

# Optimal Complexity and Certification of Bregman First-Order Methods

Radu-Alexandru Dragomir · Adrien B. Taylor ·  
Alexandre d'Aspremont\* · Jérôme Bolte\*

Last revised on September 29, 2020

**Abstract** We provide a lower bound showing that the  $O(1/k)$  convergence rate of the NoLips method (a.k.a. Bregman Gradient or Mirror Descent) is optimal for the class of problems satisfying the relative smoothness assumption. This assumption appeared in the recent developments around the Bregman Gradient method, where acceleration remained an open issue.

The main inspiration behind this lower bound stems from an extension of the performance estimation framework of Drori and Teboulle (Mathematical Programming, 2014) to Bregman first-order methods. This technique allows computing worst-case scenarios for NoLips in the context of relatively-smooth minimization. In particular, we used numerically generated worst-case examples as a basis for obtaining the general lower bound.

## 1 Introduction

We consider the constrained minimization problem

$$\min_{x \in C} f(x) \tag{P}$$

where  $f$  is a convex continuously differentiable function and  $C$  is a nonempty closed convex subset of  $\mathbb{R}^n$ . In large-scale settings, first-order methods are particularly popular due to their simplicity and their low cost per iteration.

The (projected) gradient descent (PG) is a classical method for solving (P), and consists in successively minimizing quadratic approximations of  $f$ , with

$$x_{k+1} = \operatorname{argmin}_{u \in C} f(x_k) + \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{2\lambda} \|u - x_k\|^2, \tag{PG}$$

where  $\|\cdot\|$  is the Euclidean norm. Although standard, there is often no good reason for making such approximations, beyond our capability of solving this intermediate optimization problem. In other words, this traditional approximation typically does not reflect neither the geometry of  $f$  nor that of  $C$ . A powerful generalization of PG consists in performing instead a *Bregman gradient step*

$$x_{k+1} = \operatorname{argmin}_{u \in C} f(x_k) + \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{\lambda} D_h(u, x_k), \tag{BG}$$

\* Last two authors listed in alphabetical order

Radu-Alexandru Dragomir  
Université Toulouse I Capitole, Toulouse & D.I. École Normale Supérieure, Paris, France. radu-alexandru.dragomir@inria.fr

Adrien B. Taylor  
INRIA, D.I. École Normale Supérieure, Paris, France. adrien.taylor@inria.fr

Alexandre d'Aspremont  
CNRS & D.I., UMR 8548, École Normale Supérieure, Paris, France. aspremon@ens.fr

Jérôme Bolte  
TSE (Université Toulouse 1 Capitole), Toulouse, France. jbolte@ut-capitole.fr

where the Euclidean distance has been replaced by the *Bregman distance*  $D_h(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle$  induced by some strictly convex and continuously differentiable *kernel function*  $h$ . A well-chosen  $h$  allows designing first-order algorithms adapted to the geometry of the constraint set and/or the objective function. Of course, a conflicting goal is to choose  $h$  such that each iteration (BG) can be solved efficiently in practice, discarding choices such as  $h = f$  (for which performing an iteration would be as hard as solving the original problem).

Recently, Bauschke et al. [4] introduced a natural condition for analyzing this scheme, assuming that the inner objective in the iteration (BG) is an upper bound on  $f$ . This ensures that performing an iteration decreases the function values  $f(x_k)$ . This assumption, known as relative smoothness (precisely defined in Def. 2 below), generalizes the standard  $L$ -smoothness assumption implied by Lipschitz continuity of  $\nabla f$ . The Bregman gradient algorithm, also called NoLips in the setting of [4], is thus a natural extension of gradient descent (PG) to objective functions whose geometry is better modeled by a non-quadratic kernel  $h$ . Practical examples of relative smoothness arise in Poisson inverse problems [4], quadratic inverse problems [10], rank minimization [14] and regularized higher-order tensor methods [33].

*Can we accelerate NoLips?* In the Euclidean setting where  $h(x) = \frac{1}{2}\|x\|^2$ , accelerated projected gradient methods exhibit faster convergence than the vanilla projected gradient algorithm. These methods, which can be traced back to Nesterov [31], are proven to be *optimal* for  $L$ -smooth functions and have found a number of successful applications, in e.g., imaging [7]. A natural question is therefore to understand whether the NoLips algorithm can be accelerated in the relatively-smooth setting. This question has been raised in several works, including that of Bauschke, Bolte and Teboulle [4, Section 6], that of Lu, Freund and Nesterov [26, Section 3.4], and the survey of Teboulle [39, Section 6]. Partial answers have already been provided under somewhat strict additional regularity assumptions (see e.g., [1, 41, 22] and discussions in the sequel), while the general case was apparently still open, and relevant in practical applications. In this work, we produce a lower complexity bound proving that NoLips is *optimal* for the general relatively-smooth setting, and therefore that generic acceleration is impossible.

In order to do so, we adopt the standard *black-box model* used for studying complexity of first-order methods [30]. We consider that both  $f$  and  $h$  are described by first-order oracles, so as to obtain generic complexity results, and we look for worst-case *couples* of functions  $(f, h)$  satisfying the relative smoothness assumption. A central idea in our approach is the fact that, when studying the worst-case behavior of Bregman methods in the relatively-smooth setting,  $f$  and  $h$  can get arbitrarily close to some *limiting pathological nonsmooth functions*.

*Obtaining worst-case scenarios of Bregman first-order methods.* To obtain the lower complexity bound, we start by empirically inspecting the worst-case behaviors of NoLips. In other words, we show that worst-case scenarios (i.e., worst-case pairs of functions  $(f, h)$ ) can be generated numerically through appropriate semidefinite programs (SDP).

The problems of computing such worst-case scenarios are usually referred to as *performance estimation problems* (PEPs), and were pioneered by [18] in the context of smooth convex minimization. An additional attractive feature of this approach is that feasible points to their dual problems naturally correspond to worst-case guarantees. For our purposes, we adapt the PEP framework to the setting of Bregman methods and relatively-smooth functions, and showcase the approach by providing worst-case examples for NoLips, along with the corresponding worst-case guarantees coming from its dual. Finally, the very simple and pathological worst-case functions for NoLips served as an inspiration for developing the more general lower bound for Bregman first-order schemes.

Discovering worst-case functions for NoLips is not the only interest of PEPs, as they also allow us to explore worst-case behaviors and convergence bounds for a variety of first-order methods, in a variety of settings, as we illustrate in the sequel.

## 1.1 Contributions and paper organization

The main contribution of this work is twofold. First, we provide a lower bound showing that it is impossible to generically accelerate Bregman gradient methods under the appropriate oracle model. More precisely, we show that the  $O(1/k)$  convergence rate on function values of NoLips is *optimal* in the relatively-smooth setting. As mentioned earlier, the family of worst-case functions that we used for

developing the lower bound was inspired by numerical solutions to a series of Performance Estimation Problems (PEPs).

For this purpose, we developed PEP techniques for Bregman settings. It required extending the analysis of [37] to handle classes of differentiable (but not necessarily  $L$ -smooth) and strictly convex functions. While we present the analysis on the basic NoLips algorithm for readability purposes, our results and methodology can be applied to various Bregman methods, such as inertial variants [1], or the Bregman proximal point scheme for convex minimization and monotone inclusions [20, 12]. Besides discovering worst-case examples, PEPs can be used for obtaining bounds with various convergence criteria, as we showcase by proving a new rate on the Bregman divergence between successive iterates for NoLips.

The paper is organized as follows. After introducing the setup in Section 2, we prove the optimality of NoLips in Section 3. We expose the framework of computer-aided analysis of Bregman methods in Section 4, including several applications in Section 4.5. We point out that Sections 3 and 4 are both of independent interest and can be read separately.

## 1.2 Related work

*Bregman methods.* The idea of using non-Euclidean geometries induced by convex kernels can be traced back to the work of Nemirovskii and Yudin [30]. For nonsmooth objectives, it gave birth to the mirror descent algorithm [8, 6, 23], which generalizes the subgradient method to non-Euclidean geometries. It has been proven to be particularly efficient for minimization on the unit simplex, where choosing the *entropy kernel* turns out to be much more effective and scalable than the squared Euclidean norm. This approach has been very successful in online learning; see [11, Chap. 5] and references therein. The use of Bregman distances has also been thoroughly studied for interior proximal methods [13, 38, 20, 1].

