathbf{w}_0\mathbf{g}_k^\top$ and $\mathbf{g}_i\mathbf{g}_j^\top$ for $j = 0, \ldots, k$ and $i = 0, \ldots, N-1$). It follows that $\beta_{i,0}, \ldots, \beta_{i,i-1} = 0$.

(ii) Using a similar argument: it follows from $\lambda_{N-1,\star} = 0$ that $\lambda_{N-2,N-1} = \lambda_{\star,N-1} = 0$. There is therefore no diagonal element corresponding to the entry $\mathbf{g}_{N-1}\mathbf{g}_{N-1}^\top$ in S'', and the corresponding off-diagonal elements should be zero as well due to the constraint $S'' \succeq 0$. Hence $\alpha_{N-1,0}, \ldots \alpha_{N-1,N-2} = 0$.

(iii) Using (i) and (ii), and for $\{\alpha_{i,j}\}$, the couple $(\tau, \{\lambda_{i,j}\})$ is feasible for (dual-SDP-R) by construction, following the reformulation steps of S in Appendix B. It follows that $\text{UB}_{\mu,L}(\{\alpha_{i,j}\}) \leq \tau$ and Lemma 5 allows reaching the desired claim. □

It is relatively straightforward to establish that ITEM is a solution to (14), given that (i) ITEM achieves the lower complexity bound (see Section 2.3), that (ii) ITEM is a fixed-step first-order method, and that (iii) all the inequalities involved in the proof of Theorem 1 and Theorem 3 are used in the relaxation procedure (R). We conclude this section by the corresponding formal statement.

Theorem 4 *Let $N \in \mathbb{N}$, and $0 \leq \mu < L < \infty$. Algorithm* (3) *is a solution to* (Minimax-R).

Proof We exhibit a solution to (Minimax-R) and show that it corresponds to ITEM in Appendix C. □

Numerical examples of the design procedure are provided in Appendix E, including optimized methods for different objectives, like function values, for which we provide the slightly adapted design strategy in Appendix D. Code for reproducing the results are provided in Section 4.

4 Conclusion

In this work, we provided the *Information-Theoretic Exact Method* (ITEM), a first-order method whose worst-case guarantee exactly matches the lower bound for minimizing smooth strongly convex functions. Furthermore, we showed how to develop such methods constructively, through *performance estimation problems* and semidefinite programming.

We believe that obtaining accelerated first-order methods as solutions to minimax problems certainly brings perspectives and a systematic approach to accelerated methods in first-order convex optimization, similar to the design procedure for obtaining Chebyshev methods for quadratic minimization (see e.g., the review of [d'Aspremont et al., 2021]). In addition, we think that the conceptual simplicity of the shapes and proofs of such optimized methods render them attractive as textbook examples for illustrating the acceleration phenomenon. In particular, it appeared as very surprising to us that both sequences $\{y_k\}$ and $\{z_k\}$ now have relatively clear interpretations: z_k's are optimal for optimizing the distance to an optimal solution, whereas y_k's are essentially optimal for optimizing function values (see [Kim and Fessler, 2016]).

Those methods might as well serve as an inspiration for further developments on this topic, for designing accelerated methods in other settings, and for alternate performance criterion, as showcased numerically in Appendix E.

Finally, let us mention that extending methods such as the Optimized Gradient Method, the Triple Momentum Method, and the Information-Theoretic Exact Method to more general situations, possibly involving constraints, for instance, seems less straightforward compared to other acceleration schemes. We leave this question for future works.

Software Source code for helping to reproduce the slightly algebraic passage in Section 2 can be found in

`https://github.com/AdrienTaylor/Optimal-Gradient-Method`

together with implementations in the Performance Estimation Toolbox [Taylor et al., 2017a] for validating the potential from Lemma 1, the final bound from Theorem 3, and the constructive procedure of Lemma 6.

Acknowledgments The authors would like to thank Shuvomoy Das Gupta for pointing out a few typos and for nice suggestions for improvements.

References

Nikhil Bansal and Anupam Gupta. Potential-function proofs for gradient methods. *Theory of Computing*, 15(1):1–32, 2019.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Saman Cyrus, Bin Hu, Bryan Van Scoy, and Laurent Lessard. A robust accelerated optimization algorithm for strongly convex functions. In *2018 Annual American Control Conference (ACC)*, pages 1376–1381. IEEE, 2018.

Alexandre d'Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. Technical report, HAL, 2021.

Etienne De Klerk, François Glineur, and Adrien B. Taylor. On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. *Optimization Letters*, 11(7):1185–1199, 2017.

Yoel Drori. The exact information-based complexity of smooth convex minimization. *Journal of Complexity*, 39:1–16, 2017.

Yoel Drori and Adrien B. Taylor. Efficient first-order methods for convex minimization: a constructive approach. *Mathematical Programming*, 184(1):183–220, 2020.

Yoel Drori and Adrien B. Taylor. On the oracle complexity of smooth strongly convex minimization. *preprint arXiv:2101.09740*, 2021.

Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Math. Program. Ser. A*, 145:451–482, 2014.

Alexander V. Gasnikov and Yurii E. Nesterov. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58(1):48–64, 2018.

Dennis Gramlich, Christian Ebenbauer, and Carsten W Scherer. Convex synthesis of accelerated gradient algorithms for optimization and saddle point problems using lyapunov functions. *preprint arXiv:2006.09946*, 2020.

Donghwan Kim and Jeffrey A. Fessler. Optimized first-order methods for smooth convex minimization. *Math. Program. Ser. A*, 159(1):81–107, 2016.

Donghwan Kim and Jeffrey A Fessler. On the convergence analysis of the optimized gradient method. *Journal of optimization theory and applications*, 172(1):187–205, 2017.

Donghwan Kim and Jeffrey A. Fessler. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *Journal of optimization theory and applications*, 2020.

Laurent Lessard and Peter Seiler. Direct synthesis of iterative algorithms with bounds on achievable worst-case convergence rate. In *2020 American Control Conference (ACC)*, pages 119–125. IEEE, 2020.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

J. Löfberg. YALMIP : A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD Conference*, 2004.

APS Mosek. The MOSEK optimization software. *Online at http://www.mosek.com*, 54, 2010.

Arkadi Nemirovski. Optimization II: Numerical methods for nonlinear continuous optimization. *Lecture notes, http://www2.isye.gatech.edu/~nemirovs/Lect_OptII.pdf*, 1999.

Arkadi S. Nemirovskii. Information-based complexity of linear operator equations. *Journal of Complexity*, 8(2):153–175, 1992.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Math. Dokl.*, 27(2):372–376, 1983.

Yurii Nesterov. *Introductory lectures on convex optimization: a basic course*. Applied optimization. Kluwer Academic Publishers, 2004. ISBN 9781402075537.

Chanwoo Park, Jisun Park, and Ernest K Ryu. Factor-$\sqrt{2}$ acceleration of accelerated gradient methods. *preprint arXiv:2102.07366*, 2021.

