RELATIVELY SMOOTH CONVEX OPTIMIZATION BY FIRST-ORDER METHODS, AND APPLICATIONS[*]

HAIHAO LU[†], ROBERT M. FREUND[‡], AND YURII NESTEROV[§]

Abstract. The usual approach to developing and analyzing first-order methods for smooth convex optimization assumes that the gradient of the objective function is uniformly smooth with some Lipschitz constant L. However, in many settings the differentiable convex function $f(\cdot)$ is not uniformly smooth—for example, in D-optimal design where $f(x) := -\ln \det(HXH^T)$ and $X := \mathbf{D}\mathrm{iag}(x)$, or even the univariate setting with $f(x) := -\ln(x) + x^2$. In this paper we develop a notion of "relative smoothness" and relative strong convexity that is determined relative to a user-specified "reference function" $h(\cdot)$ (that should be computationally tractable for algorithms), and we show that many differentiable convex functions are relatively smooth with respect to a correspondingly fairly simple reference function $h(\cdot)$. We extend two standard algorithms—the primal gradient scheme and the dual averaging scheme—to our new setting, with associated computational guarantees. We apply our new approach to develop a new first-order method for the D-optimal design problem, with associated computational complexity analysis. Some of our results have a certain overlap with the recent work [H. H. Bauschke, J. Bolte, and M. Teboulle, *Math. Oper. Res.*, 42 (2017), pp. 330–348].

1. Introduction, definition of "relative smoothness," and basic properties.

1.1. Traditional setup for smooth first-order methods.
Our optimization problem of interest is

(1)
$$P: \quad f^* := \quad \begin{aligned} &\text{minimum}_x \quad f(x) \\ &\text{s.t.} \quad x \in Q \, , \end{aligned}$$

where $Q \subseteq \mathbb{E}$ is a closed convex set in the finite-dimensional vector space $\mathbb{E}$ with inner product $\langle \cdot \, , \, \cdot \rangle$, and $f(\cdot) : Q \to \mathbb{R}$ is a differentiable convex function.

There are by now very many first-order methods for tackling the optimization problem (1); see, for example, [15], [22], [19]; virtually all such methods are designed to solve (1) when the gradient of $f(\cdot)$ satisfies a uniform Lipschitz condition on Q, namely, there exists a constant $L_f < \infty$ for which

(2)
$$\|\nabla f(x) - \nabla f(y)\|_* \leq L_f \|x - y\| \quad \text{for all } x, y \in Q \, ,$$

where $\| \cdot \|$ is a given norm on $\mathbb{E}$ and $\| \cdot \|_*$ denotes the usual dual norm. For example, consider the standard gradient descent scheme, which presumes the norm in (2) is

[†]MIT Department of Mathematics, Cambridge, MA 02139 (haihao@mit.edu).

[‡]MIT Sloan School of Management, Cambridge, MA 02139 (rfreund@mit.edu).

[§]Department of Mathematical Engineering, Université Catholique de Louvain, Louvain-La-Neuve 1348, Belgium (yurii.nesterov@uclouvain.be).

Euclidean, and uses the following update:

$$(3) \qquad x^{i+1} \leftarrow \arg\min_{x \in Q} \left\{ f(x^i) + \langle \nabla f(x^i), x - x^i \rangle + \frac{L_f}{2} \|x - x^i\|_2^2 \right\} \ .$$

One can prove for the standard gradient descent scheme that after k iterations it holds for any $x \in Q$ that

$$(4) \qquad f(x^k) - f(x) \ \leq \ \frac{L_f \|x - x^0\|_2^2}{2k} \ ,$$

which is an $O(1/k)$ sublinear rate of convergence [15], [19]. Furthermore, if $f(\cdot)$ is also uniformly μ_f-strongly convex for some $\mu_f > 0$, namely,

$$(5) \qquad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu_f}{2} \|y - x\|_2^2 \quad \text{for all } x, y \in Q \ ,$$

then one can prove linear convergence for the gradient descent scheme (see [15], [19]), i.e., for any $x \in Q$ we have that

$$(6) \qquad f(x^k) - f(x) \ \leq \ \frac{L_f}{2} \left(1 - \frac{2\mu_f}{L_f + \mu_f} \right)^k \|x - x^0\|_2^2 \ .$$

More general versions of first-order methods are not restricted to the Euclidean ($\|\cdot\|_2$) norm, and use a differentiable "prox function" $h(\cdot)$, which is a 1-strongly convex function on Q, to define a Bregman distance

$$(7) \qquad D_h(y, x) := h(y) - h(x) - \langle \nabla h(x), y - x \rangle \quad \text{for all } x, y \in Q$$

which as a result satisfies

$$D_h(y, x) \geq \tfrac{1}{2} \|y - x\|^2 \ .$$

The standard primal gradient scheme (with Bregman distance) (see [22]), has the following update formula

$$(8) \qquad x^{i+1} \leftarrow \arg\min_{x \in Q} \left\{ f(x^i) + \langle \nabla f(x^i), x - x^i \rangle + L_f D_h(x, x^i) \right\} \ .$$

Notice in (8) by construction that the update requires the capability of solving instances of a subproblem of the general form

$$(9) \qquad x_{\text{new}} \leftarrow \arg\min_{x \in Q} \{ \langle c, x \rangle + h(x) \}$$

for suitable iteration-specific values of c; indeed, (8) is an instance of the subproblem (9) with $c = \frac{1}{L_f} \nabla f(x^i) - \nabla h(x^i)$ at iteration i. It is especially important to note that the primal gradient scheme is somewhat meaningless whenever we do not have the capability to efficiently solve (9), a point which we will return to later on. In a typical design and implementation of a first-order method for solving (1), one attempts to specify the norm $\|\cdot\|$ and the strongly convex prox function $h(\cdot)$ in consideration of the shape of the feasible domain Q while also ensuring that the subproblem (9) is efficiently solvable.

Regarding computational guarantees, one can prove for the primal gradient scheme that after k iterations it holds for any $x \in Q$ that

$$(10) \qquad f(x^k) - f(x) \ \leq \ \frac{L_f D_h(x, x^0)}{k} \ ,$$

which is an exact generalization of (4); see [22], [14].

We emphasize that standard first-order methods as stated above for solving (1) require that $f(\cdot)$ be uniformly smooth on Q, that is, that there is a finite value of the Lipschitz constant L_f as defined in (2), in order to ensure associated computational guarantees. However, there are many differentiable convex functions in practice that do not satisfy a uniform smoothness condition. Consider $f(x) := -\ln\det(HXH^T)$ with $X := \mathbf{D}\mathrm{iag}(x)$ in D-optimal design on the feasible set $Q = \{x \in \mathbb{R}^n : \langle e, x \rangle = 1, \ x \geq 0\}$, or $f(x) = |x|^3$ or $f(x) = x^4$ on the feasible set $Q = \mathbb{R}$, or $f(x) = -\ln(x) + x^2$ on $Q = \mathbb{R}_{++}$. Of course, if the algorithm iterates have monotone decreasing objective function values (which is provably the case for most smooth first-order methods), it then is sufficient just to ensure that $f(\cdot)$ is smooth on some level set of $f(\cdot)$. Nevertheless, even in this case the constant L_f may be huge. For instance, let $f(x) = -\ln(x) + x^2$ on $Q = \mathbb{R}_{++}$, and consider the level set $\{x : f(x) \leq 10\}$. Then one still has $L_f \approx \exp^{20}$ on this level set, which is not reasonable for practical use.

Notice that unlike quadratic functions, the second-order terms of the functions in the above examples vary dramatically on Q—and especially as $x \to \partial Q$ (or as x goes to infinity in Q). It therefore becomes unreasonable to use a uniform bound of the form L_f to upper-bound second-order information.

Motivated by the above drawbacks in standard first-order methods, we develop a notion of "relative smoothness" and relative strong convexity, relative to a given "reference function" $h(\cdot)$ and which does not require the specification of any particular norm—and indeed $h(\cdot)$ need not be either strictly or strongly convex. Armed with relative smoothness and relative strong convexity, we demonstrate the capability to solve a more general class of differentiable convex optimization problems (without uniform Lipschitz continuous gradients), and we also demonstrate linear convergence results for both a primal gradient scheme and a dual averaging scheme when the function is both relatively smooth and relatively strongly convex.

There is a certain overlap of ideas and results herein with the paper [6] by Bauschke, Bolte, and Teboulle. For starters, the relative smoothness condition definition in the present paper in Definition 1.1 is equivalent to the (LC) condition in [6] except that [6] also requires the reference function $h(\cdot)$ to be essentially smooth and strictly convex, which we do not need in this paper. The main developments in [6] are based on generalizing a key descent lemma and applying this generalization to tackle (additive) composite optimization problems using the primal gradient scheme (called the NoLips Algorithm in [6]) with associated complexity analysis involving a symmetry measure of the Bregman distance $D_h(\cdot, \cdot)$. These results are then illustrated in the application of composite optimization to Poisson inverse problems. While the NoLips Algorithm in [6] is structurally the same as Algorithm 1 herein, they are both instantiations of the standard primal gradient scheme; however, as will be seen in section 3 here, we do not need any symmetry measure in constructing step sizes or in the complexity analysis. The paper [28] by Zhou, Liang, and Shen also tackles composite optimization using the standard primal gradient scheme which therein is called PGA-$\mathcal{B}$, with a focus on demonstrating equivalence of proximal gradient and proximal point methods more broadly. Here we develop measures of relative smoothness and also relative strong convexity, which can improve the computational guarantees of the primal gradient scheme; see Theorem 3.1. We further present computational guarantees for the dual averaging scheme [17] in Theorem 3.2. In section 2 we show that many differentiable convex functions are relatively smooth with respect to a correspondingly fairly simple reference function $h(\cdot)$ that is easy to construct and for which algorithmic computations can efficiently be performed. In section 4 we apply our approach to develop a new first-order method for the D-optimal design problem,

with associated computational complexity analysis. Throughout the current paper, we compare and clarify similarities and differences between our work and [6] in the context of the specific contributions as they arise.

1.2. Relative smoothness and relative strong convexity. Let $h(\cdot)$ be any given differentiable convex function (it need not be strongly nor even strictly convex) defined on Q. We will henceforth refer to $h(\cdot)$ as the reference function. We define relative smoothness and "relative strong convexity" of $f(\cdot)$ relative to $h(\cdot)$ using the Bregman distance (7) associated with $h(\cdot)$ as follows.