The introduction of the relative smoothness assumption in [4] has provided a way to adapt the Bregman kernel to the geometry of the objective function  $f$  and thus extend the domain of application of the Bregman Gradient method. Subsequent work has focused on nonconvex extensions [10], linear convergence rates under additional assumptions [26, 3], and inertial variants [22, 29].

*Black-box model and lower complexity bounds.* The first-order black-box model, developed initially in the works of Nemirovskii [30] and later Nesterov [32] has allowed to prove optimal complexity for several classes of problems in first-order optimization [15]. These results usually rely on well-chosen *worst-case functions* whose structure makes them difficult to minimize for all methods within a given class. Our worst-case instances are obtained from pointwise maxima of affine functions, reminiscent of lower bounds for nonsmooth convex minimization [30, 42]. Our construction then involves smoothing those piecewise affine functions, making them differentiable. This technique is also used in the very related work of Guzman and Nemirovskii [21], which studies lower bounds for minimization of convex functions that are smooth with respect to  $\ell_p$  norms. To the best of our knowledge, the lower bound obtained in the sequel is not a particular case of those in [21], as smoothness with respect to a certain norm is different from relative smoothness with respect to the same (squared) norm, beyond the  $\ell_2$ -norm.

*Performance estimation problems.* The PEP methodology, proposed initially by [18], was already used to discover optimal methods and corresponding lower bounds in other settings: for smooth convex minimization [18, 24, 15, 17], nonsmooth convex minimization [19, 17], and stochastic optimization [16].

## 1.3 Notation

We use  $\bar{C}$  to denote the closure of a set  $C$ ,  $\text{int} C$  for its interior and  $\partial C$  for its boundary. We denote  $(e_1, \dots, e_n)$  the canonical basis of  $\mathbb{R}^n$ , and for  $p \in \{1, \dots, n\}$  we write  $E_p = \text{Span}(e_1, \dots, e_p)$  the set of vectors supported by the first  $p$  coordinates.  $\mathbf{S}_n$  denotes the set of symmetric matrices of size  $n$ . If (P) is an optimization problem, then  $\text{val}(\text{P})$  stands for its (possibly infinite) value.

Subscripts on a vector denote the iteration counter, while a superscript such as  $x^{(i)}$  denotes the  $i$ -th coordinate. The set  $I = \{0, 1, \dots, N, *\}$  is often used to index the first  $N$  iterates of an optimization algorithm as well as the optimal point:

$$\{x_i\}_{i \in I} = \{x_0, x_1, \dots, x_N, x_*\}.$$

We use the standard notation  $\langle \cdot, \cdot \rangle$  for the Euclidean inner product, and  $\| \cdot \|$  for the corresponding Euclidean norm. For a vector  $x \in \mathbb{R}^n$ , we write  $\|x\|_\infty = \max_{i=1\dots n} |x^{(i)}|$  for its  $\ell_\infty$  norm. Other notations are standard from convex analysis; see e.g., [34, 5].

## 2 Algorithmic setup

In this section, we introduce the base ingredients and technical assumptions on  $f$  and  $h$  that are used within Bregman first-order methods.

### 2.1 Kernel functions

Let  $C$  be a nonempty closed convex subset of  $\mathbb{R}^n$ . The first step in defining Bregman methods is the choice of a *kernel* (or reference) function  $h$  on  $C$ .

**Definition 1 (Kernel function)** A function  $h : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  is called a kernel function on  $C$  if

- (i)  $h$  is closed convex proper (c.c.p.),
- (ii)  $\overline{\text{dom } h} = C$ ,
- (iii)  $h$  is continuously differentiable and strictly convex on  $\text{int dom } h \neq \emptyset$ .

A kernel function  $h$  induces a *Bregman distance*  $D_h$  defined as

$$D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle \quad \forall x \in \text{dom } h, y \in \text{dom } \nabla h.$$

Note that  $D_h$  is not a distance in the classical sense, however it enjoys a separation property; due to the strict convexity of  $h$  we have  $D_h(x, y) \geq 0 \quad \forall x \in \text{dom } h, y \in \text{dom } \nabla h$ , and  $D_h(x, y) = 0$  iff  $x = y$ .

*Examples.* We list some of the most classical examples of kernel functions:

- the **Euclidean kernel**  $h(x) = \frac{1}{2}\|x\|^2$  with domain  $\mathbb{R}^n$ , and for which  $D_h(x, y) = \frac{1}{2}\|x - y\|^2$  is the Euclidean distance,
- the **Boltzmann-Shannon entropy**  $h(x) = \sum_i x^{(i)} \log x^{(i)}$  extended to 0 by setting  $0 \log 0 = 0$ , whose domain is thus  $\mathbb{R}_+^n$ ,
- the **Burg entropy**  $h(x) = \sum_i -\log x^{(i)}$  with domain  $\mathbb{R}_{++}^n$ ,
- the **quartic kernel**  $h(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2$  with domain  $\mathbb{R}^n$  [10].

We refer the reader to [4, 26] for more examples. It should be emphasized that, while a kernel function is differentiable on the interior of its domain, it is not required to be differentiable on the boundary. For instance, the Boltzmann-Shannon entropy is continuous but not differentiable at 0. Moreover, the domain of  $h$  can be closed, such as for the Boltzmann-Shannon entropy, or open, as for the Burg entropy.

*Convex conjugate.* If  $h$  is a kernel function, we define its convex conjugate  $h^*$  as

$$h^*(y) = \sup_{u \in \mathbb{R}^n} \langle u, y \rangle - h(u)$$

If, for every  $y \in \mathbb{R}^n$ , the supremum in the definition of  $h^*(y)$  is attained, then  $h^*$  is differentiable and its gradient satisfies for every  $u \in \text{dom } \nabla h^*$

$$\nabla h^*(y) = \operatorname{argmax}_{u \in \mathbb{R}^n} \langle u, y \rangle - h(u).$$

## 2.2 Relatively-smooth optimization problems

We now recall the framework of relatively-smooth optimization [4, 26] for solving the minimization problem

$$\min_{x \in C} f(x) \quad (\text{P})$$

For simplicity, we present the setting without nonsmooth regularization term; our lower bound is a fortiori valid for the Bregman *proximal* gradient algorithm designed for solving composite problems [4, Eq. (12)].

Let us first state our blanket assumptions.

### Assumption 1

- (i)  $h$  is a kernel function on  $C$ ,
- (ii)  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  is a closed convex proper function such that  $\text{dom } h \subset \text{dom } f$  and which is continuously differentiable on  $\text{dom } \nabla h$ ,
- (iii) For every  $\lambda > 0$ ,  $x \in \text{int dom } h$  and  $p \in \mathbb{R}^n$ , the problem

$$\min_{u \in C} \langle p, u - x \rangle + \frac{1}{\lambda} D_h(u, x)$$

has a unique minimizer, which lies in  $\text{dom } \nabla h$ ,

- (iv) The problem has at least one minimizer, i.e.,  $\text{argmin}_C f \neq \emptyset$ .

Condition (iii) is standard and ensures well-posedness of Bregman gradient methods. It is satisfied if, for instance,  $h$  is strongly convex or supercoercive [4, Lemma 2]. In addition to these assumptions, the central property we need in order to apply the Bregman gradient method is the so-called relative smoothness [4, 26].

**Definition 2 (Relative smoothness)** Let  $h$  be a kernel function on  $C$ , and  $f$  a function such that  $\text{dom } h \subset \text{dom } f$ . We say that  $f$  is smooth relative to  $h$  if there exists a constant  $L > 0$  such that

$$Lh - f \quad \text{is convex on } \text{dom } h. \quad (\text{LC})$$

Relative smoothness allows to build a simple global majorant of  $f$ ; indeed, (LC) implies that (see, e.g, [4])

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + LD_h(x, y) \quad \forall x \in \text{dom } h, y \in \text{dom } \nabla h,$$

and the NoLips method consists in successively minimizing this upper approximation.

We list below some examples of relatively-smooth problems.