Adrien Taylor. *Convex Interpolation and Performance Estimation of First-order Methods for Convex Optimization*. PhD thesis, Université catholique de Louvain, 2017.

Adrien Taylor and Francis Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In *Proceedings of the 2019 Conference on Learning Theory (COLT)*, volume 99, pages 2934–2992, 2019.

Adrien B. Taylor, Julien M Hendrickx, and François Glineur. Performance estimation toolbox (PESTO): automated worst-case analysis of first-order optimization methods. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1278–1283. IEEE, 2017a.

Adrien B. Taylor, Julien M Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2):307–345, 2017b.

Onur Toker and Hitay Ozbay. On the NP-hardness of solving bilinear matrix inequalities and simultaneous stabilization with static output feedback. In *Proceedings of 1995 American Control Conference*, volume 4, pages 2525–2526. IEEE, 1995.

Charles F. Van Loan and Gene H. Golub. *Matrix computations*. Johns Hopkins University Press Baltimore, 1983.

Bryan Van Scoy, Randy A. Freeman, and Kevin M. Lynch. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1):49–54, 2018.

Ashia C. Wilson, Benjamin Recht, and Michael I. Jordan. A Lyapunov analysis of momentum methods in optimization. *preprint arXiv:1611.02635*, 2016.

Kaiwen Zhou, Anthony Man-Cho So, and James Cheng. Boosting first-order methods by shifting objective: New schemes with faster worst case rates. In *Advances in Neural Information Processing Systems (NeuRIPS)*, 2020.

A Alternate parametrization for first-order methods

In this section, we show that any method (7) can be reparametrized as (8) (and vice-versa), using the identity (9). First note that the equivalence is clear for $k = 1$, as

$$\begin{aligned} w_1 - x_\star &= w_0 - x_\star - \tfrac{h_{1,0}}{L}(\nabla \tilde{f}(w_0) + \mu(w_0 - x_\star)) \\ &= (w_0 - x_\star)(1 - \tfrac{\mu}{L} h_{1,0}) - \tfrac{h_{1,0}}{L}\nabla \tilde{f}(w_0), \end{aligned}$$

and hence the equivalence holds for $k = 1$. Now, assuming the equivalence holds at iteration k, let us check that it holds at iteration $k+1$, that is assume $w_i - x_\star = (w_0 - x_\star)\left(1 - \frac{\mu}{L}\sum_{j=0}^{i-1}\alpha_{i,j}\right) - \frac{1}{L}\sum_{j=0}^{i-1}\alpha_{i,j}\nabla \tilde{f}(w_j)$ for $0 \leq i \leq k$ and compute

$$\begin{aligned} w_{k+1} - x_\star =& w_k - x_\star - \tfrac{1}{L}\sum_{i=0}^{k} h_{k+1,i}(\nabla \tilde{f}(w_i) + \mu(w_i - x_\star)) \\ =& (w_0 - x_\star)\left(1 - \tfrac{\mu}{L}\sum_{i=0}^{k-1}\alpha_{k,i} - \tfrac{\mu}{L}\sum_{i=0}^{k} h_{k+1,i} + \tfrac{\mu^2}{L^2}\sum_{i=0}^{k}\sum_{j=0}^{k-1} h_{k+1,i}\alpha_{i,j}\right) \\ & - \tfrac{1}{L}\sum_{i=0}^{k-1}\alpha_{k,i}\nabla \tilde{f}(w_i) - \tfrac{1}{L}\sum_{i=0}^{k} h_{k+1,i}\nabla \tilde{f}(w_i) + \tfrac{\mu}{L^2}\sum_{i=0}^{k}\sum_{j=0}^{k-1} h_{k+1,i}\alpha_{i,j}\nabla \tilde{f}(w_j) \end{aligned}$$

by reverting the ordering of the double sums, renaming the indices, and reordering, we get

$$\begin{aligned} w_{k+1} - x_\star =& (w_0 - x_\star)\left(1 - \tfrac{\mu}{L}\sum_{i=0}^{k-1}\alpha_{k,i} - \tfrac{\mu}{L}\sum_{i=0}^{k} h_{k+1,i} + \tfrac{\mu^2}{L^2}\sum_{j=0}^{k-1}\sum_{i=j+1}^{k} h_{k+1,i}\alpha_{i,j}\right) \\ & - \tfrac{1}{L}\sum_{i=0}^{k-1}\alpha_{k,i}\nabla \tilde{f}(w_i) - \tfrac{1}{L}\sum_{i=0}^{k} h_{k+1,i}\nabla \tilde{f}(w_i) + \tfrac{\mu}{L^2}\sum_{j=0}^{k-1}\sum_{i=j+1}^{k} h_{k+1,i}\alpha_{i,j}\nabla \tilde{f}(w_j) \\ =& (w_0 - x_\star)\left(1 - \tfrac{\mu}{L}\sum_{i=0}^{k-1}\alpha_{k,i} - \tfrac{\mu}{L}\sum_{i=0}^{k} h_{k+1,i} + \tfrac{\mu^2}{L^2}\sum_{i=0}^{k-1}\sum_{j=i+1}^{k} h_{k+1,j}\alpha_{j,i}\right) \\ & - \tfrac{1}{L}\sum_{i=0}^{k-1}\alpha_{k,i}\nabla \tilde{f}(w_i) - \tfrac{1}{L}\sum_{i=0}^{k} h_{k+1,i}\nabla \tilde{f}(w_i) + \tfrac{\mu}{L^2}\sum_{i=0}^{k-1}\sum_{j=i+1}^{k} h_{k+1,j}\alpha_{j,i}\nabla \tilde{f}(w_i) \\ =& (w_0 - x_\star)\left(1 - \tfrac{\mu}{L}h_{k+1,k} - \tfrac{\mu}{L}\sum_{i=0}^{k-1}\left(\alpha_{k,i} + h_{k+1,i} - \tfrac{\mu}{L}\sum_{j=i+1}^{k} h_{k+1,j}\alpha_{j,i}\right)\right) \\ & - \tfrac{1}{L}h_{k+1,k}\nabla \tilde{f}(w_k) - \tfrac{1}{L}\sum_{i=0}^{k-1}\left(\alpha_{k,i} + h_{k+1,i} - \tfrac{\mu}{L}\sum_{j=i+1}^{k} h_{k+1,j}\alpha_{j,i}\right)\nabla \tilde{f}(w_i). \end{aligned}$$

From this last reformulation, the choice (9), that is

$$\alpha_{k+1,i} = \begin{cases} h_{k+1,k} & \text{if } i = k \\ h_{k+1,i} + \alpha_{k,i} - \frac{\mu}{L}\sum_{j=i+1}^{k} h_{k+1,j}\alpha_{j,i} & \text{if } 0 \leq i < k, \end{cases}$$

allows enforcing the coefficients of all independent terms $(w_0 - x_\star)$, $\nabla \tilde{f}(w_0), \ldots, \nabla \tilde{f}(w_k)$ to be equal in both (7) and (8), reaching the desired statement. In addition, note that this change of variable is reversible.