DEFINITION 1.1. $f(\cdot)$ is L-smooth relative to $h(\cdot)$ on Q if for any $x, y \in \text{int } Q$, there is a scalar L for which

$$(11) \qquad f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + L D_h(y, x) \ .$$

DEFINITION 1.2. $f(\cdot)$ is μ-strongly convex relative to $h(\cdot)$ on Q if for any $x, y \in \text{int } Q$, there is a scalar $\mu \geq 0$ for which

$$(12) \qquad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \mu D_h(y, x) \ .$$

(Here and elsewhere int Q denotes the interior of Q. In cases where Q has no interior, one can instead use the relative interior of Q.) Note that relative smoothness and relative strong convexity of $f(\cdot)$ are defined relative to the reference function $h(\cdot)$ directly; no norm is involved in the definitions, so that smoothness/strong convexity does not depend on any norm. Furthermore, $h(\cdot)$ is not presumed to have any special properties by itself such as strict or (traditional) strong convexity; rather the key structural properties involve how $f(\cdot)$ behaves relative to $h(\cdot)$. The definition of relative smoothness above is equivalent to the (LC) condition in [6], but [6] requires the reference function to be essentially smooth and strictly convex, which we do not need.

The following proposition presents equivalent definitions of relative smoothness and relative strong convexity. In the case when both $f(\cdot)$ and $h(\cdot)$ are twice differentiable, parts (a-iii) and (b-iii) of the proposition demonstrate that the above definitions are equivalent to

$$\mu \nabla^2 h(x) \preceq \nabla^2 f(x) \preceq L \nabla^2 h(x) \ \text{ for all } x \in \text{int } Q \ ,$$

which is an intuitively simple condition on the Hessian matrices of the two functions.

PROPOSITION 1.1. *The following conditions are equivalent:*
  (a-i) *$f(\cdot)$ is L-smooth relative to $h(\cdot)$;*
  (a-ii) *$Lh(\cdot) - f(\cdot)$ is a convex function on Q;*
  (a-iii) *under twice differentiability $\nabla^2 f(x) \preceq L \nabla^2 h(x)$ for any $x \in \text{int } Q$;*
  (a-iv) *$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L \langle \nabla h(x) - \nabla h(y), x - y \rangle$ for all $x, y \in \text{int } Q$.*
The following conditions are equivalent:
  (b-i) *$f(\cdot)$ is μ-strongly convex relative to $h(\cdot)$;*
  (b-ii) *$f(\cdot) - \mu h(\cdot)$ is a convex function on Q;*
  (b-iii) *under twice differentiability $\nabla^2 f(x) \succeq \mu \nabla^2 h(x)$ for any $x \in \text{int } Q$;*
  (b-iv) *$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \langle \nabla h(x) - \nabla h(y), x - y \rangle$ for all $x, y \in \text{int } Q$.*

The first part of Proposition 1.1 is almost equivalent to Proposition 1 of [6].

Proof. For $x \in Q$ define $\phi(x) := Lh(x) - f(x)$. Using (11) and (7) it follows that (a-i) holds if and only if $\phi(x) \geq \phi(y) + \langle \nabla \phi(y), x - y \rangle$ for all $x, y \in Q$, which is

equivalent to the convexity of $\phi(\cdot) = Lh(\cdot) - f(\cdot)$ from Theorem 2.1.2 of [15], thus showing that (a-i) $\Leftrightarrow$ (a-ii). It follows from Theorem 2.1.3 of [15] applied to $\phi(\cdot)$ that $\phi(\cdot)$ is convex if and only if $\langle \nabla\phi(x) - \nabla\phi(y), x - y \rangle \geq 0$ for all $x, y \in Q$, which shows that (a-ii) $\Leftrightarrow$ (a-iv). If $f(\cdot)$ and $h(\cdot)$ are twice differentiable, then it follows from Theorem 2.1.4 of [15] that (a-ii) $\Leftrightarrow$ (a-iii).

Similar proofs can be applied for part (b). $\qquad\square$

For notational convenience, let us denote by $f(\cdot) \preceq h(\cdot)$ that $h(\cdot) - f(\cdot)$ is a convex function, whereby this also means $f(\cdot)$ is 1-smooth with respect to $h(\cdot)$ from Proposition 1.1. Similarly $f(\cdot) \succeq h(\cdot)$ means $f(\cdot) - h(\cdot)$ is a convex function and so $f(\cdot)$ is 1-strongly convex with respect to $h(\cdot)$. (In the case when both $f(\cdot)$ and $h(\cdot)$ are twice differentiable, the relation "$\cdot \succeq \cdot$" on two functions is consistent with the Löwner partial order on the Hessians of these two functions from Propositon 1.1.) Then the condition that $f(\cdot)$ is L-smooth with respect to $h(\cdot)$ is equivalent to $f(\cdot) \preceq Lh(\cdot)$; similarly the condition that $f(\cdot)$ is μ-strongly convex with respect to $h(\cdot)$ is equivalent to $f(\cdot) \succeq \mu h(\cdot)$. In addition, relative smoothness and relative strong convexity are each transitive, so that $f(\cdot) \preceq g(\cdot)$ and $g(\cdot) \preceq h(\cdot)$ implies that $f(\cdot) \preceq h(\cdot)$.

We can also work with sums and linear transformations of relatively smooth and/or relatively strongly convex functions, as the next proposition states.

PROPOSITION 1.2.
1. If $f_1(\cdot) \preceq L_1h(\cdot)$ and $f_2(\cdot) \preceq L_2h_2(\cdot)$, then for all $\alpha, \beta \geq 0$ it holds that $f(\cdot) := \alpha f_1(\cdot) + \beta f_2(\cdot) \preceq h(\cdot) := \alpha L_1h_1(\cdot) + \beta L_2h_2(\cdot)$.
2. If $f_1(\cdot) \succeq \mu_1h_1(\cdot)$ and $f_2(\cdot) \succeq \mu_2h_2(\cdot)$, then for all $\alpha, \beta \geq 0$ it holds that $f(\cdot) := \alpha f_1(\cdot) + \beta f_2(\cdot) \succeq h(\cdot) := \alpha\mu_1h_1(\cdot) + \beta\mu_2h_2(\cdot)$.
3. If $f(\cdot) \preceq h(\cdot)$, and A is a linear transformation of appropriate dimension, then $\phi_f(x) := f(Ax) \preceq \phi_h(x) := h(Ax)$.
4. If $f(\cdot) \succeq h(\cdot)$, and A is a linear transformation of appropriate dimension, then $\phi_f(x) := f(Ax) \succeq \phi_h(x) := h(Ax)$.

Proof. The proofs of the first two arguments follow directly from the definitions of relative smoothness and relative strong convexity in Definitions 1.1 and 1.2. The proofs of the last two arguments follow from the equivalent definition (a-iv) and (b-iv) in Proposition 1.1. $\qquad\square$

1.3. Constructive algorithmic setup. Let us now discuss criteria for choosing the reference function $h(\cdot)$ in the context of computational schemes for solving the optimization problem (1). To be concrete, consider a simple primal gradient scheme as shown in Algorithm 1. Note that this scheme is essentially as described in the update formula (8), except that the uniform smoothness constant L_f is replaced by the relative smoothness parameter L of $f(\cdot)$ with respect to the reference function $h(\cdot)$ as defined in Definition 1.1, and the only formal requirement for $h(\cdot)$ is that the pair $(f(\cdot), h(\cdot))$ must satisfy the conditions of Definition 1.1.

Algorithm 1 Primal Gradient Scheme with reference function $h(\cdot)$.

Initialize. Initialize with $x^0 \in Q$. Let L, $h(\cdot)$ satisfying Definition 1.1 be given.
At iteration i :
Perform Updates. Compute $\nabla f(x^i)$,

$$x^{i+1} \leftarrow \arg\min_{x \in Q}\{f(x^i) + \langle \nabla f(x^i), x - x^i \rangle + LD_h(x, x^i)\} \ .$$

In order to efficiently execute the update step in Algorithm 1 we also require of $h(\cdot)$ that the subproblem (9) is efficiently solvable for any given c. In summary, to solve the optimization problem (1) using Algorithm 1, we need to specify a reference function $h(\cdot)$ that has the following two properties:

(i) $f(\cdot)$ is L-smooth relative to $h(\cdot)$ on Q, and

(ii) the subproblem (9) always has a solution, and the solution is efficiently computable.

In section 2 we will see how this can be done for several useful classes of problems that are not otherwise solvable by traditional first-order methods that require uniform Lipschitz continuity of the gradient. In section 3 we analyze the computational guarantees associated with the primal gradient scheme (Algorithm 1) as well as a dual averaging scheme. In section 4, we apply the computational guarantees of section 3 to the D-optimal design problem.

Notation. For a vector x, $X = \mathbf{D}\mathrm{iag}(x)$ denotes the diagonal matrix with the coefficients of x along the diagonal. For a symmetric matrix A, $\mathrm{diag}(A)$ denotes the vector of the diagonal coefficients of A, and $\mathbf{M}\mathrm{diag}(A)$ denotes the diagonal matrix whose diagonal coefficients correspond to the diagonal coefficients of A. Unless otherwise specified, the norm of a matrix is the operator norm using ℓ_2 norms. The ℓ_p norm of a vector x is denoted by $\|x\|_p$. For symmetric matrices, "$\succeq$" denotes the Löwner partial order. In a mild double use of notation, $f(\cdot) \succeq h(\cdot)$ denotes $f(\cdot) - h(\cdot)$ is a convex function, and the appropriate meaning of $\succeq$ will be obvious in context. Let e denote the vector of 1's whose dimension is dictated by context. Let $\Delta_n := \{x \in \mathbb{R}^n : \langle e, x \rangle = 1, \ x \geq 0\}$ denote the standard unit simplex in $\mathbb{R}^n$. Given two matrices A and B of the same order, let $A \circ B$ denote the Hadamard (i.e., componentwise) product of A and B; see, for example, Anstreicher [2]. Let exp denote the base of the natural logarithm.

2. Examples of relatively smooth optimization problems. Here we show several classes of optimization problems (1) for which one can easily construct a reference function $h(\cdot)$ with the two properties mentioned above, namely, (i) $f(\cdot)$ is L-smooth relative to $h(\cdot)$ for an easily determined value L, and (ii) the subproblem (9) is efficiently solvable.