- **Euclidean case.** When choosing the Euclidean kernel  $h(x) = \frac{1}{2}\|x\|^2$ , (LC) reduces to the usual descent lemma and holds for instance when the gradient of  $f$  is Lipschitz continuous. To avoid ambiguity, we refer to this standard Euclidean smoothness as *L-smoothness*.
- **Classical mirror descent setting.** In some previous work on Bregman methods [1, 41], it is assumed that  $f$  has a Lipschitz continuous gradient with constant  $\tilde{L}$  and that the kernel  $h$  is  $\sigma$ -strongly convex. This is a particular case of relative smoothness, since we have

$$\begin{aligned} \begin{cases} \nabla f \text{ is Lipschitz continuous with constant } \tilde{L} \\ h \text{ is } \sigma\text{-strongly convex} \end{cases} &\implies \begin{cases} \frac{\tilde{L}}{2}\|\cdot\|^2 - f \text{ is convex} \\ h - \frac{\sigma}{2}\|\cdot\|^2 \text{ is convex} \end{cases} \\ &\implies \frac{\tilde{L}}{\sigma}(h - \frac{\sigma}{2}\|\cdot\|^2) + (\frac{\tilde{L}}{2}\|\cdot\|^2 - f) \text{ is convex} \\ &\implies \frac{\tilde{L}}{\sigma}h - f \text{ is convex} \\ &\implies f \text{ is smooth relative to } h \text{ with constant } \tilde{L}/\sigma. \end{aligned}$$

- **Poisson inverse problems.** More recent examples include functions that are not  $L$ -smooth in the Euclidean sense, such as the Kullback-Leibler divergence between some observation  $b \in \mathbb{R}^m$  and a linear measurement  $Ax$  of an unknown source vector  $x \in \mathbb{R}^n$ :

$$f(x) = D_{\text{KL}}(b, Ax) = \sum_{j=1}^m b_j \log\left(\frac{b_j}{A_j x}\right) - A_j x + b_j.$$

Minimizing  $f$  on the nonnegative orthant allows the recovery of a signal corrupted with Poisson noise, which is a fundamental problem in imaging sciences [9]. In this setting,  $f$  is not  $L$ -smooth since its Hessian diverges around the origin. However, it can be shown to be relatively-smooth with respect to the Burg entropy  $h(x) = \sum_i -\log(x^{(i)})$  (see [4]).

- **Quartic functions.** A large class of problems in phase recovery and low-rank matrix optimization involve minimizing polynomials of degree 4. These polynomials are not globally  $L$ -smooth but are relatively-smooth with respect to the quartic kernel  $h(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2$  (see [10, 14]).

We use the following convenient notation to characterize the class of relatively-smooth problems.

**Definition 3** We say that the couple of functions  $(f, h)$  is a relatively-smooth instance, and write  $(f, h) \in \mathcal{B}_L(C)$  if

- (i)  $f$  and  $h$  satisfy Assumption 1,
- (ii)  $Lh - f$  is convex on  $C$ .

Finally, let us denote by  $\mathcal{B}_L$  the union of  $\mathcal{B}_L(C)$  for all closed convex sets  $C$ :

$$\mathcal{B}_L = \bigcup_{n \geq 1} \bigcup_{\substack{C \subset \mathbb{R}^n \\ C \text{ closed convex}}} \mathcal{B}_L(C)$$

### 2.3 The NoLips/Bregman Gradient algorithm

Previous assumptions allow defining the Bregman Gradient (BG)/NoLips algorithm for minimizing  $f$ . For simplicity, we only consider the constant step size method.

---

**Algorithm 1** Bregman Gradient (BG) / NoLips [4]

---

**Input:**  $(f, h) \in \mathcal{B}_L(C)$ ,  $x_0 \in \text{int dom } h$ , step size  $\lambda \in (0, 1/L]$ .  
**for**  $k = 0, 1, \dots$  **do**

$$x_{k+1} = \underset{u \in \mathbb{R}^n}{\text{argmin}} \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{\lambda} D_h(u, x_k) \quad (1)$$

**end for**

---

Using first-order optimality conditions, update (1) can alternatively be written as

$$x_{k+1} = \nabla h^* [\nabla h(x_k) - \lambda \nabla f(x_k)] \quad (2)$$

involving the gradient  $\nabla h^*$  which is usually referred to as the *mirror map*. The three operations  $\nabla f$ ,  $\nabla h$  and  $\nabla h^*$  are the basic building blocks of Bregman-type methods, which we now define formally.

### 2.4 Defining a class of Bregman first-order methods

For proving a general lower bound for relatively-smooth optimization, we need to specify the oracle model and the class of methods under consideration.

We adopt the first-order black-box model, where information about a function can be gained by calling an *oracle* returning the value and gradient of  $f$  at a given point. In the Bregman setting, we assume that we also have access to the first-order oracles of the kernel function  $h$  and its conjugate  $h^*$ .

**Definition 4** An algorithm  $\mathcal{A}$  is called a Bregman first-order algorithm if, for a given problem instance  $(f, h) \in \mathcal{B}_L$  and number of iterations  $T \in \mathbb{N}$ , it generates at each time step  $t \in \{0, \dots, T\}$ , a set of primal points  $\mathcal{X}_t$  and dual points  $\mathcal{Y}_t$  from the following process:

1. Set  $\mathcal{X}_0 = \{x_0\}$ , where  $x_0 \in \text{int dom } h$  is some initialization point, and  $\mathcal{Y}_0 = \{\nabla f(x_0), \nabla h(x_0)\}$ .
2. For each  $t = 1, \dots, T$ , perform one of the two following operations:
  - either call the **primal oracle**  $(\nabla f, \nabla h)$  at some point  $x_t$  chosen such as

$$x_t \in \text{Span}(\mathcal{X}_{t-1}) \cap \text{dom } \nabla h$$

and update the dual set as

$$\mathcal{Y}_t = \mathcal{Y}_{t-1} \cup \{\nabla f(x_t), \nabla h(x_t)\}.$$

- Or call the **mirror oracle**  $\nabla h^*$  at some dual point  $y_t$  chosen such as

$$y_t \in \text{Span}(\mathcal{Y}_{t-1})$$

with

$$\nabla h^*(y_t) = \underset{u \in C}{\text{argmin}} h(u) - \langle y_t, u \rangle$$

and update the primal set as

$$\mathcal{X}_t = \mathcal{X}_{t-1} \cup \{\nabla h^*(y_t)\}.$$

3. Output some point  $x_T \in \text{Span}(\mathcal{X}_T)$ .

Such structural assumptions on the class of algorithms are classical from complexity analyses of Euclidean first-order methods and are used to prove e.g., optimality of accelerated first order methods [32]. Definition 4 is a natural extension to the Bregman setting, allowing additional uses of the oracles associated with the kernel function  $h$ . This model can often be relaxed through the use of more involved information theoretic arguments, see e.g., [30, 21, 15, 42].

Here, we focus on Definition 4 as it is general enough to encompass all Bregman-type methods that only use oracles for  $\nabla f, \nabla h$ , which we call the *primal oracles*, the map  $\nabla h^*$ , which we call the *mirror oracle*, as well as linear operations. One can verify that all known Bregman gradient methods, including NoLips and inertial variants such as IGA [1] or the recent algorithm in [22], fit in this model.

Observe that, as NoLips performs one primal oracle call and one mirror call per iteration, an iteration of NoLips corresponds actually to *two time steps* of the formal procedure in Definition 4. This is why, in order to avoid ambiguity, we state our lower bound as a function of the number of oracle calls.

### 3 Convergence rate and optimality of NoLips

In this section, we start by recalling the  $O(1/k)$  convergence rate bound for the NoLips algorithm in the setting where  $(f, h) \in \mathcal{B}_L(C)$ . We then proceed to prove that NoLips is an *optimal* algorithm for the class  $\mathcal{B}_L(C)$ , by showing that this rate is also a *lower bound* for a generic class of Bregman gradient algorithms that we define below. The key elements for proving the lower bound were empirically discovered through the solution to a Performance Estimation Problem (PEP), which is detailed in Section 4.

#### 3.1 Upper bound

We first state the  $O(1/k)$  convergence rate for NoLips. It slightly differs with the one from [4], as it is improved by a factor of 2 and does not involve the so-called *symmetry coefficient*.

**Theorem 1 (NoLips convergence rate)** *Let  $L > 0$ ,  $C$  be a nonempty closed convex subset of  $\mathbb{R}^n$  and  $(f, h) \in \mathcal{B}_L(C)$  be an relatively-smooth instance. Then the sequence  $\{x_k\}_{k \geq 0}$  generated by Algorithm 1 with constant step size  $\lambda \in (0, 1/L]$  satisfies for all  $k \geq 0$*

$$f(x_k) - f(u) \leq \frac{D_h(u, x_0)}{\lambda k} \tag{3}$$

for every  $u \in \text{dom } h$ .

*Remark 1* Let  $x_* \in \operatorname{argmin}_C f$ . In order to take  $u = x_*$  in Equation (3) and obtain a bound on the suboptimality gap  $f(x_k) - f(x_*)$ , we need  $x_*$  to belong to the domain of  $h$ . In most cases, this condition is trivially satisfied. However, it can fail if  $x_*$  lies on the boundary of  $C$  and  $\operatorname{dom} h$  is open, such as for the Burg entropy.