B Algebraic manipulations of (dual-SDP-R)

In this section, we reformulate (dual-SDP-R) for enabling us optimizing both on $\alpha_{i,j}$'s and $\lambda_{i,j}$'s simultaneously. For doing that, let us start by conveniently noting that

$$S(\tau,\{\lambda_{i,j},\{\alpha_{i,j}\}\})\succeq 0 \Leftrightarrow \begin{pmatrix} S'(\tau,\{\lambda_{i,j}\},\{\alpha_{i,j}\}) & \mathbf{w}_N \\ \mathbf{w}_N^\top & 1\end{pmatrix}\succeq 0,$$

with $S'(\tau,\{\lambda_{i,j}\},\{\alpha_{i,j}\}) = S(\tau,\{\lambda_{i,j}\},\{\alpha_{i,j}\})+\mathbf{w}_N\mathbf{w}_N^\top$, using a standard Schur complement (see, e.g., [Van Loan and Golub, 1983]). The motivation underlying this reformulation is that this *lifted* linear matrix inequality depends linearly on $\{\alpha_{N,i}\}_i$'s. Indeed, the coefficients of the last iteration only appear through the term $\mathbf{w}_N$, which is not present in S' (details below).

We only consider the case $N>1$ below. In the case $N=1$, the (6) can be solved without the following simplifications. Let us develop the expression of $S'(.)$ as follows

$$\begin{aligned}
S'(\tau,\{\lambda_{i,j}\},\{\alpha_{i,j}\}) =& \tau\mathbf{w}_0\mathbf{w}_0^\top + \tfrac{1}{2(L-\mu)}\left(\lambda_{N-1,\star}\,\mathbf{g}_{N-1}\mathbf{g}_{N-1}^\top + \sum_{i=0}^{N-1}\lambda_{\star,i}\mathbf{g}_i\mathbf{g}_i^\top + \sum_{i=0}^{N-2}\lambda_{i,i+1}(\mathbf{g}_i-\mathbf{g}_{i+1})(\mathbf{g}_i-\mathbf{g}_{i+1})^\top\right)\\
&- \tfrac{1}{2}\lambda_{\star,0}(\mathbf{g}_0\mathbf{w}_0^\top+\mathbf{w}_0\mathbf{g}_0^\top) - \tfrac{1}{2}\sum_{i=1}^{N-1}(\lambda_{i-1,i}+\lambda_{\star,i})(\mathbf{g}_i\mathbf{w}_i^\top+\mathbf{w}_i\mathbf{g}_i^\top)\\
&+ \tfrac{1}{2}\sum_{i=0}^{N-2}\lambda_{i,i+1}(\mathbf{g}_{i+1}\mathbf{w}_i^\top+\mathbf{w}_i\mathbf{g}_{i+1}^\top)\\
=& \tau\mathbf{w}_0\mathbf{w}_0^\top + \tfrac{1}{2(L-\mu)}\left(\lambda_{N-1,\star}\,\mathbf{g}_{N-1}\mathbf{g}_{N-1}^\top + \sum_{i=0}^{N-1}\lambda_{\star,i}\mathbf{g}_i\mathbf{g}_i^\top + \sum_{i=0}^{N-2}\lambda_{i,i+1}(\mathbf{g}_i-\mathbf{g}_{i+1})(\mathbf{g}_i-\mathbf{g}_{i+1})^\top\right)\\
&- \tfrac{1}{2}\lambda_{\star,0}(\mathbf{g}_0\mathbf{w}_0^\top+\mathbf{w}_0\mathbf{g}_0^\top) - \tfrac{1}{2}\sum_{i=1}^{N-2}\lambda_{i,i+1}(\mathbf{g}_i\mathbf{w}_i^\top+\mathbf{w}_i\mathbf{g}_i^\top)\\
&- \tfrac{1}{2}\lambda_{N-1,\star}(\mathbf{g}_{N-1}\mathbf{w}_{N-1}^\top+\mathbf{w}_{N-1}\mathbf{g}_{N-1}^\top) + \tfrac{1}{2}\sum_{i=0}^{N-2}\lambda_{i,i+1}(\mathbf{g}_{i+1}\mathbf{w}_i^\top+\mathbf{w}_i\mathbf{g}_{i+1}^\top),
\end{aligned}$$

where we used $\lambda_{i-1,i}+\lambda_{\star,i}=\lambda_{i,i+1}$ (for $i=1,\dots,N-2$), and $\lambda_{N-2,N-1}+\lambda_{\star,N-1}=\lambda_{N-1,\star}$ for obtaining the second equality. Substituting the expressions for $\mathbf{w}_i$'s, we arrive to

$$\begin{aligned}
S'(\tau,\{\lambda_{i,j}\},\{\alpha_{i,j}\}) =& \tau\mathbf{w}_0\mathbf{w}_0^\top + \tfrac{1}{2(L-\mu)}\left(\lambda_{N-1,\star}\,\mathbf{g}_{N-1}\mathbf{g}_{N-1}^\top + \sum_{i=0}^{N-1}\lambda_{\star,i}\mathbf{g}_i\mathbf{g}_i^\top + \sum_{i=0}^{N-2}\lambda_{i,i+1}(\mathbf{g}_i-\mathbf{g}_{i+1})(\mathbf{g}_i-\mathbf{g}_{i+1})^\top\right)\\
&- \tfrac{1}{2}\lambda_{\star,0}(\mathbf{g}_0\mathbf{w}_0^\top+\mathbf{w}_0\mathbf{g}_0^\top) - \sum_{i=1}^{N-2}\lambda_{i,i+1}\left(1-\tfrac{\mu}{L}\sum_{j=0}^{i-1}\alpha_{i,j}\right)\tfrac{1}{2}(\mathbf{g}_i\mathbf{w}_0^\top+\mathbf{w}_0\mathbf{g}_i^\top)\\
&+ \sum_{i=1}^{N-2}\lambda_{i,i+1}\sum_{j=0}^{i-1}\tfrac{\alpha_{i,j}}{L}\tfrac{1}{2}(\mathbf{g}_i\mathbf{g}_j^\top+\mathbf{g}_j\mathbf{g}_i^\top) - \lambda_{N-1,\star}\left(1-\tfrac{\mu}{L}\sum_{j=0}^{N-2}\alpha_{N-1,j}\right)\tfrac{1}{2}(\mathbf{g}_{N-1}\mathbf{w}_0^\top+\mathbf{w}_0\mathbf{g}_{N-1}^\top)\\
&+ \lambda_{N-1,\star}\sum_{j=0}^{N-2}\tfrac{\alpha_{N-1,j}}{L}\tfrac{1}{2}(\mathbf{g}_{N-1}\mathbf{g}_j^\top+\mathbf{g}_j\mathbf{g}_{N-1}^\top) + \sum_{i=0}^{N-2}\lambda_{i,i+1}\left(1-\tfrac{\mu}{L}\sum_{j=0}^{i-1}\alpha_{i,j}\right)\tfrac{1}{2}(\mathbf{g}_{i+1}\mathbf{w}_0^\top+\mathbf{w}_0\mathbf{g}_{i+1}^\top)\\
&- \sum_{i=0}^{N-2}\lambda_{i,i+1}\sum_{j=0}^{i-1}\tfrac{\alpha_{i,j}}{L}\tfrac{1}{2}(\mathbf{g}_{i+1}\mathbf{g}_j^\top+\mathbf{g}_j\mathbf{g}_{i+1}^\top),
\end{aligned}$$

where we simply expressed each $\mathbf{w}_k$ in two terms, one with the contribution of $\mathbf{w}_0$, and the other with the contributions of $\mathbf{g}_i$'s. Although not pretty, one can observe that S' is still bilinear in $\{\alpha_{i,j}\}$ and $\{\lambda_{i,j}\}$. This expression can be largely simplified, but this form suffices for the purposes in this work.