2.1. Optimization over $\mathbb{R}^n$ with $\|\nabla^2 f(x)\|$ growing as a polynomial in $\|x\|_2$. Suppose that $f(\cdot)$ is a twice-differentiable convex function on $Q := \mathbb{R}^n$ and let $\|\nabla^2 f(x)\|$ denote the operator norm of $\nabla^2 f(x)$ with respect to the ℓ_2-norm on $\mathbb{R}^n$. Suppose that $\|\nabla^2 f(x)\| \leq p_r(\|x\|_2)$ where $p_r(\alpha) = \sum_{i=0}^{r} a_i \alpha^i$ is an r-degree polynomial of α. Let

$$(13) \qquad\qquad h(x) := \tfrac{1}{r+2}\|x\|_2^{r+2} + \tfrac{1}{2}\|x\|_2^2 \ .$$

Then the following proposition states that $f(\cdot)$ is L-smooth relative to $h(\cdot)$ for an easily computable value L. This implies that no matter how fast the Hessian of $f(\cdot)$ grows as $\|x\|_2 \to \infty$, $f(\cdot)$ can still be smooth relative to the simple reference function $h(\cdot)$, even though $\nabla f(\cdot)$ need not exhibit uniform Lipschitz continuity.

PROPOSITION 2.1. *Suppose $f(\cdot)$ is twice differentiable and satisfies $\|\nabla^2 f(x)\| \leq p_r(\|x\|_2)$, where $p_r(\alpha)$ is an r-degree polynomial of α. Let L be such that $p_r(\alpha) \leq L(1 + \alpha^r)$ for $\alpha \geq 0$. Then $f(\cdot)$ is L-smooth relative to $h(x) = \frac{1}{r+2}\|x\|_2^{r+2} + \frac{1}{2}\|x\|_2^2$.*

Proof. It follows from elementary rules of differentiation that

$$\nabla^2 h(x) = (1 + \|x\|_2^r)I + (r+1)\|x\|_2^{r-2} x x^T \ \succeq \ (1 + \|x\|_2^r)I \ \succeq \ \tfrac{1}{L} p_r(\|x\|_2)I \ \succeq \ \tfrac{1}{L}\nabla^2 f(x) \ ,$$

and so $f(\cdot)$ is L-smooth relative to $h(\cdot)$ by part (iii) of Proposition 1.1. ☐

Utilizing the additivity property in Proposition 1.2 together with Proposition 2.1, one concludes that virtually every twice-differentiable convex function on $\mathbb{R}^n$ is L-smooth relative to some simple polynomial function of $\|x\|_2$.

Remark 2.1. Suppose $p_r(\alpha) = \sum_{i=0}^r a_i \alpha^i$. In Proposition 2.1, one simple way to set L is to use $L = \sum_{i=0}^r |a_i|$. Then

$$(14) \qquad p_r(\alpha) \leq \begin{cases} \sum_{i=0}^r |a_i| & \text{for } 0 \leq \alpha \leq 1 \ , \\ \sum_{i=0}^r |a_i| \alpha^r & \text{for } \alpha \geq 1 \ , \end{cases}$$

whereby $p_r(\alpha) \leq \max\{L, L\alpha^r\} \leq L(1 + \alpha^r)$ for $\alpha \geq 0$.

Solving the subproblem (9). Let us see how we can solve the subproblem (9) for this class of optimization problems. The subproblem (9) can be written as

$$(15) \qquad \min_{x \in \mathbb{R}^n} \ \langle c, x \rangle + \frac{1}{r+2} \|x\|_2^{r+2} + \frac{1}{2}\|x\|^2 \ ,$$

and the first-order optimality conditions are simply

$$c + (1 + \|x\|_2^r)x = 0 \ ,$$

whereby $x = -\theta c$ for some $\theta \geq 0$, and it remains to simply determine the value of the nonnegative scalar θ. If $c = 0$, then $x = 0$ satisfies the optimality conditions. For $c \neq 0$, notice from above that θ must satisfy

$$1 - \theta - \|c\|_2^r \cdot \theta^{r+1} = 0 \ ,$$

which is a univariate polynomial in θ with a unique positive root. For $r = 1, 2, 3$, this root can be computed in closed form. Otherwise, the root can be computed (up to machine precision) using any scalar root-finding method.

Remark 2.2. We can incorporate in problem (15) a simple set constraint $x \in Q$ provided that we can easily compute the Euclidean projection on Q. In the case when $h(\cdot)$ is a convex function of $\|x\|_2^2$, the subproblem (9) can be converted to a 1-dimensional convex optimization problem; see Appendix A.1 for details.

A more specific example. Let $f(x) := \frac{1}{4}\|Ax - b\|_4^4 + \frac{1}{2}\|Cx - d\|_2^2$. Then $\nabla^2 f(x) = 3A^T D^2(x) A + C^T C$, where $D(x) = \mathbf{D}\text{iag}(Ax - b)$. Let us show that $f(x)$ is L-smooth relative to

$$h(x) := \frac{1}{4}\|x\|_2^4 + \frac{1}{2}\|x\|_2^2$$

on $Q = \mathbb{R}^n$ for $L = 3\|A\|^4 + 6\|A\|^3\|b\|_2 + 3\|A\|^2\|b\|_2^2 + \|C\|^2$. To see this, notice first that

$$\|\nabla^2 f(x)\| \leq 3\|A\|^2(\|b\|_2 + \|A\|\|x\|_2)^2 + \|C\|^2$$

$$= \left(3\|A\|^2\|b\|_2^2 + \|C\|^2\right) + 6\|A\|^3\|b\|_2\|x\|_2 + 3\|A\|^4\|x\|_2^2 \ ,$$

which is a 2-degree polynomial in $\|x\|_2$ with coefficients $a_0 = 3\|A\|^2\|b\|_2^2 + \|C\|^2$, $a_1 = 6\|A\|^3\|b\|_2$, and $a_2 = 3\|A\|^4$. Therefore, following Remark 2.1, it suffices to set

$$L = \sum_{i=0}^2 a_i = 3\|A\|^4 + 6\|A\|^3\|b\|_2 + 3\|A\|^2\|b\|_2^2 + \|C\|^2 \ .$$

An example with non-Lipschitz μ-strong convexity. Let $f(x) := \frac{1}{4}\|Ex\|_2^4 + \frac{1}{4}\|Ax - b\|_4^4 + \frac{1}{2}\|Cx - d\|_2^2$, and let σ_E and σ_C denote the smallest singular values of E and C, respectively, and let us suppose that $\sigma_E > 0$ and $\sigma_C > 0$. Then $\nabla^2 f(x) = \|Ex\|_2^2 E^T E + 2E^T Exx^T E^T E + 3A^T D^2(x)A + C^T C$, where $D(x) = \mathbf{D}\text{iag}(Ax - b)$. Let us show that $f(x)$ is L-smooth and μ-strongly convex relative to

$$h(x) := \tfrac{1}{4}\|x\|_2^4 + \tfrac{1}{2}\|x\|_2^2$$

on $Q = \mathbb{R}^n$ for $L = 3\|E\|^4 + 3\|A\|^4 + 6\|A\|^3\|b\|_2 + 3\|A\|^2\|b\|_2^2 + \|C\|^2$ and $\mu = \min\{\frac{\sigma_E^4}{3}, \sigma_C^2\}$. Similarly to what we have above,

$$\|\nabla^2 f(x)\| \le \|E\|^4\|x\|_2^2 + 2\|E\|^4\|x\|_2^2 + 3\|A\|^2(\|b\|_2 + \|A\|\|x\|_2)^2 + \|C\|^2$$
$$= \left(3\|A\|^2\|b\|_2^2 + \|C\|^2\right) + 6\|A\|^3\|b\|_2\|x\|_2 + \left(3\|E\|^4 + 3\|A\|^4\right)\|x\|_2^2 ,$$

which is a 2-degree polynomial in $\|x\|_2$ with coefficients $a_0 = 3\|A\|^2\|b\|_2^2 + \|C\|^2$, $a_1 = 6\|A\|^3\|b\|_2$, and $a_2 = 3\|E\|^4 + 3\|A\|^4$. Therefore following Remark 2.1 it suffices to set

$$L = \sum_{i=0}^{2} a_i = 3\|E\|^4 + 3\|A\|^4 + 6\|A\|^3\|b\|_2 + 3\|A\|^2\|b\|_2^2 + \|C\|^2 .$$

On the other hand,

$$\nabla^2 f(x) \succeq \|Ex\|_2^2 E^T E + C^T C \succeq \sigma_E^4\|x\|_2^2 I + \sigma_C^2 I \succeq \mu\left(1 + 3\|x\|_2^2\right)I$$
$$\succeq \mu\left((1 + \|x\|_2^2)I + 2xx^T\right) = \mu\nabla^2 h(x)$$

(where the last matrix inequality follows since $\|x\|_2^2 I \succeq xx^T$), and thus $f(x)$ is μ-strongly convex relative to $h(x)$.

Remark 2.3. In place of the simple reference function $h(\cdot)$ in (13) one can instead consider a "recentered" version of the form

$$h(x) = h_{x^c}(x) := \tfrac{1}{r+2}\|x - x^c\|_2^{r+2} + \tfrac{1}{2}\|x - x^c\|_2^2 ,$$

where the "center" value x^c is suitably chosen to align $f(\cdot)$ with $h(\cdot)$ and possibly attain better values of L and μ. Note that introducing the given center value x^c does not increase the difficulty of solving the subproblem (9). We illustrate this idea with a simple univariate example. Suppose that our objective function is $f(x) = x^4 - 4x^3 + 7x^2 - 5x + 3$. From the results in section 2.1 we know we can use the reference function $h_1(x) := \frac{1}{4}x^4 + \frac{1}{2}x^2$. We can also translate x by the center point $x^c := 1$ and use the reference function $h_2(x) := \frac{1}{4}(x-1)^4 + \frac{1}{2}(x-1)^2$. Straightforward calculation yields values of $L = L_1 = 9 + \sqrt{73} \approx 17.5440$ for $h_1(\cdot)$ and $L = L_2 = 4$ for $h_2(\cdot)$, whereby $h_2(\cdot)$ yields a better value of L than $h_1(\cdot)$ for this example.