The proof of Theorem 1, whose analytical form has been inferred from solving a PEP, is provided in Section 4.5.1. This result extends the  $O(1/k)$  rate of Euclidean gradient descent for  $L$ -smooth functions to the relatively-smooth setting.

*Faster algorithms under additional assumptions.* It is natural to ask whether an *accelerated* Bregman algorithm can be obtained, with a better convergence rate than  $O(1/k)$ . This has already been achieved under additional regularity assumptions, as follows

- in the Euclidean setting, when  $h(x) = \frac{1}{2}\|x\|^2$  and  $f$  is  $L$ -smooth, the seminal accelerated gradient method of Nesterov [31] enjoys a  $O(1/k^2)$  convergence rate, which is optimal for this class of functions [32].
- When  $h$  is a strongly convex kernel with closed domain and  $f$  is  $L$ -smooth (which, as discussed in Section 2.2, is a particular case of relative smoothness), the Improved Interior Gradient Algorithm (IGA) [1] also admits a  $O(1/k^2)$  convergence rate using the same momentum technique as Nesterov-type methods.
- Recently, [22] proposed an accelerated Bregman proximal gradient algorithm with rate  $O(1/k^\gamma)$ , where  $\gamma \in [1, 2]$  is determined by some crucial *triangle scaling property* of the Bregman distance, whose genericity is unclear.

However, the existence of an accelerated algorithm for the general relatively-smooth setting was still an open question prior to this work. Indeed, many applications such as Poisson inverse problems [4] or D-optimal design [26] do not satisfy the supplementary assumptions made in the works mentioned above. In the next section, we prove that, up to a constant factor of 2, the bound (3) is not improvable in general for Bregman-type methods, making NoLips an *optimal* algorithm in the black box setting for  $(f, h) \in \mathcal{B}_L$ .

### 3.2 A lower bound for relatively-smooth Bregman optimization

We show in Theorem 2 below that for any  $\epsilon \in (0, 1)$  and number of oracle calls  $N$ , there is a pair of functions  $(f, h) \in \mathcal{B}_L(\mathbb{R}^{2N+1})$  and some  $x_0 \in \mathbb{R}^{2N+1}$  such that for any *Bregman gradient algorithm* initialized at  $x_0$ , the output  $x_N$  returned after performing at most  $N$  oracle calls satisfies

$$f(x_N) - \min_{\mathbb{R}^{2N+1}} f \geq (1 - \epsilon) \frac{LD_h(x_0, x_*)}{2N + 1}. \quad (4)$$

*Proof intuition.* For finding an instance  $(f, h)$  which is difficult for all Bregman methods, we use two main ideas. The first is the well-known technique used by Nesterov [32] for proving that  $O(1/k^2)$  is the optimal complexity for  $L$ -smooth convex minimization. He defines a “worst function in the world” that allows any gradient method to discover only one dimension per iteration, hence *hiding* the minimizer from the algorithm in the remaining unexplored dimensions.

The second idea is more specific to our setting, and relies on the fact that the set of relatively-smooth problems  $\mathcal{B}_L(C)$  is not closed. In particular, a limit of differentiable functions need not be differentiable. Thence, we actually build a worst-case **sequence** of differentiable functions parameterized by some parameter  $\mu$ , whose limit when  $\mu \rightarrow 0$  is a nonsmooth pathological function.

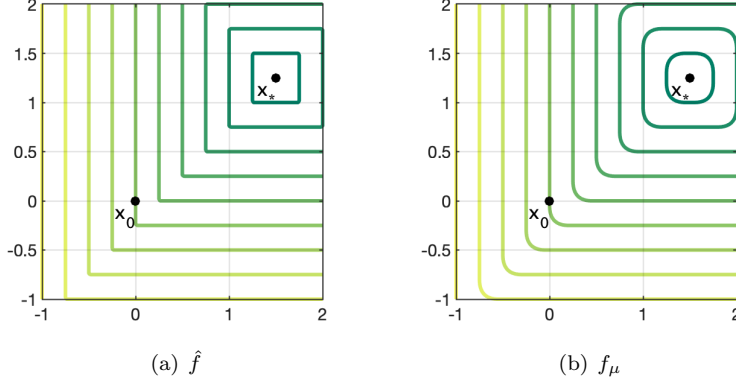
*Choosing the objective function.* Let us fix a dimension  $n \geq 1$  and a positive constant  $\eta > 0$ . Define the convex function  $\hat{f}$  for  $x \in \mathbb{R}^n$  by

$$\hat{f}(x) = \max_{i=1, \dots, n} |x^{(i)} - 1 - \frac{\eta}{i}| = \|x - x_*\|_\infty$$

which has an optimal value  $\hat{f}_* = 0$  attained at

$$x_* := (1 + \eta, 1 + \frac{\eta}{2}, \dots, 1 + \frac{\eta}{n}).$$





**Fig. 1** Level curves of function  $\hat{f}$  (left) and of its smoothed Moreau envelope  $f_\mu$  (right) for  $n = 2, \mu = 0.2$  and  $\eta = 1/2$ . Lemma 3 states that if  $\mu$  is small enough compared to  $\eta$ , the behaviors of  $\hat{f}$  and  $f_\mu$  at  $x_0 = 0$  are the same. Indeed, the size of the smoothed region where the corners are “rounded” decreases when  $\mu$  goes to 0.

The behavior of  $\hat{f}$  as a *pathological function* comes from the fact that if at least one of the coordinates of  $x$  is zero, then  $\hat{f}(x) - \hat{f}_* \geq 1$ . Let us first prove a technical lemma about the subdifferential of  $\hat{f}$ .

**Lemma 1** *Let  $x \in \mathbb{R}^n$  and  $v \in \partial \hat{f}(x)$  be a subgradient of  $\hat{f}$  at  $x$ . Then*

- (i)  $\|v\|_\infty \leq 1$ .
- (ii) Let  $i \in \{1 \dots n\}$ . If  $v^{(i)} \neq 0$  then  $|x^{(i)} - x_*^{(i)}| = \|x - x_*\|_\infty$ .

*Proof* Write  $\hat{f}$  as  $\hat{f}(x) = \max_{1 \leq i \leq n} \hat{f}_i(x)$  with  $\hat{f}_i(x) = |x^{(i)} - x_*^{(i)}|$ . Then, by [32, Lemma 3.1.10], we have

$$\partial \hat{f}(x) = \text{Conv} \{ \partial \hat{f}_i(x) | i \in I(x) \}$$

where  $I(x) = \{i \in \{1 \dots n\} | \hat{f}_i(x) = \hat{f}(x)\}$ . Hence, (i) follows immediately from the well-known property that the subgradients of the absolute value lie in  $[-1, 1]$ . (ii) is a consequence of the fact that if  $v^{(i)} \neq 0$ , then  $i \in I(x)$ , which means that  $|x^{(i)} - x_*^{(i)}| = \|x - x_*\|_\infty$ . ■

Note that  $\hat{f}$  is nonsmooth hence does not meet our assumptions. We approach it with a differentiable function by considering its Moreau envelope  $f_\mu$  given by

$$f_\mu(x) = \min_{u \in \mathbb{R}^n} \hat{f}(u) + \frac{1}{2\mu} \|x - u\|^2 \quad (5)$$

where  $\mu \in (0, 1)$  is a small parameter.  $f_\mu$  is a smoothed version of  $\hat{f}$ , which behaves similarly to  $\hat{f}$  when we choose  $\mu$  small enough. Figure 1 illustrates this phenomenon in two dimensions.

For general properties of the Moreau proximal envelope, we refer to [27]. Let us state some properties of  $f_\mu$  that we need for the analysis.

**Lemma 2**  *$f_\mu$  is a differentiable convex function, whose minimizers are the same as those of  $\hat{f}$ . Its gradient at a point  $x \in \mathbb{R}^n$  is given by  $\nabla f_\mu(x) = \mu^{-1} (x - \text{prox}_\mu^{\hat{f}}(x))$  where*

$$\text{prox}_\mu^{\hat{f}}(x) = \underset{u \in \mathbb{R}^n}{\text{argmin}} \hat{f}(u) + \frac{1}{2\mu} \|x - u\|^2$$

*is the Moreau proximal map. Moreover,  $\nabla f_\mu$  is Lipschitz continuous with constant  $1/\mu$ .*

Let us now prove the central property of  $f_\mu$ , which states that when the last  $n - p$  coordinates of  $x$  are small enough, the gradient  $\nabla f_\mu(x)$  is supported on the first  $p + 1$  coordinates. Recall that we denote  $(e_1, \dots, e_n)$  the canonical basis of  $\mathbb{R}^n$  and write, for  $p \in \{1 \dots n\}$ ,  $E_p = \text{Span}(e_1, \dots, e_p)$  and  $E_0 = \{(0, \dots, 0)\}$ .