C The Information-Theoretic Exact Method is a solution to (Minimax-R)

Proof of Theorem 4 For readability purposes, we establish the claim without explicitly computing the optimal values of the variables $\{\alpha_{i,j}\}$ and $\{\beta_{i,j}\}$ for ITEM. For avoiding this step, let us note that ITEM is clearly a fixed-step first-order method following Definition 2. Therefore, following Lemma 3, the method can also be written in the alternate parametrization (8),

using the association $w_i \leftarrow y_i$ $(i = 0, 1, \ldots, N-1)$ and $w_N \leftarrow z_N$, where y_k and z_k are the sequences defined by (3). Let $\{\alpha^\star_{i,j}\}$ denote the steps sizes corresponding to ITEM written in the form (8). We proceed to show that by choosing

$$\begin{aligned}
\lambda^\star_{\star,i} &= \frac{1-q}{L}\frac{A_{i+1}-A_i}{1+qA_N} \\
\lambda^\star_{i,i+1} &= \frac{1-q}{L}\frac{A_{i+1}}{1+qA_N} \\
\lambda^\star_{N-1,\star} &= \frac{1-q}{L}\frac{A_N}{1+qA_N} \\
\tau^\star &= \frac{1}{1+qA_N}
\end{aligned} \tag{16}$$

and setting $\{\beta^\star_{i,j}\}$ in accordance to (14), we reach a feasible solution to (Minimax-R). Note that optimality of the solution follows from the value of $\tau^\star$, which matches the lower complexity bound discussed in Section 2.3.

For establishing dual feasibility, we relate (Minimax-R) to the Lagrangian of (R). That is, denoting

$$K = S''(\tau^\star, \{\lambda^\star_{i,j}\}, \{\beta^\star_{i,j}\}) - \mathbf{w}_N \mathbf{w}_N^\top = S(\tau^\star, \{\lambda^\star_{i,j}\}, \{\alpha^\star_{i,j}\}),$$

we have for all (F, G) as in (11), by construction,

$$\begin{aligned}
&F\left(\sum_{i=0}^{N-2} \lambda^\star_{i,i+1}(\mathbf{f}_{i+1}-\mathbf{f}_i) + \sum_{i=0}^{N-1}\lambda^\star_{\star,i}\mathbf{f}_i - \lambda^\star_{N-1,\star}\mathbf{f}_{N-1}\right) + \mathrm{Tr}(KG) \\
&= \tau^\star\|w_0 - w_\star\|^2 - \|w_N - w_\star\|^2 \\
&\quad + \sum_{i=0}^{N-2}\lambda^\star_{i,i+1}\left[\tilde f(w_{i+1}) - \tilde f(w_i) + \langle \nabla\tilde f(w_{i+1}); w_i - w_{i+1}\rangle + \tfrac{1}{2(L-\mu)}\|\nabla\tilde f(w_i) - \nabla\tilde f(w_{i+1})\|^2\right] \\
&\quad + \sum_{i=0}^{N-1}\lambda^\star_{\star,i}\left[\tilde f(w_i) - \tilde f_\star + \langle\nabla\tilde f(w_i), w_\star - w_i\rangle + \tfrac{1}{2(L-\mu)}\|\nabla\tilde f(w_i)\|^2\right] \\
&\quad + \lambda^\star_{N-1,\star}\left[\tilde f_\star - \tilde f(w_{N-1}) + \tfrac{1}{2(L-\mu)}\|\nabla\tilde f(w_{N-1})\|^2\right].
\end{aligned}$$

Using the association $w_i \leftarrow y_i$ $(i = 0, 1, \ldots, N-1)$ and $w_N \leftarrow z_N$, as well as $f(y_i) = \tilde f(y_i) + \frac{\mu}{2}\|y_i - x_\star\|^2$, it follows from Lemma 1 (and in particular the weighted sum in its proof) that for $i = 1, \ldots, N-1$

$$\begin{aligned}
&\lambda^\star_{\star,i}\left[\tilde f(w_i) - \tilde f_\star + \langle\nabla\tilde f(w_i), w_\star - w_i\rangle + \tfrac{1}{2(L-\mu)}\|\nabla\tilde f(w_i)\|^2\right] \\
&\quad + \lambda^\star_{i-1,i}\left[\tilde f(w_i) - \tilde f(w_{i-1}) + \langle\nabla\tilde f(w_i); w_{i-1} - w_i\rangle + \tfrac{1}{2(L-\mu)}\|\nabla\tilde f(w_i) - \nabla\tilde f(w_{i-1})\|^2\right] = \frac{1}{L+\mu A_N}(\phi_{i+1} - \phi_i),
\end{aligned}$$

as well as

$$\begin{aligned}
\lambda^\star_{\star,0}\left[\tilde f(w_0) - \tilde f_\star + \langle\nabla\tilde f(w_0), w_\star - w_0\rangle + \tfrac{1}{2(L-\mu)}\|\nabla\tilde f(w_0)\|^2\right] &= \frac{1}{L+\mu A_N}(\phi_1 - \phi_0), \\
\lambda^\star_{N-1,\star}\left[\tilde f_\star - \tilde f(w_{N-1}) + \tfrac{1}{2(L-\mu)}\|\nabla\tilde f(w_{N-1})\|^2\right] &= -\frac{(1-q)A_N}{L+\mu A_N}\psi_{N-1}.
\end{aligned}$$