2.2. D-optimal design problem. Given a matrix $H \in \mathbb{R}^{m \times n}$ of rank m, where $n \ge m + 1$, the D-optimal design problem is

$$D: \quad f^* = \min_x \quad f(x) := -\ln\det\left(HXH^T\right)$$
(16)
$$\text{s.t.} \quad \langle e, x \rangle = 1 ,$$
$$x \ge 0 ,$$

where recall $X := \mathbf{D}\text{iag}(x)$. In statistics, the D-optimal design problem corresponds to maximizing the determinant of the Fisher information matrix $\mathbb{E}(hh^T)$; see [12], [4]. And in computational geometry, D-optimal design arises as a Lagrangian dual problem of the minimum volume covering ellipsoid problem, which dates back at least

60 years to [9]; see Todd [21] for a modern treatment. Indeed, (16) is useful in a variety of different application areas, for example, computational statistics [8] and data mining [13]. In terms of algorithms for solving (16), Khachiyan and Todd [11] proposed a theory-oriented scheme based on interior-point methods; see also Zhang [27] as well as [20] for more practical treatments using interior-point methods. Khachiyan [10] later proposed and analyzed a first-order method (equivalent to the Frank–Wolfe method) to solve (16), which led to other works along this line including Yildirim [25] and Ahipasaoglu, Sun, and Todd [1]. The complexity analysis in these papers is very specialized for the D-optimal design problem. In contrast, we will show how the primal gradient scheme (Algorithm 1) can be applied to the D-optimal design problem; furthermore, in section 4 we will apply the complexity analysis of section 3 for the primal gradient scheme to the setup of D-optimal design, along with a comparison of our convergence guarantees with the guarantees from prior literature.

Notice that (16) is an instance of (1) with $Q = \Delta_n := \{x \in \mathbb{R}^n : \langle e, x \rangle = 1, \ x \geq 0\}$. Although strictly speaking, $f(\cdot)$ in (16) is not defined everywhere on the relative boundary of Q and hence does not have gradients or Hessians everywhere on the relative boundary of Q, this will not be of concern. For $f(\cdot)$ in (16) let us choose the reference function $h(\cdot)$ to be the logarithmic barrier function, namely,

$$h(x) := -\sum_{j=1}^{n} \ln(x_j) \ ,$$

which is defined on the positive orthant $\mathbb{R}^n_{++}$. The following proposition states that $f(\cdot)$ is 1-smooth relative to $h(\cdot)$.

PROPOSITION 2.2. *Suppose $f(x) = -\ln\det(HXH^T)$, where $X = \mathbf{D}\text{iag}(x)$. Then $f(\cdot)$ is 1-smooth relative to $h(x) = -\sum_{j=1}^{n}\ln(x_j)$ on $\mathbb{R}^n_{++}$.*

Proof. The gradient of $f(\cdot)$ is $\nabla f(x) = \text{diag}(-C)$ and the Hessian of $f(\cdot)$ is $\nabla^2 f(x) = C \circ C$, where $C := H^T(HXH^T)^{-1}H$. Let $U = HX^{\frac{1}{2}}$; then $U^T(UU^T)^{-1}U \preceq I$ since the left side of this matrix inequality is a projection operator, whereby $X^{\frac{1}{2}}H^T(HXH^T)^{-1}HX^{\frac{1}{2}} \preceq I$. Multiplying this matrix inequality on the left and right by $X^{-\frac{1}{2}}$ then shows that $C \preceq X^{-1}$. Therefore,

$$(17) \qquad \nabla^2 f(x) = C \circ C \preceq C \circ X^{-1} \preceq X^{-1} \circ X^{-1} = X^{-2} = \nabla^2 h(x) \ ,$$

where the first and the second matrix inequality above each follows from the fact that $C \preceq X^{-1}$ and the Hadamard product of two symmetric positive semidefinite matrices is also a symmetric positive semidefinite matrix. The result then follows using property (a-iii) of Proposition 1.1. □

Solving the subproblem (9). Let us see how we can solve the subproblem (9) for Q and $h(\cdot)$ given above. The subproblem (9) can be written as

$$\min_{x \in \Delta_n} \ \langle c, x \rangle - \sum_{j=1}^{n} \ln(x_j) \ ,$$

and the first-order optimality conditions are simply

$$x > 0, \ \langle e, x \rangle = 1, \ \text{and} \ c - X^{-1}e = -\theta e$$

for some scalar multiplier θ. Given θ, it then follows that $x_j = 1/(c_j + \theta)$ for $j = 1, \ldots, n$, and it remains to simply determine the value of the scalar θ. Now notice

that θ must satisfy

$$(18) \qquad d(\theta) := \sum_{j=1}^{n} \frac{1}{c_j + \theta} \;\; - 1 \;=\; 0$$

for some θ in the interval $\mathcal{F} := (-\min_j\{c_j\}, \infty)$. Notice that $d(\cdot)$ is strictly decreasing on $\mathcal{F}$, and $d(\theta) \to +\infty$ as $\theta \searrow -\min_j\{c_j\}$ and $d(\theta) \to -1$ as $\theta \to \infty$, whereby (18) has a unique solution in $\mathcal{F}$. Furthermore, as suggested by results in Ye [24] or [7], one can use Newton's method (or any other suitable scalar solution-finding method) to efficiently compute the solution of (18) (up to machine precision) on the interval $\mathcal{F}$.

2.3. Generalized volumetric function optimization. For a given integer parameter $p > 0$, let us also study optimization on the simplex of the following generalization of the volumetric barrier function:

$$(19) \qquad \begin{aligned} \min_x \quad & f_p(x) = \ln\det\left(HX^{-p}H^T\right) \\ \text{s.t.} \quad & \langle e, x \rangle = 1 \;, \\ & x \geq 0 \;, \end{aligned}$$

where the integer p is the parameter of the volumetric function $f_p(\cdot)$, and $H \in \mathbb{R}^{m \times n}$ is a rank-m matrix, where $n \geq m + 1$. Here the feasible region is $Q = \Delta_n$. Note that $f_p(\cdot)$ is a convex function when $p \geq 0$ (and $f_p(\cdot)$ is a concave function when $p = -1$).

Similarly to the D-optimal design problem, $f_p(\cdot)$ is not defined everywhere on the boundary of $\mathbb{R}^n_+$, but this will not be a concern. The reference function $h(\cdot)$ we choose is the logarithmic barrier function, namely,

$$h(x) := -\sum_{j=1}^{n} \ln(x_j) \;,$$

which is defined on $\mathbb{R}^n_{++}$. The following proposition states that $f_p(\cdot)$ is $p(p+1)$-smooth relative to $h(\cdot)$.

PROPOSITION 2.3. $f_p(\cdot)$ is $p(p+1)$-smooth relative to $h(x) = -\sum_{j=1}^{n} \ln(x_j)$ on $\mathbb{R}^n_{++}$.

Proof. By elementary calculus, the gradient of $f_p(\cdot)$ is

$$\nabla f_p(x) = -p \cdot \operatorname{diag}\left(X^{-1/2-p/2}CX^{-1/2-p/2}\right) \;,$$

and the Hessian of $f_p(\cdot)$ is

$$\nabla^2 f_p(x) = p(p+1)\mathbf{M}\mathrm{diag}(X^{-1-p/2}CX^{-1-p/2}) - p^2 X^{-1-p/2}(C \circ C)X^{-1-p/2} \;,$$

where $C := H^T(HX^{-p}H^T)^{-1}H$, and $\mathbf{M}\mathrm{diag}(M)$ denotes the diagonal matrix whose entries are the diagonal components of the matrix M. Let $U = HX^{-p/2}$; then $U^T(UU^T)^{-1}U \preceq I$ since the left side of this matrix inequality is a projection operator. Therefore each diagonal component of $U^T(UU^T)^{-1}U$ does not exceed 1, whereby we have $\mathbf{M}\mathrm{diag}\left(U^T(UU^T)^{-1}U\right) \preceq I$. Therefore,

$$\begin{aligned} \nabla^2 f_p(x) &\preceq p(p+1)\mathbf{M}\mathrm{diag}\left(X^{-1-p/2}CX^{-1-p/2}\right) \\ &= p(p+1)X^{-1}\mathbf{M}\mathrm{diag}\left(U^T(UU^T)^{-1}U\right)X^{-1} \\ &\preceq p(p+1)X^{-2} \\ &= p(p+1)\nabla^2 h(x) \;, \end{aligned}$$

where the first inequality follows from the fact that the Hadamard product of two symmetric positive semidefinite matrices is also a symmetric positive semidefinite matrix and C is a positive semidefinite matrix, and the first equation follows since X is itself a diagonal matrix. The result then follows by property (iii) of Proposition 1.1. $\square$

Solving the subproblem (9). Using $h(x) = -\sum_{j=1}^n \ln(x_j)$, the subproblem (9) here is identical to that for the D-optimal design problem, since the reference function $h(\cdot)$ and the feasible domain Q are the same. Therefore the methodology discussed in section 2.2 applies here as well.

Remark 2.4. By setting $H = A^T$ and using Proposition 1.2, it can also be shown that $\hat{f}(x) := \ln\det(A^T \mathbf{D}\mathrm{iag}(Ax - b)^{-p}A)$ is $p(p+1)$-smooth relative to $h(x) := -\sum_i \ln(A_i x - b_i)$. When $p = 2$ this is the volumetric barrier function on the set $Q = \{x \in \mathbb{R}^n : Ax \geq b\}$; see [23], [3].

2.4. Optimization over $Q \subset (0, u]^n$ with $\|\nabla^2 f(x)\|$ growing as a polynomial in $\sum_{i=1}^n \frac{1}{x_i}$. Suppose that $f(\cdot)$ is a twice-differentiable convex function on $Q \subset (0, u]^n$ and that $\|\nabla^2 f(x)\| \leq q_s(\sum_{i=1}^n \frac{1}{x_i})$, where $q_s(\alpha) = \sum_{i=0}^s a_i \alpha^i$ is an s-degree polynomial in α. (Recall $\|\nabla^2 f(x)\|$ denotes the operator norm of $\nabla^2 f(x)$ with respect to the ℓ_2-norm on $\mathbb{R}^n$.) Let

$$h(x) := \frac{u^3}{2(s+1)} \left(\sum_{i=1}^n \frac{1}{x_i} \right)^{s+1} .$$

Then the following proposition states that $f(\cdot)$ is L-smooth relative to $h(\cdot)$ for an easily computable value L. This implies that no matter how fast $\nabla f(x)$ grows as x approaches the open boundary of the region $(0, u]^n$, $f(\cdot)$ is smooth relative to the simple reference function $h(\cdot)$, even though $\nabla f(\cdot)$ need not exhibit uniform Lipschitz continuity on Q.