**Lemma 3** Assume that  $\mu \in (0, 1)$  and  $\eta > 4\mu n^2$ . Let  $p \in \{0 \dots n-1\}$ . For any vector  $x \in \mathbb{R}^n$  such that

$$\max_{i=p+1, \dots, n} |x^{(i)}| \leq \mu$$

we have that  $\nabla f_\mu(x) \in E_{p+1}$ . In addition, we have  $\|\nabla f_\mu(x)\|_\infty \leq 1$ .

*Proof* Take  $x \in \mathbb{R}^n$  such that  $\max_{i=p+1, \dots, n} |x_i| \leq \mu$ . By Lemma 2,  $\nabla f_\mu$  is given by

$$\nabla f_\mu(x) = \frac{1}{\mu}(x - \text{prox}_{\hat{f}}^\mu(x)) \quad (6)$$

Write  $y = \text{prox}_{\hat{f}}^\mu(x)$ . The optimality condition defining the proximal map yields

$$y - x + \mu v = 0 \quad (7)$$

where  $v \in \partial \hat{f}(y)$ , and therefore the combination of (6) and (7) implies

$$\nabla f_\mu(x) = v \in \partial \hat{f}(y). \quad (8)$$

Now, let us assume by contradiction that  $\nabla f_\mu(x)$  is not in  $E_{p+1}$ , meaning that there exists an index  $l \in \{p+2 \dots n\}$  such that  $v^{(l)} \neq 0$ . It follows from Lemma 1 that  $|(y - x_*)^{(l)}| = \|y - x_*\|_\infty$ . Hence we have in particular that  $|y^{(l)} - x_*^{(l)}| \geq |y^{(p+1)} - x_*^{(p+1)}|$ . Using Condition (7) to replace  $y$  we get

$$|x_*^{(l)} + \mu v^{(l)} - x^{(l)}| \geq |x_*^{(p+1)} + \mu v^{(p+1)} - x^{(p+1)}|,$$

and recalling the definition of  $x_*$  we have

$$|1 + \frac{\eta}{l} + \mu v^{(l)} - x^{(l)}| \geq |1 + \frac{\eta}{p+1} + \mu v^{(p+1)} - x^{(p+1)}|.$$

By Lemma 1,  $\|v\|_\infty \leq 1$ , so for all  $i$  we have  $1 + \mu v^{(i)} \geq 1 - \mu \|v\|_\infty \geq 0$ . In addition, we assumed that  $\max_{i=p+1, \dots, n} |x^{(i)}| \leq \mu < \frac{\eta}{4n^2}$  which implies  $\frac{\eta}{i} - x^{(i)} \geq 0$  for all  $i \geq p+1$ . Therefore, both terms inside the absolute values are nonnegative, it follows that we can drop absolute values and

$$\begin{aligned} \mu(v^{(l)} - v^{(p+1)}) &\geq \frac{\eta}{p+1} - \frac{\eta}{l} + x^{(l)} - x^{(p+1)} \\ &\geq \eta \cdot \frac{l - (p+1)}{l(p+1)} - 2\mu \\ &\geq \frac{\eta}{l(p+1)} - 2\mu \\ &\geq \frac{\eta}{n^2} - 2\mu, \end{aligned} \quad (9)$$

and therefore

$$v^{(l)} - v^{(p+1)} \geq \frac{\eta}{\mu n^2} - 2 > 2,$$

because we assumed  $\eta > 4\mu n^2$ . This is a contradiction since  $(v^{(l)} - v^{(p+1)}) \leq 2\|v\|_\infty \leq 2$ . Finally, the second part of the lemma is a consequence of (8) and  $\|v\|_\infty \leq 1$ . ■

We also need the following lemma for relating the values of  $\hat{f}$  and  $f_\mu$ .

**Lemma 4** Let  $\mu > 0$  and  $x \in \mathbb{R}^n$ . Then  $f_\mu(x) \geq \hat{f}(x) - \mu$ .

*Proof* Write  $y = \text{prox}_{\hat{f}}^\mu(x)$ . By definition of  $f_\mu$  and the proximal map we have

$$\begin{aligned} f_\mu(x) &= \hat{f}(y) + \frac{1}{2\mu} \|y - x\|^2 \\ &\geq \hat{f}(y) \\ &= \|y - x_*\|_\infty \\ &\geq \|x - x_*\|_\infty - \|x - y\|_\infty. \end{aligned}$$

Recall the optimality conditions defining the proximal map can be written as

$$\mu^{-1}(x - y) \in \partial f(y),$$

and, since all subgradients of  $\hat{f}$  have coordinates smaller than 1 (Lemma 1), we reach  $\|x - y\|_\infty \leq \mu$ . It follows that  $f_\mu(x) \geq \|x - x_*\|_\infty - \|x - y\|_\infty \geq \|x - x_*\|_\infty - \mu = \hat{f}(x) - \mu$ , which concludes the proof. ■

*Choosing the kernel.* As for the objective function  $f_\mu$ , let us pick a family of kernels  $h_\mu$ , whose behavior approach those of a nonsmooth function as  $\mu \rightarrow 0$ .

Let us first define a unidimensional convex function  $\phi_\mu : \mathbb{R} \rightarrow \mathbb{R}$  by

$$\phi_\mu(t) = \begin{cases} t - \mu/2 & \text{if } t \geq \mu, \\ \frac{1}{2\mu}t^2 & \text{elsewhere.} \end{cases}$$

Note that  $\phi_\mu$  is sometimes known as the *Huber function*, which is a smooth approximation of the absolute value and also appears as a worst-case function for first-order methods in  $L$ -smooth minimization [37].

Define  $d_\mu : \mathbb{R}^n \rightarrow \mathbb{R}$  through

$$d_\mu(x) = \frac{\mu}{2}\|x\|^2 + \sum_{i=1}^n \phi_\mu(x^{(i)}), \quad x \in \mathbb{R}^n. \quad (10)$$

$d_\mu$  is a differentiable strictly convex function, whose gradient satisfies, for  $x \in \mathbb{R}^n$  and  $i \in \{1 \dots n\}$ ,

$$\nabla d_\mu(x)^{(i)} = \mu x^{(i)} + \min(1, x^{(i)}/\mu).$$

From the expression above, we can deduce two crucial properties that we need in the sequel: for  $x \in \mathbb{R}^n$  and  $i \in \{1 \dots n\}$ , we have

$$\nabla d_\mu(x)^{(i)} = 0 \quad \text{if and only if} \quad x^{(i)} = 0, \quad (11)$$

$$|\nabla d_\mu(x)^{(i)}| \leq 1 \quad \text{implies} \quad |x^{(i)}| \leq \mu. \quad (12)$$

Let  $L > 0$ . We define the kernel  $h_\mu$  for  $x \in \mathbb{R}^n$  as

$$h_\mu(x) = \frac{1}{L} (f_\mu(x) + d_\mu(x)). \quad (13)$$

By construction,  $Lh_\mu - f_\mu$  is convex, so the relative smoothness property holds. It is easy to see that Assumption 1 is satisfied as  $h_\mu$  is strongly convex, so we have  $(f_\mu, h_\mu) \in \mathcal{B}_L(\mathbb{R}^n)$ .

*Proving the zero-preserving property of the oracles.* Now that the functions are defined, we are ready to prove that all oracles involved in the Bregman algorithm allow to discover *only one dimension per oracle call*.