In addition, noting that $\phi_0 = L\|w_0 - w_\star\|^2$ allows reaching the following reformulation

$$\begin{aligned}
&F\left(\sum_{i=0}^{N-2} \lambda^\star_{i,i+1}(\mathbf{f}_{i+1}-\mathbf{f}_i) + \sum_{i=0}^{N-1}\lambda^\star_{\star,i}\mathbf{f}_i - \lambda^\star_{N-1,\star}\mathbf{f}_{N-1}\right) + \mathrm{Tr}(KG) \\
&= \frac{1}{L+\mu A_N}\left(\phi_0 - (1-q)A_N\psi_{N-1} + \sum_{i=0}^{N-1}(\phi_{i+1} - \phi_i)\right) - \|z_N - w_\star\|^2 \\
&= 0,
\end{aligned}$$

where the last equality follows from $\phi_N = (1-q)A_N\psi_{N-1} + (L+\mu A_N)\|z_N - w_\star\|^2$. Therefore, ITEM is a solution to (Minimax-R). In more direct terms of the SDP (Minimax-R), one can verify that

$$\begin{aligned}
&\lambda^\star_{\star,0} - \lambda^\star_{0,1} = \frac{1-q}{L}\left(\frac{A_1}{1+qA_N} - \frac{A_1}{1+qA_N}\right) = 0 \\
&\lambda^\star_{i-1,i} + \lambda^\star_{\star,i} - \lambda^\star_{i,i+1} = \frac{1-q}{L}\left(\frac{A_i}{1+qA_N} + \frac{A_{i+1}-A_i}{1+qA_N} - \frac{A_{i+1}}{1+qA_N}\right) = 0 \qquad \text{for } i = 1, \ldots, N-2 \\
&\lambda^\star_{N-2,N-1} + \lambda^\star_{\star,N-1} - \lambda^\star_{N-1,\star} = \frac{1-q}{L}\left(\frac{A_{N-1}}{1+qA_N} + \frac{A_N - A_{N-1}}{1+qA_N} - \frac{A_N}{1+qA_N}\right) = 0,
\end{aligned}$$

the previous computations therefore imply that $K = 0$ for ITEM, and hence $S''(\tau, \{\lambda_{i,j}\}, \{\beta_{i,j}\}) = \mathbf{w}_N\mathbf{w}_N^\top \succeq 0$. □

D An SDP formulation for optimizing function values

In this section, we show how to adapt the methodology developed in Section 3 for a family of alternate design criteria, which include $(f(w_N)-f_\star)/\|w_0-w_\star\|^2$ and $(f(w_N)-f_\star)/(f(w_0)-f_\star)$ (for which numerical examples are provided respectively in Section E.1 and Section E.2). The developments slightly differ from those required for optimizing $\|w_N-w_\star\|^2/\|w_0-w_\star\|^2$; and we decided not to present a unified version in the core of the text, for readability purposes. In particular, the set of selected inequalities is slightly different, altering the linearization procedure.

The criteria we deal with in this section are of the form

$$\frac{f(w_N)-f_\star}{c_w\|w_0-w_\star\|^2+c_f(f(w_0)-f_\star)}=\frac{\tilde{f}(w_N)-f_\star+\frac{\mu}{2}\|w_N-w_\star\|^2}{c_w\|w_0-w_\star\|^2+c_f(\tilde{f}(w_0)-f_\star+\frac{\mu}{2}\|w_0-w_\star\|^2)}.$$

As the steps are essentially the same as detailed in Section 3, we proceed without providing much detail. We start with the discrete version, using the set $I=\{\star,0,\dots,N\}$

$$\begin{aligned}\max_{\substack{\{(w_i,g_i,f_i)\}_{i\in I}\\ d\in\mathbb{N}}}\ & f_N-f_\star+\tfrac{\mu}{2}\|w_N-w_\star\|^2 \\ \text{s.t. } & c_w\|w_0-w_\star\|^2+c_f(f_0-f_\star+\tfrac{\mu}{2}\|w_0-w_\star\|^2)=1,\ g_\star=0 \\ & w_k \text{ generated by (8)} && \text{for } k=1,\dots,N \\ & f_i\geq f_j+\langle g_j;w_i-w_j\rangle+\tfrac{1}{2(L-\mu)}\|g_i-g_j\|^2 && \text{for all } i,j\in I.\end{aligned}$$

The upper bound we use is now very slightly different (the selected subset of constraints is not the same as that of (R))

$$\begin{aligned}\mathrm{UB}_{\mu,L}(\{\alpha_{i,j}\})=\max_{\substack{\{(w_i,g_i,f_i)\}_{i\in I}\\ d\in\mathbb{N}}}\ & f_N-f_\star+\tfrac{\mu}{2}\|w_N-w_\star\|^2 \\ \text{s.t. } & c_w\|w_0-w_\star\|^2+c_f(f_0-f_\star+\tfrac{\mu}{2}\|w_0-w_\star\|^2)=1,\ g_\star=0 \\ & w_k \text{ generated by (8)} && \text{for } k=1,\dots,N \\ & f_i\geq f_{i+1}+\langle g_{i+1};w_i-w_{i+1}\rangle+\tfrac{1}{2(L-\mu)}\|g_i-g_{i+1}\|^2 && \text{for } i=0,\dots,N-1 \\ & f_\star\geq f_i+\langle g_i,w_\star-w_{i+1}\rangle+\tfrac{1}{2(L-\mu)}\|g_{i+1}\|^2 && \text{for } i=0,\dots,N\end{aligned}$$

The corresponding SDP can be written using a similar couple (G,F)

$$G=\begin{pmatrix}\|w_0-w_\star\|^2 & \langle g_0;w_0-w_\star\rangle & \langle g_1;w_0-w_\star\rangle & \dots & \langle g_N;w_0-w_\star\rangle\\ \langle g_0;w_0-w_\star\rangle & \|g_0\|^2 & \langle g_1;g_0\rangle & \dots & \langle g_N;g_0\rangle\\ \langle g_1;w_0-w_\star\rangle & \langle g_1;g_0\rangle & \|g_1\|^2 & \dots & \langle g_N;g_1\rangle\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \langle g_N;w_0-w_\star\rangle & \langle g_N;g_0\rangle & \langle g_N;g_1\rangle & \dots & \|g_N\|^2\end{pmatrix}$$

$$F=\begin{pmatrix}f_0-f_\star\\ f_1-f_\star\\ \vdots\\ f_N-f_\star\end{pmatrix},$$

and the similar notations

$$\mathbf{w}_0=e_1\in\mathbb{R}^{N+2},\quad \mathbf{g}_i=e_{i+2}\in\mathbb{R}^{N+2},\quad \mathbf{f}_i=e_{i+1}\in\mathbb{R}^{N+1},$$

with $i=0,\dots,N$ and e_i being the unit vector whose ith component is equal to 1. In addition, we can also denote by

$$\mathbf{w}_k=\mathbf{w}_0\left(1-\tfrac{\mu}{L}\sum_{i=0}^{k-1}\alpha_{k,i}\right)-\sum_{i=0}^{k-1}\tfrac{\alpha_{k,i}}{L}\mathbf{g}_i.$$