PROPOSITION 2.4. *Suppose $f(\cdot)$ is twice differentiable on Q and satisfies $\|\nabla^2 f(x)\| \leq q_s(\sum_{i=1}^n \frac{1}{x_i})$ where $q_s(\alpha)$ is an s-degree polynomial in α. Let L be such that $q_s(\alpha) \leq L\alpha^s$ for all $\alpha \geq \frac{n}{u}$. Then $f(\cdot)$ is L-smooth relative to $h(x) = \frac{u^3}{2(s+1)}(\sum_{i=1}^n \frac{1}{x_i})^{s+1}$.*

Proof. Let $X := \mathbf{D}\mathrm{iag}(x)$, and it follows from elementary rules of differentiation that

$$(20) \qquad \nabla^2 h(x) = u^3 \left(\sum_{i=1}^n \frac{1}{x_i} \right)^s X^{-3} + \frac{u^3 s}{2} \left(\sum_{i=1}^n \frac{1}{x_i} \right)^{s-1} X^{-2} ee^T X^{-2} .$$

Therefore

$$(21) \quad \nabla^2 h(x) \succeq u^3 \left(\sum_{i=1}^n \frac{1}{x_i} \right)^s X^{-3} \succeq \left(\sum_{i=1}^n \frac{1}{x_i} \right)^s I \succeq \frac{1}{L} q_s \left(\sum_{i=1}^n \frac{1}{x_i} \right) I \succeq \frac{1}{L} \nabla^2 f(x) ,$$

where the second matrix inequality uses $u \geq x_i$ and the third matrix inequality is due to $\sum_{i=1}^n \frac{1}{x_i} \geq \sum_{i=1}^n \frac{1}{u} = \frac{n}{u}$. Therefore $f(\cdot)$ is L-smooth relative to $h(\cdot)$ by part (iii) of Proposition 1.1. $\square$

Remark 2.5. Suppose $q_s(\alpha) = \sum_{i=0}^s a_i \alpha^i$. In Proposition 2.4, one simple way to set L is to use $L = \sum_{i=0}^s |a_i| \left(\frac{u}{n} \right)^{i-s}$. This implies for $\alpha \geq \frac{n}{u}$ that

$$(22) \qquad q_s(\alpha) \leq \sum_{i=0}^s |a_i| \alpha^i \leq \left(\sum_{i=0}^s |a_i| \left(\frac{u}{n} \right)^{i-s} \right) \alpha^s = L\alpha^s .$$

Solving the subproblem (9). Let us see how we can solve the subproblem (9) for this class of optimization problems. After rescaling c by $u^3/2$, the subproblem (9) can be equivalently written as

$$(23) \qquad \min_{x \in (0,u]^n} \langle c, x \rangle + \frac{1}{s+1} \left(\sum_{i=1}^{n} \frac{1}{x_i} \right)^{s+1} .$$

Let $\theta = \left(\sum_{i=1}^{n} \frac{1}{x_i} \right)^s$, then the optimality conditions for (23) can be written as

$$(24) \qquad x_i = \begin{cases} u & \text{if} \quad c_i \leq \frac{\theta}{u^2} , \\ \sqrt{\frac{\theta}{c_i}} & \text{for} \quad c_i > \frac{\theta}{u^2} \end{cases}$$

for $i = 1, \ldots, n$. For a given $\theta > 0$, define $x_i(\theta)$ using the above rule (24), and it remains to simply determine the value of the positive scalar θ in the interval $\mathcal{F} := [\left(\frac{n}{u}\right)^s, \infty)$ that satisfies

$$(25) \qquad d(\theta) := \theta - \left(\sum_{i=1}^{n} \frac{1}{x_i(\theta)} \right)^s = 0 .$$

Notice that $d(\cdot)$ is strictly increasing on $\mathcal{F}$, and $d\left(\left(\frac{n}{u}\right)^s\right) \leq 0$ (since $x_i(\theta) \leq u$ for any θ) and $d(\theta) \to \infty$ as $\theta \to \infty$. Therefore (25) has a unique solution in $\mathcal{F}$, which can be solved with high accuracy using any suitable root-finding method, for example, binary search combined with a 1-dimensional Newton's method.

Remark 2.6. In a sense, there are basically two ways that a twice-differentiable convex function can fail to have a uniformly Lipschitz gradient: (i) when the Hessian grows without limit as $\|x\| \to \infty$ and/or (ii) when the Hessian grows without limit as $x \to x^0 \in \partial Q$. Section 2.1 has provided a mechanism for constructing a reference function $h(\cdot)$ for case (i) when the growth is polynomial, and section 2.4 has provided such a mechanism for case (ii) when the growth is polynomial. By utilizing the additivity and linear transformation properties of relative smoothness in Proposition 1.2, it should be possible to construct suitable reference functions for many convex functions of interest.

3. Computational analysis for the primal gradient scheme and the dual averaging scheme. In this section we present computational guarantees for two algorithms: the primal gradient scheme (Algorithm 1) as well as a dual averaging scheme (Algorithm 2).

3.1. Analysis of primal gradient scheme (Algorithm 1). Our main result for the primal gradient scheme is the following sublinear and linear convergence bounds.

THEOREM 3.1. *Consider the primal gradient scheme (Algorithm* 1*). If $f(\cdot)$ is L-smooth and μ-strongly convex relative to $h(\cdot)$ for some $L > 0$ and $\mu \geq 0$, then for all $k \geq 1$ and $x \in Q$, sequence $\{f(x^k)\}$ is monotonically decreasing, and the following inequality holds:*

$$(26) \qquad f(x^k) - f(x) \leq \frac{\mu D_h(x, x^0)}{\left(1 + \frac{\mu}{L-\mu}\right)^k - 1} \leq \frac{L-\mu}{k} D_h(x, x^0) ,$$

where, in the case when $\mu = 0$, the middle expression is defined in the limit as $\mu \to 0^+$.

The first inequality in (26) shows linear convergence when $\mu > 0$; indeed, in this case it holds that

$$(27) \qquad \frac{\mu D_h(x, x^0)}{\left(1 + \frac{\mu}{L-\mu}\right)^k - 1} \leq L \left(1 - \frac{\mu}{L}\right)^k D_h(x, x^0) \ .$$

(This inequality holds trivially for $k = 1$, and induction on k establishes the result for $k \geq 2$.) Furthermore, when k is large the -1 term in the denominator of the left-hand side can be ignored which yields the asymptotic bound $\mu \left(1 - \frac{\mu}{L}\right)^k D_h(x, x^0)$. The second inequality in (26) shows an $O(1/k)$ sublinear convergence rate. In particular, the convergence rate in (26) is $\frac{L}{k} D_h(x, x^0)$ when $\mu = 0$.

Note that Algorithm 1 herein and the NoLips algorithm in [6] as well as algorithm PGA-$\mathcal{B}$ in [28] are structurally identical (they are all instantiations of the primal gradient methodology). However, the step-size rule in [6] as well as the complexity analysis in [6] depends on a symmetry measure of $D_h(\cdot, \cdot)$, namely, $\alpha := \min_{x, y \neq x} D_h(x, y) / D_h(y, x)$, whereas there is no such dependence here. The instantiation of Algorithm 1 in [6] uses a smaller "step size" of $(1 + \alpha)/2L$ as opposed to $1/L$ in the update computation in Algorithm 1 (since it must always hold that $\alpha \leq 1$), and [6] proves a computational guarantee of $f(x^k) - f(x) \leq \frac{2L}{(1+\alpha)k} D_h(x, x^0)$. The bound in Theorem 3.1 is better than this symmetry-based bound, but only by a multiplicative constant factor $(1+\alpha)/2$ when $\mu = 0$; it is of course far better (linear convergence rather than sublinear convergence) when $\mu > 0$.

The proof of the bound in Theorem 3.1 relies on the following standard "three-point property."

LEMMA 3.1 (three-point property of Tseng [22]). *Let $\phi(x)$ be a convex function, and let $D_h(\cdot, \cdot)$ be the Bregman distance for $h(\cdot)$. For a given vector z, let*

$$z^+ := \arg \min_{x \in Q} \{\phi(x) + D_h(x, z)\} \ .$$

Then

$$\phi(x) + D_h(x, z) \geq \phi(z^+) + D_h(z^+, z) + D_h(x, z^+) \quad for \ all \ x \in Q \ .$$

Proof of Theorem 3.1. Define a parameter sequence

$$C_k := \frac{1}{\sum_{i=1}^{k} \left(\frac{L}{L-\mu}\right)^i} \overset{(\cdot)}{=} \frac{\mu}{L \left(\left(1 + \frac{\mu}{L-\mu}\right)^k - 1\right)} \ ,$$

where the second equality "$(\cdot)$" follows from elementary geometric series analysis, and holds only when $\mu > 0$. In particular, $C_k = \frac{1}{k}$ if $\mu = 0$. For any $x \in Q$ and $i \geq 1$ we have

$$
\begin{aligned}
(28) \quad f(x^i) &\leq f(x^{i-1}) + \langle \nabla f(x^{i-1}), x^i - x^{i-1} \rangle + L D_h(x^i, x^{i-1}) \\
&\leq f(x^{i-1}) + \langle \nabla f(x^{i-1}), x - x^{i-1} \rangle + L D_h(x, x^{i-1}) - L D_h(x, x^i) \\
&\leq f(x) + (L - \mu) D_h(x, x^{i-1}) - L D_h(x, x^i) \ ,
\end{aligned}
$$

where the first inequality follows from the definition of L-smoothness relative to $h(\cdot)$, the second inequality is due to the three-point property with

$$\phi(x) = \tfrac{1}{L} \left\langle \nabla f(x^{i-1}), x - x^{i-1} \right\rangle$$

and $z = x^{i-1}$, $z^+ = x^i$, and the last inequality uses the μ-strong convexity of $f(\cdot)$ relative to $h(\cdot)$, which implies $\langle \nabla f(x^{i-1}), x - x^{i-1} \rangle \le f(x) - f(x^{i-1}) - \mu D_h(x, x^{i-1})$. Substituting $x = x^{i-1}$ in (28) shows in particular that $f(x^i) \le f(x^{i-1})$ which proves monotonicity of the sequence $\{f(x^i)\}$.