**Proposition 1 (Zero-preserving property of  $\nabla f_\mu, \nabla h_\mu, \nabla h_\mu^*$ )** *Assume that  $\mu \in (0, 1)$  and  $\eta > 4\mu n^2$ . Let  $p \in \{0 \dots n-1\}$ , and  $x \in \mathbb{R}^n \cap E_p$  a vector supported by the  $p$  first coordinates. Then*

$$\nabla f_\mu(x), \nabla h_\mu(x), \nabla h_\mu^*(x) \in E_{p+1}.$$

*Proof* Let  $x \in E_p$ . Then  $x$  satisfies the assumption of Lemma 3 which proves that  $\nabla f_\mu(x) \in E_{p+1}$ . By Property (11) of  $d_\mu$ , we also have that  $\nabla d_\mu(x) \in E_p$ , which allows us to conclude that

$$\nabla h_\mu(x) = L^{-1} (\nabla f_\mu(x) + \nabla d_\mu(x)) \in E_{p+1}.$$

It remains to prove the result for  $\nabla h_\mu^*(x)$ . Write  $z = \nabla h_\mu^*(x)$ , which amounts to say that  $\nabla h_\mu(z) = x$ , that is

$$\nabla f_\mu(z) + \nabla d_\mu(z) = Lx$$

using (13). Let  $l \in \{p+1 \dots n\}$ . We have  $x \in E_p$ , hence the  $l$ -th coordinate of  $x$  is zero and

$$\nabla f_\mu(z)^{(l)} + \nabla d_\mu(z)^{(l)} = 0.$$

Using the second part of Lemma 3, we have that  $\|\nabla f_\mu(z)\|_\infty \leq 1$ ; it follows that

$$|\nabla d_\mu(z)^{(l)}| \leq 1,$$

which implies that  $|z^{(l)}| \leq \mu$ , by property (12) of  $d_\mu$ . Since this holds for any  $l \geq p+1$ , we have established

$$\max_{l=p+1, \dots, n} |z^{(l)}| \leq \mu.$$

Applying Lemma 3 to  $z$ , we obtain that  $\nabla f_\mu(z) \in E_{p+1}$ . Remembering that  $\nabla h_\mu(z) = x \in E_p$  by construction, we get

$$\nabla d_\mu(z) = L\nabla h_\mu(z) - \nabla f_\mu(z) \in E_{p+1}.$$

By Property (11) of  $d_\mu$ , it follows that  $z \in E_{p+1}$ , which concludes the proof. ■

We can now use Proposition 1 inductively to state a lower bound on the performance of any Bregman gradient algorithm applied to  $(f_\mu, h_\mu)$ .

**Proposition 2** *Let  $N \geq 1$  and choose the dimension  $n = 2N + 1$ . Let  $\mu \in (0, 1)$  and  $\eta > 4\mu^2$ . Consider the functions  $f_\mu, h_\mu : \mathbb{R}^n \rightarrow \mathbb{R}$  defined in (5) and (13) respectively. Then, for any Bregman gradient method satisfying Definition 4, applied to  $(f_\mu, h_\mu)$  and initialized at  $x_0 = (0, \dots, 0)$ , the output  $\bar{x}$  returned after performing at most  $N$  calls to each one of the primal and mirror oracles satisfies*

$$f_\mu(\bar{x}) - \min_{\mathbb{R}^n} f_\mu \geq \frac{LD_{h_\mu}(x_*, x_0)}{2N + 1} \cdot \frac{1 - \mu}{1 + \mu + \eta + \frac{\mu}{2}(1 + \eta)^2}.$$

*Proof* The zero-preserving property and the structure of Bregman gradient algorithms described in Definition 4 implies that the set of primal points  $\mathcal{X}_t$  and dual points  $\mathcal{Y}_t$  at iteration  $t$  are supported by the  $t$  first coordinates, i.e.,

$$\mathcal{X}_t, \mathcal{Y}_t \subset E_t.$$

Indeed, since we initialized  $\mathcal{X}_0 = \{x_0\} \subset E_0$ , this follows by induction. Assume that at time  $t$ , we have  $\mathcal{X}_t, \mathcal{Y}_t \subset E_t$ . If the primal oracle is chosen at iteration  $t + 1$ , since the query point  $x_{t+1}$  is taken as a linear combination of points in  $\mathcal{X}_t$  it also lies in  $E_t$ , and thus Proposition 1 states that the new dual vectors  $\nabla f_\mu(x_{t+1}), \nabla h_\mu(x_{t+1})$  belong to  $E_{t+1}$ . If, on the other hand, the mirror oracle is chosen, then with the same argument we have that  $y_{t+1} \in E_t$  and by Proposition 1 that  $\nabla h_\mu^*(y_{t+1}) \in E_{t+1}$ .

Now, because the algorithm has called at most  $N$  times each oracle, it has performed at most  $2N$  steps and thus the output point satisfies  $\bar{x} \in E_{2N}$ , which means that  $\bar{x}^{(2N+1)} = 0$ .

We use Lemma 4 to relate  $f_\mu(\bar{x})$  and  $\hat{f}(\bar{x})$ . Recalling that  $\min f_\mu = \hat{f}_* = 0$ , we get

$$\begin{aligned} f_\mu(\bar{x}) - \min_{\mathbb{R}^n} f_\mu &= f_\mu(\bar{x}) \\ &\geq \hat{f}(\bar{x}) - \mu \\ &\geq |\bar{x}^{(2N+1)} - x_*^{(2N+1)}| - \mu \\ &= 1 + \frac{\eta}{2N + 1} - \mu \\ &\geq 1 - \mu, \end{aligned} \tag{14}$$

where we used the definition of  $\hat{f}$  and the fact that  $\bar{x}^{(2N+1)} = 0$ .

Let us now upper bound the initial diameter. Remembering that  $Lh_\mu = f_\mu + d_\mu$  in (13), we have

$$LD_{h_\mu}(x_*, x_0) = D_{f_\mu}(x_*, x_0) + D_{d_\mu}(x_*, x_0).$$

by definition of the Bregman distance. To deal with the first term, we recall that  $f_\mu(x_*) = 0$  and write

$$\begin{aligned} D_{f_\mu}(x_*, x_0) &= f_\mu(x_*) - f_\mu(x_0) - \langle \nabla f_\mu(x_0), x_* - x_0 \rangle \\ &= -f_\mu(x_0) - \langle \nabla f_\mu(x_0), x_* - x_0 \rangle \\ &\leq -\hat{f}(x_0) + \mu - \langle \nabla f_\mu(x_0), x_* - x_0 \rangle \\ &= -1 - \eta + \mu - \langle \nabla f_\mu(x_0), x_* - x_0 \rangle, \end{aligned}$$

where we used again Lemma 4 at  $x_0 = (0, \dots, 0)$ . Now, Lemma 3 applies to  $x_0$  with  $p = 0$  and allows to state that  $\nabla f_\mu(x_0) \in E_1$  and that  $\|\nabla f_\mu(x_0)\|_\infty \leq 1$ . Therefore

$$|\langle \nabla f_\mu(x_0), x_* - x_0 \rangle| = |\nabla f_\mu(x_0)^{(1)} (x_*^{(1)} - x_0^{(1)})| \leq |x_*^{(1)} - x_0^{(1)}| = 1 + \eta.$$

Hence we have the following upper bound

$$\begin{aligned} D_{f_\mu}(x_*, x_0) &\leq -1 - \eta + \mu + |\langle \nabla f_\mu(x_0), x_* - x_0 \rangle| \\ &\leq \mu. \end{aligned} \tag{15}$$

The second term can be directly computed from Definition (10) of  $d_\mu$ , recalling that  $x_*^{(i)} \geq 1 \geq \mu$  for  $i \in \{0 \dots n\}$ ,

$$\begin{aligned}
D_{d_\mu}(x_*, x_0) &= d_\mu(x_*) - d_\mu(x_0) - \langle \nabla d_\mu(x_0), x_* - x_0 \rangle \\
&= d_\mu(x_*) \\
&= \sum_{k=1}^{2N+1} \left[ \frac{\mu}{2} \left(1 + \frac{\eta}{k}\right)^2 + 1 + \frac{\eta}{k} - \frac{\mu}{2} \right] \\
&\leq (2N+1) \left[ \frac{\mu}{2} (1 + \eta)^2 + \eta + 1 \right].
\end{aligned} \tag{16}$$

Combining (15) and (16) gives

$$\begin{aligned}
LD_{h_\mu}(x_*, x_0) &= D_{f_\mu}(x_*, x_0) + D_{d_\mu}(x_*, x_0) \\
&\leq \mu + (2N+1) \left[ \frac{\mu}{2} (1 + \eta)^2 + \eta + 1 \right] \\
&\leq (2N+1) \left[ \mu + \frac{\mu}{2} (1 + \eta)^2 + \eta + 1 \right].
\end{aligned}$$

This bound, along with (14), yields

$$f_\mu(\bar{x}) - \min_{\mathbb{R}^n} f_\mu \geq 1 - \mu \geq \frac{LD_{h_\mu}(x_*, x_0)}{2N+1} \cdot \frac{1 - \mu}{1 + \mu + \eta + \frac{\mu}{2}(1 + \eta)^2}$$

whence the desired result. ■

Since constants  $\mu, \eta$  can be taken arbitrarily small, we now use Proposition 1 to show that the bound can be approached to any precision and thus prove our main result.