A dual formulation of $\mathrm{UB}_{\mu,L}$ is given by (we directly included the Schur complement)

$$\begin{aligned}\mathrm{UB}_{\mu,L}(\{\alpha_{i,j}\})=\min_{\tau,\lambda_{i,j}\geq 0}\ & \tau, \\ \text{s.t. } & \begin{pmatrix}\bar{S}'(\tau,\{\lambda_{i,j}\},\{\alpha_{i,j}\}) & \sqrt{\mu}\mathbf{w}_N\\ \sqrt{\mu}\mathbf{w}_N^\top & 2\end{pmatrix}\succeq 0, \\ & \tau\, c_f\,\mathbf{f}_0+\sum_{i=0}^{N-1}\lambda_{i,i+1}(\mathbf{f}_{i+1}-\mathbf{f}_i)+\sum_{i=0}^{N}\lambda_{\star,i}\mathbf{f}_i=\mathbf{f}_N\end{aligned}$$

with

$$\bar{S}'(\tau,\{\lambda_{i,j}\},\{\alpha_{i,j}\}) = \tau\,(c_w + c_f\tfrac{\mu}{2})\mathbf{w}_0\mathbf{w}_0^\top + \sum_{i=0}^{N}\frac{\lambda_{\star,i}}{2}\left(-\mathbf{g}_i\mathbf{w}_i^\top - \mathbf{w}_i\mathbf{g}_i^\top + \tfrac{1}{L-\mu}\mathbf{g}_i\mathbf{g}_i^\top\right)$$
$$+ \sum_{i=0}^{N-1}\frac{\lambda_{i,i+1}}{2}\left(\mathbf{g}_{i+1}(\mathbf{w}_i-\mathbf{w}_{i+1})^\top + (\mathbf{w}_i-\mathbf{w}_{i+1})\mathbf{g}_{i+1}^\top + \tfrac{1}{L-\mu}(\mathbf{g}_i-\mathbf{g}_{i+1})(\mathbf{g}_i-\mathbf{g}_{i+1})^\top\right).$$

Note that that the equality constraint corresponds to

$$\begin{aligned} &c_f + \lambda_{\star,0} - \lambda_{0,1} = 0\\ &\lambda_{i-1,i} + \lambda_{\star,i} - \lambda_{i,i+1} = 0 \qquad \text{for } i = 1,\dots,N-1\\ &\lambda_{N-1,N} + \lambda_{\star,N} = 1. \end{aligned}$$

We perform a some additional work on $\bar{S}'$ (whose dependency on $\{\alpha_{i,j}\}$ is implicit through the dependency on $\{\mathbf{w}_i\}$), as before

$$\begin{aligned}\bar{S}'(\tau,\{\lambda_{i,j}\},\{\alpha_{i,j}\}) &= \tau\,(c_w + c_f\tfrac{\mu}{2})\mathbf{w}_0\mathbf{w}_0^\top + \frac{1}{2(L-\mu)}\left(\sum_{i=0}^{N}\lambda_{\star,i}\mathbf{g}_i\mathbf{g}_i^\top + \sum_{i=0}^{N-1}\lambda_{i,i+1}(\mathbf{g}_i-\mathbf{g}_{i+1})(\mathbf{g}_i-\mathbf{g}_{i+1})^\top\right)\\ &\quad - \sum_{i=0}^{N}\lambda_{\star,i}\tfrac{1}{2}\left(\mathbf{g}_i\mathbf{w}_i^\top + \mathbf{w}_i\mathbf{g}_i^\top\right) + \sum_{i=0}^{N-1}\lambda_{i,i+1}\tfrac{1}{2}\left(\mathbf{g}_{i+1}(\mathbf{w}_i-\mathbf{w}_{i+1})^\top + (\mathbf{w}_i-\mathbf{w}_{i+1})\mathbf{g}_{i+1}^\top\right)\\ &= \tau\,(c_w + c_f\tfrac{\mu}{2})\mathbf{w}_0\mathbf{w}_0^\top + \frac{1}{2(L-\mu)}\left(\sum_{i=0}^{N}\lambda_{\star,i}\mathbf{g}_i\mathbf{g}_i^\top + \sum_{i=0}^{N-1}\lambda_{i,i+1}(\mathbf{g}_i-\mathbf{g}_{i+1})(\mathbf{g}_i-\mathbf{g}_{i+1})^\top\right)\\ &\quad - \lambda_{\star,0}\tfrac{1}{2}(\mathbf{g}_0\mathbf{w}_0^\top + \mathbf{w}_0\mathbf{g}_0^\top) - \sum_{i=1}^{N}(\lambda_{\star,i}+\lambda_{i-1,i})\tfrac{1}{2}(\mathbf{g}_i\mathbf{w}_i^\top + \mathbf{w}_i\mathbf{g}_i^\top)\\ &\quad + \sum_{i=0}^{N-1}\lambda_{i,i+1}\tfrac{1}{2}(\mathbf{g}_{i+1}\mathbf{w}_i^\top + \mathbf{w}_i\mathbf{g}_{i+1}^\top)\\ &= \tau\,(c_w + c_f\tfrac{\mu}{2})\mathbf{w}_0\mathbf{w}_0^\top + \frac{1}{2(L-\mu)}\left(\sum_{i=0}^{N}\lambda_{\star,i}\mathbf{g}_i\mathbf{g}_i^\top + \sum_{i=0}^{N-1}\lambda_{i,i+1}(\mathbf{g}_i-\mathbf{g}_{i+1})(\mathbf{g}_i-\mathbf{g}_{i+1})^\top\right)\\ &\quad - \lambda_{\star,0}\tfrac{1}{2}(\mathbf{g}_0\mathbf{w}_0^\top + \mathbf{w}_0\mathbf{g}_0^\top) - \sum_{i=1}^{N-1}\lambda_{i,i+1}\tfrac{1}{2}(\mathbf{g}_i\mathbf{w}_i^\top + \mathbf{w}_i\mathbf{g}_i^\top) - \tfrac{1}{2}(\mathbf{g}_N\mathbf{w}_N^\top + \mathbf{w}_N\mathbf{g}_N^\top)\\ &\quad + \sum_{i=0}^{N-1}\lambda_{i,i+1}\tfrac{1}{2}(\mathbf{g}_{i+1}\mathbf{w}_i^\top + \mathbf{w}_i\mathbf{g}_{i+1}^\top),\end{aligned}$$

where we used $\lambda_{\star,i} + \lambda_{i-1,i} = \lambda_{i,i+1}$ (for $i = 1,\dots,N-1$) and $\lambda_{\star,N} + \lambda_{N-1,N} = 1$. Now, making the dependence on $\alpha_{i,j}$'s explicit again, we arrive to