It then follows using induction and (28) that

(29)
$$\sum_{i=1}^{k} \left(\frac{L}{L-\mu} \right)^i f(x^i) \le \sum_{i=1}^{k} \left(\frac{L}{L-\mu} \right)^i f(x) + L D_h(x, x^0) - \left(\frac{L}{L-\mu} \right)^k L D_h(x, x^k) .$$

Using the monotonicity of $f(x^i)$ and the nonnegativity of $D_h(x, x^k)$, this implies that

(30)
$$\left(\sum_{i=1}^{k} \left(\frac{L}{L-\mu} \right)^i \right) (f(x^k) - f(x)) \le L D_h(x, x^0) - \left(\frac{L}{L-\mu} \right)^k L D_h(x, x^k) \le L D_h(x, x^0) .$$

By substituting in the equality

$$\sum_{i=1}^{k} \left(\frac{L}{L-\mu} \right)^i = \frac{1}{C_k}$$

in (30) and rearranging, we obtain

(31)
$$f(x^k) - f(x) \quad \le \quad C_k L D_h(x, x^0) = \frac{\mu D_h(x, x^0)}{\left(1 + \frac{\mu}{L-\mu} \right)^k - 1} .$$

The proof of the second inequality in (26) follows by noting that $(1 + \frac{\mu}{L-\mu})^k \ge 1 + \frac{k\mu}{L-\mu}$. □

3.2. Dual averaging scheme and analysis. Another algorithm for solving our optimization problem (1) is the dual averaging scheme [17], which we present here in Algorithm 2. Somewhat akin to the primal gradient scheme, the update step in the dual averaging scheme also requires the solution of a subproblem exactly of the form (9). Notice that we need the coefficient μ of strong convexity in order to implement Algorithm 2, in contrast to the primal gradient scheme (Algorithm 1). One can always conservatively set $\mu \leftarrow 0$ in Algorithm 2 if no reasonable lower bound on the best value of μ is known.

Algorithm 2 Dual Averaging Scheme with reference function $h(\cdot)$.

Initialize. Let L, μ, and $h(\cdot)$ satisfying Definitions 1.1 and 1.2 be given.
    Let x^0 be the "$h(\cdot)$-center" of Q, namely, $x^0 \leftarrow \arg\min_{x \in Q}\{h(x)\}$, satisfying $h(x^0) = 0$.

At iteration k :
Perform Updates. Compute $f(x^k)$, $\nabla f(x^k)$, $a_{k+1} = \frac{1}{L-\mu} \left(\frac{L}{L-\mu} \right)^k$, and

$$x^{k+1} \leftarrow \arg\min_{x \in Q} \left\{ h(x) + \sum_{i=0}^{k} a_{i+1} \left(f(x^i) + \langle \nabla f(x^i), x - x^i \rangle + \mu D_h(x, x^i) \right) \right\} .$$

We have the following result regarding computational guarantees for the dual averaging scheme.

THEOREM 3.2. *Consider the dual averaging scheme (Algorithm 2). If $f(\cdot)$ is L-smooth and μ-strongly convex relative to $h(\cdot)$ with $L > \mu$, then for all $k \geq 1$ and $x \in Q$, the following inequality holds:*

$$
(32) \qquad \min_{i=1,\ldots,k}\{f(x^i)\} - f(x) \leq \frac{\mu h(x)}{\left(1 + \frac{\mu}{L-\mu}\right)^k - 1} \leq \frac{L - \mu}{k} h(x) ,
$$

where in the case $\mu = 0$, the middle expression is defined as the limit as $\mu \to 0^+$.

Similarly to the result in Theorem 3.1, the first inequality in (32) shows linear convergence when $\mu > 0$, since

$$
(33) \qquad \frac{\mu h(x)}{\left(1 + \frac{\mu}{L-\mu}\right)^k - 1} \leq L\left(1 - \frac{\mu}{L}\right)^k h(x) ;
$$

this follows using identical logic to (27).

Proof of Theorem 3.2. Define

$$
\psi_k(x) := h(x) + \sum_{i=0}^{k-1} a_{i+1}\left(f(x^i) + \langle \nabla f(x^i), x - x^i \rangle + \mu D_h(x, x^i)\right)
$$

for $k \geq 0$ and $\psi_k^* := \min_{x \in Q} \psi_k(x)$, whereby $x^k = \arg\min_{x \in Q} \psi_k(x)$ and $\psi_k(x^k) = \psi_k^*$. It follows from the definition of relative strong convexity (Definition 1.2) that for any $x \in Q$,

$$
(34) \qquad \psi_k^* \leq h(x) + A_k f(x) ,
$$

where

$$
A_k := \sum_{i=0}^{k-1} a_{i+1} \stackrel{(\cdot)}{=} \frac{1}{\mu}\left[\left(1 + \frac{\mu}{L-\mu}\right)^k - 1\right]
$$

for all $k \geq 0$, and where the second equality $(\cdot)$ above follows from elementary geometric series analysis and holds only when $\mu > 0$; note that $A_k = \frac{k}{L}$ when $\mu = 0$.

The function $\psi_k(\cdot)$ is a sum of a linear function and the reference function $h(\cdot)$ multiplied by the coefficient $1 + \mu A_k$. Therefore $(1 + \mu A_k)h(\cdot)$ and $\psi_k(\cdot)$ define the same Bregman distance, whereby for any $x \in Q$ it holds that
$$
(35)
$$
$$
(1 + \mu A_k)D_h(x, x^k) = D_{\psi_k}(x, x^k) = \psi_k(x) - \psi_k(x^k) - \langle \nabla \psi_k(x^k), x - x^k \rangle \leq \psi_k(x) - \psi_k^* ,
$$

where the last inequality utilizes $\psi_k(x^k) = \psi_k^*$ as well as the first order optimality condition of $x^k = \arg\min_{x \in Q} \psi_k(x)$. Therefore

$$
\psi_{k+1}^*
$$
$$
= \psi_{k+1}(x^{k+1})
$$
$$
= \psi_k(x^{k+1}) + a_{k+1}\left(f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \mu D_h(x^{k+1}, x^k)\right)
$$
$$
\geq \psi_k^* + a_{k+1}\left(f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \left(\mu + \frac{1}{a_{k+1}}(1 + \mu A_k)\right)D_h(x^{k+1}, x^k)\right) ,
$$

where the last inequality uses (35) with $x = x^{k+1}$. Taking into account that $\mu + \frac{1}{a_{k+1}}(1 + \mu A_k) = \frac{1 + \mu A_{k+1}}{a_{k+1}} = \frac{1}{a_{k+1}}(\frac{L}{L-\mu})^{k+1} = L$, and using the relative smoothness of $f(\cdot)$ (Definition 1.1), we obtain

$$\psi_{k+1}^* \geq \psi_k^* + a_{k+1} f(x^{k+1}) \ .$$

It then follows by induction that

$$(36) \qquad \sum_{i=0}^{k-1} a_{i+1} f(x^{i+1}) \leq \psi_k^* \leq h(x) + A_k f(x),$$

where the second inequality is from (34). The proof is completed by rearranging (36) and taking the minimum over i. $\qquad\square$

3.3. On optimization problems with a composite function. Sometimes we are interested in solving the *composite* optimization problem [18]:

$$(37) \qquad P: \quad f^* := \quad \text{minimum}_x \quad f(x) + P(x)$$
$$\text{s.t.} \quad x \in Q$$

under the same assumptions on $f(\cdot)$ and Q as in (1), but now the objective function includes another function $P(\cdot)$ that is assumed to be convex but not necessarily differentiable, and for which the following subproblem is efficiently solvable:

$$(38) \qquad x_{\text{new}} \leftarrow \arg\min_{x \in Q}\{\langle c, x \rangle + P(x) + h(x)\}$$

for any given c. Under this assumption it is straightforward to show that Algorithm 1 naturally extends to cover the case of the composite optimization problem (37) (see [6] and [28]) and that the computational guarantee in Theorem 3.1 extends to composite optimization as well. (Indeed, when $\mu = 0$ this extension is implied in principle from [28].) It turns out that one can actually view composite optimization as working with the objective function $\bar{f}(\cdot)$ that is 1-smooth relative to the reference function $\bar{h}(\cdot) := Lh(\cdot) + P(\cdot)$. However, the definition of the reference function $h(\cdot)$ has been premised on $h(\cdot)$ being differentiable on Q, which might not hold for $\bar{h}(\cdot)$ as just defined. This can all be taken care of by a suitable modification of the theory; see Appendix A.2 for details.

3.4. Questions: Accelerated methods, conjugate duality, choosing the reference function. We have shown here in section 3 that the computational guarantees of two standard first-order methods for smooth optimization—the primal gradient scheme and the dual averaging scheme—extend in precise ways to the case when $f(\cdot)$ is L-smooth relative to the reference function $h(\cdot)$. The proof techniques used here suggest that very many other first-order algorithms for smooth optimization should extend similarly with analogous computational guarantees. However, we have not been able to extend any accelerated methods, i.e., methods that attain an $O(1/k^2)$ convergence guarantee such as [16], [15], [22], to the relatively smooth case. One avenue for further research is to answer the question of whether one can develop computational guarantees for an accelerated method in the case when $f(\cdot)$ is L-smooth relative to the reference function $h(\cdot)$?

Another question that arises concerns conjugate (duality) theory for the setting of relatively smooth convex functions. One simple result in conjugate duality theory

is that when $f(\cdot)$ is L-smooth (relative to $h(\cdot) := \frac{1}{2}\|\cdot\|^2$) the conjugate function $f^*(\cdot)$ is $1/L$-strongly convex (relative to $h^*(\cdot) := \frac{1}{2}\|\cdot\|_*^2$); see [26]. Is there a way to develop a more general conjugate duality theory that yields an analogous result when $f(\cdot)$ is L-smooth relative to a general convex function $h(\cdot)$?

A third question is how can we choose the reference function $h(\cdot)$ in order to lower the value of the bounds in Theorems 3.1 and 3.2? Several ways to think about this question are discussed in Appendix A.3.

4. D-optimal design revisited: Computational guarantees using the primal gradient or dual averaging scheme. Let us now apply the computational guarantees for the primal gradient scheme (Theorem 3.1) and the dual averaging scheme (Theorem 3.2) to the D-optimal design optimization problem (16) discussed in section 2.2. Recall from the exposition in section 2.2 that $Q = \Delta_n$ and $f(x) = -\ln\det(HXH^T)$ is 1-smooth relative to the logarithmic barrier function

$$(39) \qquad h(x) = -\sum_{j=1}^{n}\ln(x_j) \ ,$$

and that the subproblem (9) is efficiently solvable. The following theorem presents a computational guarantee for using the primal gradient scheme to approximately solve the D-optimal design optimization problem (16).

THEOREM 4.1. *Consider using the primal gradient scheme (Algorithm 1) with the reference function* (39) *to solve the D-optimal design problem* (16) *using the initial point $x^0 = \frac{1}{n}e$, and suppose that $\varepsilon \le f(x^0) - f^*$. If*

$$k \ge \frac{2n\ln\left(\frac{2(f(x^0)-f^*)}{\varepsilon}\right)}{\varepsilon} \ ,$$

then $f(x^k) - f^* \le \varepsilon$.