**Theorem 2 (Lower complexity bound for  $\mathcal{B}_L$ )** *Let  $N \geq 1$ , a precision  $\epsilon \in (0, 1)$  and let  $x_0 \in \mathbb{R}^{2N+1}$  be a starting point. Then, there exist functions  $(f, h) \in \mathcal{B}_L(\mathbb{R}^{2N+1})$  such that for any Bregman gradient method  $\mathcal{A}$  satisfying Definition 4 and initialized at  $x_0$ , the output  $\bar{x}$  returned after performing at most  $N$  calls to each one of the primal and mirror oracles satisfies*

$$f(\bar{x}) - \min_{\mathbb{R}^{2N+1}} f \geq \frac{LD_h(x_*, x_0)}{2N+1} \cdot (1 - \epsilon).$$

*Proof* Consider a number  $N$  of oracle calls and a target precision  $\epsilon \in (0, 1)$ . Choose the functions  $f_\mu, h_\mu$  defined respectively in Equations (5) and (13) on  $\mathbb{R}^n$  with  $n = 2N + 1$ . These functions satisfy Assumption 1, since their domain is  $\mathbb{R}^n$ , they are convex, differentiable, and  $h_\mu$  is strongly convex. Moreover, relative smoothness holds because  $Lh_\mu - f_\mu = d_\mu$  is convex by construction. Hence  $(f_\mu, h_\mu) \in \mathcal{B}_L(\mathbb{R}^n)$ .

Because the class of problems  $\mathcal{B}_L(\mathbb{R}^n)$  is invariant by translation, we can assume without loss of generality that the algorithm is initialized at  $x_0 = (0, \dots, 0)$ . Recall that the only conditions our analysis imposed on the parameters  $\eta, \mu$  are that  $\mu \in (0, 1)$  and  $\eta > 4\mu n^2$ .

Let us then choose  $\eta = \epsilon/4$  and  $\mu = \eta/(5n^2) = \epsilon/(20n^2)$ . Under these conditions, Proposition 2 applies and gives that for any point  $\bar{x}$  returned by a Bregman gradient algorithm that is initialized at  $x_0$  and which performs at most  $N$  calls to each oracle we have

$$f_\mu(\bar{x}) - \min_{\mathbb{R}^{2N+1}} f_\mu \geq \frac{LD_{h_\mu}(x_*, x_0)}{2N+1} \cdot \frac{1 - \mu}{1 + \mu + \eta + \frac{\mu}{2}(1 + \eta)^2}.$$

The last term can be bounded from below, using our choice of  $\mu, \eta$ , and the fact that  $\eta < 1$ , as

$$\frac{1 - \mu}{1 + \eta + \mu + \frac{\mu}{2}(1 + \eta)^2} \geq \frac{1 - \mu}{1 + \eta + 3\mu} = \frac{1 - \frac{\epsilon}{20n^2}}{1 + \frac{\epsilon}{4} + \frac{3\epsilon}{20n^2}} \geq 1 - \epsilon$$

yielding the desired result. ■

*Remark 2* One can refine the result above in the case where the primal and mirror oracles are not used the same number of times. Indeed, if the primal oracles are called  $N_1$  times and the mirror oracle is called  $N_2$  times, then the same reasoning shows that the lower bound remains true by replacing  $2N$  with  $N_1 + N_2$ .

Our lower bound involves the relative smoothness constant  $L$  instead of the step size  $\lambda$  in (3), but it is equivalent (up to a factor 2) when choosing  $\lambda = 1/L$ , which is actually the best possible step size choice. This shows the optimality of NoLips within the class of Bregman first-order methods (up to a universal constant).

*Connection with Conditional Gradient and the  $\ell_\infty$  setting.* The worst-case function used for the lower bound involves the smoothing of an  $\ell_\infty$  norm. As pointed out by one of the referees, there might be a connection between the hardness of the relatively-smooth setting and the lower bound for smooth minimization on the  $\ell_\infty$  ball as done in Guzman and Nemirovskii [21]. This lower bound, which is also  $O(1/k)$ , is used by the authors to prove that the rate of the Conditional Gradient algorithm is near-optimal in this setting.

It might be insightful to examine connections between these settings in future works, for example by exploiting duality between Bregman gradient methods and Conditional Gradient, as in [2].

## 4 Computer-aided performance analyses of Bregman first-order methods

In this section, we extend the computer-aided performance estimation framework in [18, 35] to the setting of Bregman methods. In short, these results show how to compute the worst-case convergence rate of a given algorithm by solving a numerical optimization problem, called performance estimation problem (PEP). Solving a PEP offers several benefits, including:

1. Computing (numerically) the *exact* worst-case complexity of an algorithm on a given class of problems after a fixed number of iterations.
2. Studying the corresponding worst-case functions.
3. Inferring an analytical worst-case guarantee by obtaining a feasible point to the dual PEP. Such dual feasible points correspond to combinations of inequalities that certify the convergence bound.

Here, we focus on inferring worst-case functions. We used this methodology for guessing, and then designing, the lower bound provided in Section 3.2. However, solving PEPs is also useful for proving new convergence rates (see Section 4.5.2), or for getting quick numerical insights into the convergence properties of an algorithm, like for instance on the inertial algorithm IGA [1] (Section 4.5.3).

To use PEPs on Bregman methods, we extend the analysis in [18, 35] to deal with differentiable and/or strictly convex functions. Previous works on the topic modelled differentiability through an  $L$ -smoothness condition, and strict convexity through strong convexity, which are assumptions that we avoid in the Bregman setting. The key difference in our work is that the classes of differentiable and/or strictly convex functions are *open* sets. Thus, the worst-case functions for this class might lie on the closure of this set and exhibit some pathological nonsmooth behavior.

This section is organized as follows. In Section 4.1, we introduce the PEP framework. Sections 4.2-4.4 extend PEPs to the Bregman setting. We provide in Section 4.5 several applications, including the procedure used to find the worst-case functions involved in the proof of the general lower bound in Section 3.2.

### 4.1 Worst-case scenarios through optimization

We now formulate the task of finding the worst-case performance of Algorithm 1 as an optimization problem. We focus on the analysis of NoLips for simplicity. However, the same ideas are directly applicable to other Bregman-type algorithms like IGA [1] (see Section 4.5.3) or Bregman proximal point [20].

Recall that we write  $\mathcal{B}_L(C)$  for the set of function pairs  $(f, h)$  satisfying Assumption 1, such that  $Lh - f$  is convex on a convex set  $C$ . For simplicity, we first focus on the case when functions have full

domain, i.e.,  $C = \mathbb{R}^n$  for some  $n \geq 1$ . In this setting, the set  $\mathcal{B}_L(\mathbb{R}^n)$  can be rewritten as

$$\mathcal{B}_L(\mathbb{R}^n) = \left\{ f, h : \mathbb{R}^n \rightarrow \mathbb{R} \left| \begin{array}{l} f \text{ is convex, differentiable and has at least one minimizer,} \\ h \text{ is strictly convex and differentiable,} \\ Lh - f \text{ is convex,} \\ \forall \lambda > 0, \forall x, p \in \mathbb{R}^n, \text{ the function } u \mapsto \langle p, u - x \rangle + \frac{1}{\lambda} D_h(u, x) \\ \text{has a unique minimizer.} \end{array} \right. \right\},$$

since all constraints in Assumption 1 about the domains of  $f$  and  $h$  become irrelevant. The general case when  $C$  is a convex subset of  $\mathbb{R}^n$  can be treated along the same approach. In fact, from the perspective of performance estimation, we can show that every problem in  $\mathcal{B}_L(C)$  can be reduced to some problem in  $\mathcal{B}_L(\mathbb{R}^n)$  with equivalent convergence rate (see Appendix A for details).