$$\begin{aligned}\bar{S}'(\tau,\{\lambda_{i,j}\},\{\alpha_{i,j}\}) &= \tau\,(c_w + c_f\tfrac{\mu}{2})\mathbf{w}_0\mathbf{w}_0^\top + \frac{1}{2(L-\mu)}\left(\sum_{i=0}^{N}\lambda_{\star,i}\mathbf{g}_i\mathbf{g}_i^\top + \sum_{i=0}^{N-1}\lambda_{i,i+1}(\mathbf{g}_i-\mathbf{g}_{i+1})(\mathbf{g}_i-\mathbf{g}_{i+1})^\top\right)\\ &\quad - \lambda_{\star,0}\tfrac{1}{2}(\mathbf{g}_0\mathbf{w}_0^\top + \mathbf{w}_0\mathbf{g}_0^\top) - \sum_{i=1}^{N-1}\lambda_{i,i+1}\left(1 - \tfrac{\mu}{L}\sum_{j=0}^{i-1}\alpha_{i,j}\right)\tfrac{1}{2}(\mathbf{g}_i\mathbf{w}_0^\top + \mathbf{w}_0\mathbf{g}_i^\top)\\ &\quad + \sum_{i=1}^{N-1}\lambda_{i,i+1}\sum_{j=0}^{i-1}\alpha_{i,j}\tfrac{1}{2}(\mathbf{g}_i\mathbf{g}_j^\top + \mathbf{g}_j\mathbf{g}_i^\top) - \left(1 - \tfrac{\mu}{L}\sum_{j=0}^{N-1}\alpha_{N,j}\right)\tfrac{1}{2}(\mathbf{g}_N\mathbf{w}_0^\top + \mathbf{w}_0\mathbf{g}_N^\top)\\ &\quad + \sum_{j=0}^{N-1}\alpha_{N,j}\tfrac{1}{2}(\mathbf{g}_N\mathbf{g}_j^\top + \mathbf{g}_j\mathbf{g}_N^\top) + \sum_{i=0}^{N-1}\lambda_{i,i+1}\left(1 - \tfrac{\mu}{L}\sum_{j=0}^{i-1}\alpha_{i,j}\right)\tfrac{1}{2}(\mathbf{g}_{i+1}\mathbf{w}_0^\top + \mathbf{w}_0\mathbf{g}_{i+1}^\top)\\ &\quad - \sum_{i=0}^{N-1}\lambda_{i,i+1}\sum_{j=0}^{i-1}\alpha_{i,j}\tfrac{1}{2}(\mathbf{g}_{i+1}\mathbf{g}_j^\top + \mathbf{g}_j\mathbf{g}_{i+1}^\top)\end{aligned}$$

and it remains to remark that the change of variables

$$\beta_{i,j} = \begin{cases}\lambda_{i,i+1}\alpha_{i,j} & \text{if } 0 \le i \le N-1\\ \alpha_{N,j} & \text{if } i = N.\end{cases} \tag{17}$$

linearizes the bilinear matrix inequality, again, and it remains to solve the SDP (13) using standard packages. Numerical results for the pairs $(c_w, c_f) = (1, 0)$ and $(c_w, c_f) = (0, 1)$ are respectively provided in Section E.1 and Section E.2. A source code for implementing those SDP is provided in Section 4.

E Numerical examples

As shown in Appendix D, slight modifications of the relaxations used for obtaining (Minimax-R) allows forming tractable problems for optimizing the parameters of fixed-step methods under different optimality criteria. Although we were unable to obtain closed-form solutions to the problems arising for these alternative criteria, the resulting problems can still be approximated numerically for specific values of μ, L and N.

In the following, we provide a couple of examples that were obtained by numerically solving the first-order method design problem (Minimax-R), formulated as a linear semidefinite program using standard solvers [Löfberg, 2004, Mosek, 2010].

E.1 Optimized methods for $(f(w_N) - f_\star)/\|w_0 - w_\star\|^2$

As a first example, we consider the criterion $(f(w_N) - f_\star)/\|w_0 - w_\star\|$. The following list provides solutions obtained by solving the corresponding design problem for $N = 1, ..., 5$ with $L = 1$ and $\mu = .1$. The solutions are presented using the notations from (7) together with the corresponding worst-case guarantees.

- For a single iteration, by solving the corresponding optimization problem, we obtain a method with guarantee $\frac{f(w_1)-f_\star}{\|w_0-w_\star\|} \leq 0.1061$ and step size
$$[h^\star_{i,j}] = \begin{bmatrix} 1.4606 \end{bmatrix}.$$
This bound and the corresponding step size match the optimal step size $h_{1,0} = \frac{q+1-\sqrt{q^2-q+1}}{q}$, see [Taylor, 2017, Theorem 4.14].
- For $N = 2$ iterations, we obtain $\frac{f(w_2)-f_\star}{\|w_0-w_\star\|} \leq 0.0418$ with
$$[h^\star_{i,j}] = \begin{bmatrix} 1.5567 & \\ 0.1016 & 1.7016 \end{bmatrix}.$$
- For $N = 3$, we obtain $\frac{f(w_3)-f_\star}{\|w_0-w_\star\|} \leq 0.0189$ with
$$[h^\star_{i,j}] = \begin{bmatrix} 1.5512 & & \\ 0.1220 & 1.8708 & \\ 0.0316 & 0.2257 & 1.8019 \end{bmatrix}.$$
- For $N = 4$, we obtain $\frac{f(w_4)-f_\star}{\|w_0-w_\star\|} \leq 0.0089$, with
$$[h^\star_{i,j}] = \begin{bmatrix} 1.5487 & & & \\ 0.1178 & 1.8535 & & \\ 0.0371 & 0.2685 & 2.0018 & \\ 0.0110 & 0.0794 & 0.2963 & 1.8497 \end{bmatrix}.$$
- Finally, for $N = 5$, we obtain $\frac{f(w_5)-f_\star}{\|w_0-w_\star\|} \leq 0.0042$ with
$$[h^\star_{i,j}] = \begin{bmatrix} 1.5476 & & & & \\ 0.1159 & 1.8454 & & & \\ 0.0350 & 0.2551 & 1.9748 & & \\ 0.0125 & 0.0913 & 0.3489 & 2.0625 & \\ 0.0039 & 0.0287 & 0.1095 & 0.3334 & 1.8732 \end{bmatrix}.$$

Note that when $\mu = 0$, we recover the step size policy of the OGM by Kim and Fessler [2016]. When setting $\mu > 0$, we observe that the resulting optimized method is apparently less practical as the step sizes critically depend on the horizon N. In particular, one can observe that $h^\star_{1,0}$ varies with the horizon N.

Figure 1 illustrates the behavior of the worst-case guarantee for larger values of N and compares it to the currently best known corresponding lower bound, as well as to worst-case guarantees for TMM, Nesterov's Fast Gradient Method (FGM) for strongly convex functions, as well as to the methods generated with the SSEP procedure from [Drori and Taylor, 2020]. All the worst-case guarantees are computed numerically using the corresponding performance estimation problems (see e.g., the toolbox [Taylor et al., 2017a]), and as a result, they are tight in the sense that matching inputs to the algorithms attaining the bounds can be numerically constructed.