Proof. Let $\delta = \frac{\varepsilon}{2(f(x^0)-f^*)}$. Then $\delta \le \frac{1}{2}$ since $\varepsilon \le f(x^0) - f^*$. Let $\hat{x} := (1-\delta)x^* + \delta x^0$. It follows from the convexity of $f(\cdot)$ that

$$f(\hat{x}) \le (1-\delta)f^* + \delta f(x^0) \ ,$$

whereby

$$(40) \qquad f(\hat{x}) - f^* \le \delta(f(x^0) - f^*) \ .$$

Meanwhile,

$$(41) \qquad D_h(\hat{x}, x^0) = h(\hat{x}) - h(x^0) - \langle \nabla h(x^0), \hat{x} - x^0 \rangle = h(\hat{x}) - h(x^0)$$
$$(42) \qquad \le -n\ln\left(\frac{\delta}{n}\right) + n\ln\left(\frac{1}{n}\right) = n\ln(1/\delta) \ ,$$

where the second equality uses $\nabla h(x^0) = -n \cdot e$ which then implies $\langle \nabla h(x^0), \hat{x} - x^0 \rangle = 0$, and the inequality follows since $\hat{x} \ge (\delta/n)e$. Therefore, for k satisfying the inequality in the statement of the theorem, we have

$$\begin{aligned} f(x^k) - f^* &= f(x^k) - f(\hat{x}) + f(\hat{x}) - f^* \\ &\le \frac{D_h(\hat{x}, x^0)}{k} + \delta(f(x^0) - f^*) \\ &\le \frac{n\ln(1/\delta)}{k} + \frac{\varepsilon}{2} \\ &\le \varepsilon \ , \end{aligned}$$

TABLE 1
Comparison of the order of computational guarantees for the Frank–Wolfe method [10], [1], *and the primal gradient and dual averaging schemes (Theorem* 4.1 *and Remark* 4.1*) for D-optimal design. All constants have been suppressed in order to highlight the dependencies on particular quantities of interest. It also follows from* [10] *that* $f(x^0) - f^* \leq m \ln(n/m)$ *for* $x^0 = (1/n)e$, *which can be inserted in the above bounds as well.*

Method	Iteration bound	Operations per iteration (dense case)	Total operations bound
Frank–Wolfe method	$m \ln(f(x^0) - f^*) + \dfrac{m^2}{\varepsilon}$	mn	$m^2 n \ln(f(x^0) - f^*) + \dfrac{m^3 n}{\varepsilon}$
Primal gradient scheme or dual averaging scheme	$\dfrac{n \ln(f(x^0) - f^*)}{\varepsilon} + \dfrac{n \ln\left(\frac{1}{\varepsilon}\right)}{\varepsilon}$	$m^2 n$	$\dfrac{m^2 n^2 \ln(f(x^0) - f^*)}{\varepsilon} + \dfrac{m^2 n^2 \ln\left(\frac{1}{\varepsilon}\right)}{\varepsilon}$

where the first inequality follows from Theorem 3.1 using $x = \hat{x}$, as well as (40), the second inequality is from (41) and the definition of δ, and the third inequality follows since $k \geq \lceil 2n \ln(1/\delta) \rceil / \varepsilon$. $\qquad\square$

Remark 4.1. For the dual averaging scheme (Algorithm 2), one obtains the identical bound as in Theorem 4.1. This is proved by following virtually the same logic as above, except we use Theorem 3.2 which bounds the smallest optimality gap using $h(x) - h(x^0)$ instead of $D_h(x, x^0)$. However, it follows from (41) that these two quantities are the same in this case. Also, in the case of the dual averaging scheme, the relevant final quantity of interest is $\min_{i=1,\ldots,k} f(x^i) - f^*$ instead of $f(x^k) - f^*$.

It is instructive to compare the computational guarantees in Theorem 4.1/Remark 4.1 to those of the Frank–Wolfe method applied to D-optimal design (first analyzed by Khachiyan [10] and reevaluated in [1] based in part on work by Yildirim [25]). Table 1 shows such a comparison, where absolute constants have been suppressed in order to highlight the dependencies on particular quantities of interest. The second column of Table 1 compares the iteration bound of the methods using the starting point $x^0 = (1/n)e$, where we emphasize that ε is the target optimality gap for the D-optimal design problem. While it follows from observations in [10] that $f(x^0) - f^* \leq m \ln(n/m)$ for $x^0 = (1/n)e$, we do not show this in Table 1, as we wish to highlight where the dependence on the initial iterate arises. Examining the first column of Table 1, note that the number of iterations of the primal gradient scheme (or dual averaging scheme) can be less than that of the Frank–Wolfe method, especially when ε is not too small and when $n \ll m^2$. However, as the second column of Table 1 shows, the Frank–Wolfe method requires only mn operations per iteration in the worst—i.e., dense matrix—case, as it does a rank-1 update of a matrix inverse in the computation of $\nabla f(x^k)$, whereas the primal gradient scheme (or dual averaging scheme) requires $m^2 n$ operations per iteration in the dense case (it must recompute a matrix inverse in order to work with $\nabla f(x^k)$). Therefore the total bound on operations of the Frank–Wolfe method (shown in the last column of Table 1) is superior.

The bound for the Frank–Wolfe method applied to the D-optimal design problem is based on analysis that is uniquely designed for evaluating the D-optimal design problem, and is not part of the general theory for the Frank–Wolfe method (that we are aware of). Even though the primal gradient scheme and the dual averaging scheme have inferior computational guarantees to the Frank–Wolfe method applied to the

D-optimal design problem, they are the first (that we are aware of) first-order methods for which one has a general theory (Theorems 3.1 and 3.2) that can be meaningfully applied to yield computational guarantees for the D-optimal design problem. We hope that this analysis will spur further interest in developing improved algorithms for D-optimal design and its dual problem—the minimum volume enclosing ellipsoid problem.

Appendix A.

A.1. Solving the subproblem (9) when $h(x)$ is a convex function of $\|x\|_2^2$ and Q has simple constraints. We consider the following subproblem,

$$(43) \qquad \min_{x \in Q} \ \langle c, x \rangle + h(x) \ ,$$

where $h(x) = g(\|x\|_2^2)$ and $g(\cdot)$ is a (univariate) closed convex function of $\|x\|_2^2$. Let $y := \|x\|_2^2$ and define $\mathcal{D} := \{\|x\|_2^2 : x \in Q\} \subset \mathbb{R}$, which is the domain of $g(\cdot)$. Let $g^*(\cdot)$ denote the conjugate function of $g(\cdot)$, namely,

$$g^*(t) := \sup_{y \in \mathcal{D}}\{ty - g(y)\} \ ,$$

whose domain we denote by $\mathcal{D}^*$. Since $g(\cdot)$ is a convex function, we know from conjugacy theory [5] that $g(y) = \sup_{t \in D^*}\{ty - g^*(t)\}$. Therefore (43) becomes

$$\min_{x \in Q}\left\{\langle c, x \rangle + g(\|x\|_2^2)\right\} = \min_{x \in Q}\left\{\sup_{t \in D^*}\left\{\langle c, x \rangle + t\|x\|_2^2 - g^*(t)\right\}\right\}$$

$$= \sup_{t \in D^*}\left\{-g^*(t) + \min_{x \in Q}\left\{\langle c, x \rangle + t\|x\|_2^2\right\}\right\} \ ,$$

where the second equality above holds whenever the min and the sup operators can be exchanged (which is akin to strong duality). Notice that $\min_{x \in Q}\{\langle c, x \rangle + t\|x\|_2^2\}$ is a Euclidean projection problem. Therefore the subproblem (9) becomes a 1-dimensional concave maximization problem if the Euclidean projection problem can be easily solved and one can conveniently form and work with the univariate convex conjugate function $g^*(\cdot)$.

A.2. Extension to composite optimization. Here we discuss some details of the extension of the ideas and results of this paper to composite optimization as described in section 3.3, using the definitions $\bar{f}(\cdot) := f(\cdot) + P(\cdot)$ and $\bar{h}(\cdot) = Lh(\cdot) + P(\cdot)$ as defined in section 3.3. Note that $\bar{f}(\cdot)$ and $\bar{h}(\cdot)$ are not necessarily differentiable on Q since they include the function $P(\cdot)$. However, we can use the equivalent condition from (a-ii) of Proposition 1.1 to define relative smoothness. Let us now show how convergence results for the primal gradient scheme still hold in this more general setting using an extension of the proof of Theorem 3.1.

Let $g_P(x) \in \partial P(x)$ be a specific subgradient of $P(\cdot)$ at x, and we will use the same subgradient of $P(\cdot)$ at x when constructing a subgradient of $\bar{f}(\cdot)$ and/or $\bar{h}(\cdot)$, namely, $g_{\bar{f}}(x) := \nabla f(x) + g_P(x)$ and $g_{\bar{h}}(x) := L\nabla h(x) + g_P(x)$. Then Algorithm 1 has the following update:

$$
\begin{aligned}
(44) \qquad x^{i+1} &= \arg\min_{x \in Q}\{\bar{f}(x^i) + \langle g_{\bar{f}}(x^i), x - x^i \rangle + D_{\bar{h}}(x, x^i)\} \\
&= \arg\min_{x \in Q}\{\bar{f}(x^i) + \langle \nabla f(x^i) + g_P(x^i), x - x^i \rangle + D_{Lh}(x, x^i) \\
&\qquad\qquad + P(x) - P(x^i) - \langle g_P(x^i), x - x^i \rangle\} \\
&= \arg\min_{x \in Q}\{f(x^i) + \langle \nabla f(x^i), x - x^i \rangle + LD_h(x, x^i) + P(x)\} \ ,
\end{aligned}
$$

where in the third equality above the term involving $g_P(x^i)$ arising in $\partial \bar{f}(x^i)$ cancels out the corresponding term involving $g_P(x^i)$ arising in $\partial \bar{h}(x^i)$ as part of the expansion of $D_{\bar{h}}(x, x^i)$. There is therefore no actual need to compute $g_P(x^i) \in \partial P(x^i)$ in the update. Indeed, this update (44) corresponds exactly to the update in the NoLips algorithm [6] (up to the step size) and the PGA-$\mathcal{B}$ algorithm in [28] (up to the step size) for composite optimization.

The proof of the computational guarantee in Theorem 3.1 can be generalized directly to the composite optimization setting as follows. Let us denote

$$s_i(x) := \bar{f}(x^i) + \langle g_{\bar{f}}(x^i), x - x^i \rangle + D_{\bar{h}}(x, x^i) = f(x^i) + \langle \nabla f(x^i), x - x^i \rangle + L D_h(x, x^i) + P(x).$$

Notice that $x^{i+1} = \arg\min_{x \in Q} s_i(x)$; therefore from the first-order optimality conditions there is a subgradient $g_{s_i}(x^{i+1}) \in \partial s_i(x^{i+1})$ for which $\langle g_{s_i}(x^{i+1}), x - x^{i+1} \rangle \geq 0$ for all $x \in Q$. From the additivity property of subgradients, we can write $g_{s_i}(x^{i+1}) = \nabla f(x^i) + L\nabla h(x^{i+1}) - L\nabla h(x^i) + \bar{g}$ for some $\bar{g} \in \partial P(x^{i+1})$, and let us assign $g_P(x^{i+1}) := \bar{g} = g_{s_i}(x^{i+1}) - \nabla f(x^i) - L\nabla h(x^{i+1}) + L\nabla h(x^i)$, which then is used to define the subgradient $g_{\bar{f}}(x^{i+1})$, $g_{\bar{h}}(x^{i+1})$, and the Bregman distance $D_{\bar{h}}(x, x^{i+1})$ in the proof. Recall that the primal gradient scheme does not rely on the choice of subgradient of $P(x^{i+1})$, thus the choice of $g_P(x^{i+1})$ is only used in the proof and it is well-defined.

Utilizing the above method for specifying the subgradients of $P(\cdot)$ at each of the iterates x^i of the primal gradient scheme, we can prove the following more specialized form of the three-point property which we can use in the proof of Theorem 3.1 for the setting composite optimization.

LEMMA A.1. *For any $x \in Q$, we have for any $i \geq 0$,*

$$(45) \qquad f(x^i) + \langle g_{\bar{f}}(x^i), x^{i+1} - x^i \rangle + D_{\bar{h}}(x^{i+1}, x^i)$$
$$\leq f(x^i) + \langle g_{\bar{f}}(x^i), x - x^i \rangle + D_{\bar{h}}(x, x^i) - D_{\bar{h}}(x, x^{i+1}) \ .$$

Proof. Notice that $s_i(x) - \bar{h}(x) = f(x^i) + \langle \nabla f(x^i) - L\nabla h(x^i), x - x^i \rangle - Lh(x^i)$ and so is a linear function of x, whereby it holds that

$$(s_i(x) - \bar{h}(x)) - (s_i(x^{i+1}) - \bar{h}(x^{i+1})) = \langle \nabla(s_i - \bar{h})(x^{i+1}), x - x^{i+1} \rangle$$
$$= \langle g_{s_i}(x^{i+1}), x - x^{i+1} \rangle - \langle g_{\bar{h}}(x^{i+1}), x - x^{i+1} \rangle$$
$$\geq -\langle g_{\bar{h}}(x^{i+1}), x - x^{i+1} \rangle \ ,$$

where the inequality follows from the choice of $g_{s_i}(x^{i+1})$. Rearranging the above and recalling the definition of $s_i(x)$ then completes the proof. $\qquad\square$

The proof of Theorem 3.1 in the setting of composite optimization follows directly by replacing $h(\cdot)$, $\nabla h(\cdot)$, $f(\cdot)$, and $\nabla f(\cdot)$ by $\bar{h}(\cdot)$, $g_{\bar{h}}(\cdot)$, $\bar{f}(\cdot)$, and $g_{\bar{f}}(\cdot)$, respectively, and utilizing (45) to deduce the second inequality in (28).

A.3. Criteria for choosing the reference function $h(\cdot)$. One natural question is how can we choose $h(\cdot)$ in order to lower the value of the bound in Theorem 3.1? Let us consider the simple case when $f(\cdot)$ is twice differentiable and is not strongly convex, namely, $\mu = 0$, and $f(\cdot)$ attains its optimum at some point x^*. Then the convergence bound (26) can be rewritten as

$$f(x^k) - f(x^*) \leq \frac{1}{k} D_{Lh}(x^*, x^0) = \frac{1}{k} D_f(x^*, x^0)$$
$$+ \frac{1}{k} \left(\int_0^1 \int_0^t (x^* - x^0)^T \left[\nabla^2 (Lh - f)(x^0 + s(x^* - x^0)) \right] (x^* - x^0) \, ds \, dt \right),$$

where $\nabla^2(Lh - f)(y)$ is the Hessian of the "gap function" $Lh(\cdot) - f(\cdot)$ at the point $y \in Q$. Notice that the first term above is fixed independently of the choice of $h(\cdot)$ and L. It follows from Proposition 1.1 that if $f(\cdot)$ is L-smooth relative to $h(\cdot)$ then $\nabla^2(Lh - f)(y) \succeq 0$ for any $y \in \text{int } Q$, whereby the second term above is always nonnegative. Since we do not know x^* in most cases, in order to make the bound smaller we want the Hessian $\nabla^2(Lh - f)(y)$ to be smaller for all $y \in \text{int } Q$.

There is a trade-off between how small the Hessian $\nabla^2(Lh-f)(y)$ is and how hard it will be to solve the subproblem (9). If we choose $Lh(\cdot) = f(\cdot)$, the Hessian of the gap function is 0, but solving the subproblem (9) is as hard as solving the original problem (1). On the other hand, in standard gradient descent we use $h(\cdot) = \frac{1}{2}\|\cdot\|_2^2$ in which case the subproblem (9) can be easily solved, while the Hessian of the gap function can be huge—thus implying a poorer convergence bound. There are a number of ways to try to manage this trade-off. For example, in gradient descent with preconditioning we can use $h(\cdot) = \frac{1}{2}\|\cdot\|_B^2 := \sqrt{\langle\cdot, B\cdot\rangle}$, where B is a computationally friendly positive definite matrix—typically a diagonal matrix. The criteria for designing B usually involves (i) ensuring that solving equations with B is easy (so that the subproblem (9) can be easily solved), and (ii) B is "close to" the Hessian of $f(\cdot)$ (so that the Hessian of the gap function is small).

Acknowledgment. The authors are grateful to the three referees for their comprehensive efforts and their suggestions on ways to improve the readability of the paper.

REFERENCES

[1] S. D. AHIPASAOGLU, P. SUN, AND M. J. TODD, *Linear convergence of a modified Frank-Wolfe algorithm for computing minimum volume enclosing ellipsoids*, Optim. Methods Softw., 23 (2008), pp. 5–19.

[2] K. ANSTREICHER, *Large step volumetric potential reduction algorithms for linear programming*, Ann. Oper. Res,, 62 (1996), pp. 521–538.

[3] K. M. ANSTREICHER, *The volumetric barrier for semidefinite programming*, Math. Oper. Res., 25 (2000), pp. 365–380.

[4] C. L. ATWOOD, *Optimal and efficient designs of experiments*, Ann. Math. Statist., 40 (1969), pp. 1570–1602.

[5] M. AVRIEL, *Nonlinear Programming: Analysis and Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

[6] H. H. BAUSCHKE, J. BOLTE, AND M. TEBOULLE, *A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications*, Math. Oper. Res., 42 (2017), pp. 330–348.

[7] A. BELLONI AND R. M. FREUND, *On the second-order feasibility cone: Primal-dual representation and efficient projection*, SIAM J. Optim., 19 (2008), pp. 1073–1092.

[8] C. CROUX, G. HAESBROECK, AND P. J. ROUSSEEUW, *Location adjustment for the minimum volume ellipsoid estimator*, Statist. Comput., 12 (2002), pp. 191–200.

[9] F. JOHN, *Extremum problems with inequalities as subsidiary conditions*, in Studies and Essays, Presented to R. Courant on His 60th Birthday, Interscience, New York, 1948, pp. 187–204.

[10] L. G. KHACHIYAN, *Rounding of polytopes in the real number model of computation*, Math. Oper. Res., 21 (1996), pp. 307–320.

[11] L. G. KHACHIYAN AND M. J. TODD, *On the complexity of approximating the maximal inscribed ellipsoid for a polytope*, Math. Program., 61 (1993), pp. 137–159.

[12] J. KIEFER AND J. WOLFOWITZ, *The equivalence of two extremum problems*, Canad. J. Math., 12 (1960), pp. 363–365.

[13] E. M. KNORR, R. T. NG, AND R. H. ZAMAR, *Robust space transformations for distance-based operations*, in Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, 2001, pp. 126–135.

[14] A. NEMIROVSKY AND D. B. YUDIN, *Problem Complexity and Method Efficiency in Optimization*, Wiley, New York, 1983.

[15] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic, Boston, 2003.

[16] Y. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.

[17] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Math. Program., 120 (2009), pp. 221–259.

[18] Y. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program., 140 (2013), pp. 125–161.

[19] B. Polyak, *Introduction to Optimization*, Optimization Software, New York, 1987.

[20] P. Sun and R. M. Freund, *Computation of minimum-volume covering ellipsoids*, Oper. Res., 52 (2004), pp. 690–706.

[21] M. J. Todd, *Minimum-Volume Ellipsoids: Theory and Algorithms*, MOS-SIAM Ser. Optim. 23, SIAM, Philadelphia, 2016.

[22] P. Tseng, *On Accelerated Proximal Gradient Methods for Convex-Concave Optimization*, Technical report, MIT, Cambridge, MA, 2008.

[23] P. M. Vaidya, *A new algorithm for minimizing convex functions over convex sets*, in Proceedings of the 30th Annual Symposium on Foundations of Computer Science, 1989, IEEE Computer Society, Los Angeles, 1989, pp. 338–343.

[24] Y. Ye, *A new complexity result for minimizing a general quadratic function with a sphere constraint*, in Recent Advances in Global Optimization, C. Floudas and P. Pardalos, eds., Princeton University Press, Princeton, NJ, 1992, pp. 19–31.

[25] E. A. Yildirim, *On the minimum volume covering ellipsoid of ellipsoids*, SIAM J. Optim., 17 (2006), pp. 621–641.

[26] C. Zalinescu, *Convex Analysis in General Vector Spaces*, World Scientific, Singapore, 2002.

[27] Y. Zhang, *An Interior-Point Algorithm for the Maximum-Volume Ellipsoid Problem*, Department of Computational and Applied Mathematics, Technical report TR98-15, Rice University, Houston, TX, 1998.

[28] Y. Zhou, Y. Liang, and L. Shen, *A Unified Approach to Proximal Algorithms using Bregman Distance*, Technical report, Syracuse University, Syracuse, NY, 2016.