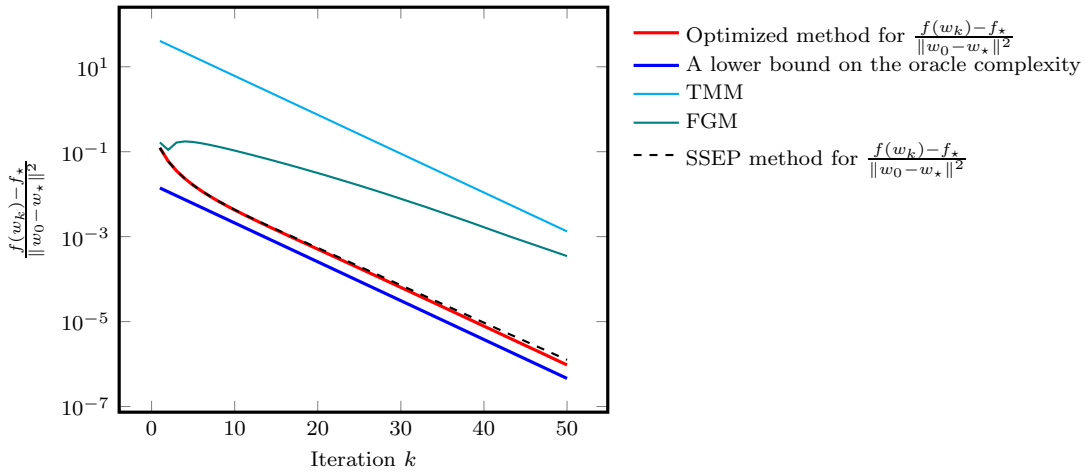

Fig. 1 Numerical comparison (for $L = 1$, $\mu = 0.01$) between (i) the worst-case guarantee of the optimized method for $\frac{f(w_k)-f_\star}{\|w_0-w_\star\|^2}$ (in red; obtained from developments in Appendix D, and numerical examples in Appendix E.1); (ii) a lower bound on the oracle complexity for this setup (in blue; presented in [Drori and Taylor, 2021, Corollary 3]), which corresponds to $\frac{f(w_k)-f_\star}{\|w_0-w_\star\|^2} \geq \mu\frac{2-\sqrt{q}}{1+\sqrt{q}}\left(1-\sqrt{q}\right)^{2k}$; (iii) the triple momentum method [Van Scoy et al., 2018] (cyan); (iv) Nesterov's fast gradient method (defined in [Nesterov, 2004, Section 2.2, "Constant Step Scheme, II"]; FGM, green), and (v) the method generated by the subspace-search elimination procedure (SSEP) from [Drori and Taylor, 2020] (dashed, black).

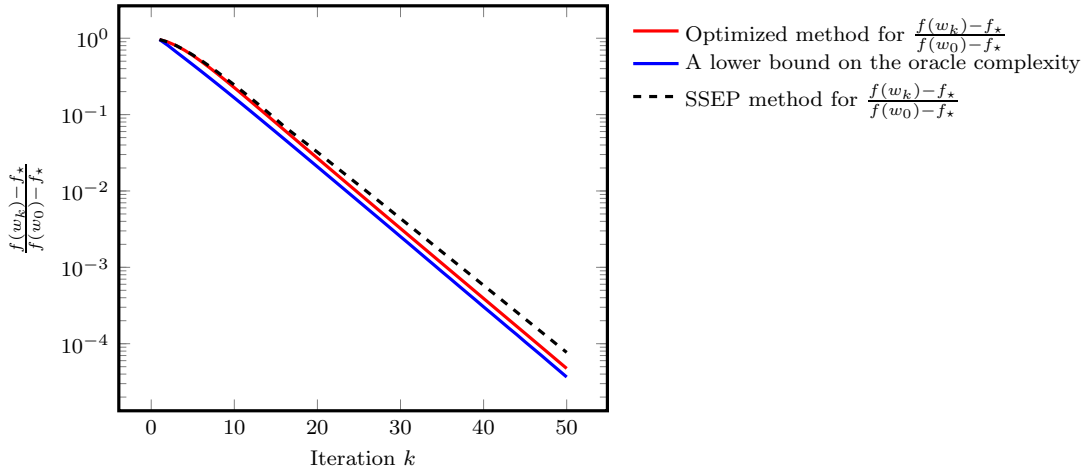

Fig. 2 Numerical comparison (for $L = 1$, $\mu = 0.01$) between (i) the worst-case guarantee of the optimized method for $\frac{f(w_k)-f_\star}{f(w_0)-f_\star}$ (in red; obtained from developments in Appendix D, and numerical examples in Appendix E.2); (ii) a lower bound on the oracle complexity for this setup (in blue; computed numerically using the procedure from [Drori and Taylor, 2021]); and (iii) a method generated by the subspace-search elimination procedure (SSEP) from [Drori and Taylor, 2020] (dashed, black).

E.2 Optimized methods for $(f(w_N) - f_\star)/(f(w_0) - f_\star)$

As in the previous section, the technique can be adapted for the criterion $(f(w_N) - f_\star)/(f(x_0) - f_\star)$, see Appendix D for details. The following step sizes were obtained by setting $L = 1$ and $\mu = .1$ and solving the resulting optimization problem from different values of N.

- For a single iteration, $N = 1$, we obtain a guarantee $\frac{f(w_1)-f_\star}{f(w_0)-f_\star} \leq 0.6694$ with the corresponding step size

$$[h^\star_{i,j}] = \begin{bmatrix} 1.8182 \end{bmatrix},$$

  which matches the known optimal step size $2/(L + \mu)$ for this setup [De Klerk et al., 2017, Theorem 4.2].
- For $N = 2$, we obtain $\frac{f(w_2)-f_\star}{f(w_0)-f_\star} \leq 0.3554$ with

$$[h^\star_{i,j}] = \begin{bmatrix} 2.0095 & \\ 0.4229 & 2.0095 \end{bmatrix}.$$

- For $N = 3$, we obtain $\frac{f(w_3)-f_\star}{f(w_0)-f_\star} \leq 0.1698$ with

$$[h^\star_{i,j}] = \begin{bmatrix} 1.9470 & & \\ 0.4599 & 2.2406 & \\ 0.1705 & 0.4599 & 1.9470 \end{bmatrix}.$$

- For $N = 4$, we obtain $\frac{f(w_4)-f_\star}{f(w_0)-f_\star} \leq 0.0789$ with

$$[h^\star_{i,j}] = \begin{bmatrix} 1.9187 & & & \\ 0.4098 & 2.1746 & & \\ 0.1796 & 0.5147 & 2.1746 & \\ 0.0627 & 0.1796 & 0.4098 & 1.9187 \end{bmatrix}.$$

- Finally, for $N = 5$, we reach $\frac{f(w_5)-f_\star}{f(w_0)-f_\star} \leq 0.0365$ with

$$[h^\star_{i,j}] = \begin{bmatrix} 1.9060 & & & & \\ 0.3879 & 2.1439 & & & \\ 0.1585 & 0.4673 & 2.1227 & & \\ 0.0660 & 0.1945 & 0.4673 & 2.1439 & \\ 0.0224 & 0.0660 & 0.1585 & 0.3879 & 1.9060 \end{bmatrix}.$$

Note that the resulting method is again apparently less practical than ITEM, as step sizes also critically depend on the horizon N; for example, observe again that the value of $h^\star_{1,0}$ depends on N. Interestingly, one can observe that the corresponding step sizes are symmetric, and that the worst-case guarantees seem to behave slightly better than in the distance problem $\|w_N - w_\star\|^2/\|w_0 - w_\star\|^2$, although their asymptotic rate has to be the same, due to the properties of strongly-convex functions. Figure 2 illustrates the worst-case guarantees of the corresponding method for larger numbers of iterations, and compares it to the lower bound.