*Performance estimation problem.* Throughout this section, we fix a number of iterations  $N \geq 1$ , a relative smoothness parameter  $L > 0$ , and a step size  $\lambda > 0$ . In the currently known analyses of NoLips, worst-case guarantees have the following form

$$f(x_N) - f(x_*) \leq \theta(N, L, \lambda) D_h(x_*, x_0). \quad (17)$$

For instance, Theorem 1 states this result with  $\theta(N, L, \lambda) = 1/(\lambda N)$  when  $\lambda \in (0, 1/L]$  (note that since we consider the case where  $C = \mathbb{R}^n$ , we can take  $x_*$  in the bound as  $x_* \in \text{dom } h$  trivially). We then naturally seek the smallest  $\theta(N, L, \lambda)$  such that the bound (17) holds for any couple  $(f, h) \in \mathcal{B}_L(\mathbb{R}^n)$ , that is, solve the optimization problem

$$\begin{aligned} & \text{maximize} && (f(x_N) - f(x_*))/D_h(x_*, x_0) \\ & \text{subject to} && (f, h) \in \mathcal{B}_L(\mathbb{R}^n), \\ & && x_* \text{ is a minimizer of } f, \\ & && x_1, \dots, x_N \text{ are generated from } x_0 \text{ by Algorithm 1 with step size } \lambda, \end{aligned} \quad (\text{PEP})$$

in the variables  $f, h, x_0, \dots, x_N, x_*, n$ . We refer to this problem as a performance estimation problem (PEP). We use the convention  $0/0 = 0$ , so that the objective is well defined when  $x_* = x_0$ . Optimizing over the dimension  $n$  to get dimension-free bounds allows for the problem to admit efficient convex reformulations, as we see in the sequel. We look for guarantees that are independent of the kernel  $h$ , and therefore  $h$  is also an optimization variable.

Let us start by simplifying the problem. First, due to the strict convexity of  $h$ , the NoLips iteration (1) can be equivalently formulated via the first-order optimality condition

$$\nabla h(x_{i+1}) = \nabla h(x_i) - \lambda \nabla f(x_i) \quad \forall i \in \{0 \dots N-1\}$$

and, since the domain is  $\mathbb{R}^n$ , the condition that  $x_*$  minimizes  $f$  reduces to requiring  $\nabla f(x_*) = 0$ . Second, the problem is homogeneous in  $(f, h)$  (i.e., from a feasible couple  $(f, h)$ , take any constant  $c > 0$  and observe that the couple  $(cf, ch)$  is also feasible with the same objective value), hence optimizing the objective function  $f(x_N) - f(x_*)$  under the additional constraint  $D_h(x_*, x_0) = 1$  produces the same optimal value as the problem above.

Finally, we use the same argument as in [18, 37] and observe that the objective of (PEP) and the algorithmic constraints mentioned above depend solely on the values of the first-order oracles of  $f$  and  $h$  at the points  $x_0, \dots, x_N, x_*$ . Denoting  $I = \{0, 1, \dots, N, *\}$  the indices associated with the points involved in the problem, we proceed to write these values as

$$\begin{aligned} \{(f_i, g_i)\}_{i \in I} &= \{(f(x_i), \nabla f(x_i))\}_{i \in I}, \\ \{(h_i, s_i)\}_{i \in I} &= \{(h(x_i), \nabla h(x_i))\}_{i \in I}. \end{aligned}$$

Using those elements, the iterations of NoLips can be expressed as  $s_{i+1} = s_i - \lambda g_i$  for  $i \in \{0 \dots N-1\}$ , and the normalization constraint  $D_h(x_*, x_0) = 1$  becomes  $h_* - h_0 - \langle s_0, x_* - x_0 \rangle = 1$ .

Using those *discrete* representations of  $f$  and  $h$ , we can reformulate (PEP) as

$$\begin{aligned}
& \text{maximize } f_N - f_* \\
& \text{subject to } f_i = f(x_i), g_i = \nabla f(x_i), \\
& \quad h_i = h(x_i), s_i = \nabla h(x_i), \quad \text{for all } i \in I \text{ and some } (f, h) \in \mathcal{B}_L(\mathbb{R}^n), \\
& \quad g_* = 0, \\
& \quad s_{i+1} = s_i - \lambda g_i \quad \text{for } i \in \{1 \dots N-1\}, \\
& \quad h_* - h_0 - \langle s_0, x_* - x_0 \rangle = 1,
\end{aligned} \tag{PEP}$$

in the variables  $n, \{(x_i, f_i, g_i, h_i, s_i)\}_{i \in I}$ . The equivalence with the initial problem is guaranteed by the first two constraints which are called the *interpolation conditions*.

It turns out that interpolation conditions for the class  $\mathcal{B}_L(\mathbb{R}^n)$  are delicate to establish, due to assumptions on  $h$ . Fortunately, there exist two classes  $\underline{\mathcal{B}}_L(\mathbb{R}^n)$  and  $\overline{\mathcal{B}}_L(\mathbb{R}^n)$  for which they can be derived. The first class is a restriction of  $\mathcal{B}_L(\mathbb{R}^n)$  where  $f$  and  $Lh - f$  are both assumed to be strictly convex:

$$\underline{\mathcal{B}}_L(\mathbb{R}^n) = \mathcal{B}_L(\mathbb{R}^n) \cap \{(f, h) : \mathbb{R}^n \rightarrow \mathbb{R} \mid f \text{ and } Lh - f \text{ are strictly convex}\}$$

whereas the second class consists in considering a relaxation with possibly nonsmooth functions:

$$\overline{\mathcal{B}}_L(\mathbb{R}^n) = \{(f, h) : \mathbb{R}^n \rightarrow \mathbb{R} \mid f \text{ and } Lh - f \text{ are convex}\}.$$

The following inclusions then directly hold

$$\underline{\mathcal{B}}_L(\mathbb{R}^n) \subset \mathcal{B}_L(\mathbb{R}^n) \subset \overline{\mathcal{B}}_L(\mathbb{R}^n).$$

With these classes, we can now define two easier problems. The first one is a restriction of (PEP) defined on the class  $\underline{\mathcal{B}}_L(\mathbb{R}^n)$ , under the additional constraint that all iterates are distinct:

$$\begin{aligned}
& \text{maximize } f_N - f_* \\
& \text{subject to } f_i = f(x_i), g_i = \nabla f(x_i), \\
& \quad h_i = h(x_i), s_i = \nabla h(x_i), \quad \text{for all } i \in I \text{ and some } (f, h) \in \underline{\mathcal{B}}_L(\mathbb{R}^n), \\
& \quad g_* = 0, \\
& \quad s_{i+1} = s_i - \lambda g_i \quad \text{for } i \in \{1 \dots N-1\}, \\
& \quad h_* - h_0 - \langle s_0, x_* - x_0 \rangle = 1, \\
& \quad x_i \neq x_j \quad \text{for } i \neq j \in I,
\end{aligned} \tag{PEP}$$

in the variables  $n, \{(x_i, f_i, g_i, h_i, s_i)\}_{i \in I}$ . The second problem is a relaxation of (PEP), where  $(f, h) \in \overline{\mathcal{B}}_L(\mathbb{R}^n)$  are possibly nonsmooth and  $g_i, s_i$  are thus *subgradients*:

$$\begin{aligned}
& \text{maximize } f_N - f_* \\
& \text{subject to } f_i = f(x_i), g_i \in \partial f(x_i), \\
& \quad h_i = h(x_i), s_i \in \partial h(x_i), \\
& \quad Ls_i - g_i \in \partial(Lh - f)(x_i) \quad \text{for all } i \in I \text{ and some } (f, h) \in \overline{\mathcal{B}}_L(\mathbb{R}^n), \\
& \quad g_* = 0, \\
& \quad s_{i+1} = s_i - \lambda g_i \quad \text{for } i \in \{1 \dots N-1\}, \\
& \quad h_* - h_0 - \langle s_0, x_* - x_0 \rangle = 1,
\end{aligned} \tag{\overline{PEP}}$$

in the variables  $n, \{(x_i, f_i, g_i, h_i, s_i)\}_{i \in I}$ . We added the technical constraint  $Ls_i - g_i \in \partial(Lh - f)(x_i)$ , which is redundant for differentiable functions; but that is necessary in order to establish interpolation conditions in the nonsmooth case.

Because of the inclusions between the feasible sets of these problems, we naturally have

$$\text{val}(\underline{\text{PEP}}) \leq \text{val}(\text{PEP}) \leq \text{val}(\overline{\text{PEP}}).$$

We prove in the sequel that  $\overline{\text{PEP}}$  can be solved via a semidefinite program and that  $\text{val}(\underline{\text{PEP}}) = \text{val}(\overline{\text{PEP}})$  (Theorem 4), allowing to reach our claims.

Note that the relaxed problem  $\overline{\text{PEP}}$  does not correspond to any practical algorithm, as NoLips is not properly defined for nonsmooth functions  $h$ . However, we see in the sequel that feasible points of this problem correspond to accumulation points of (PEP). In other words, instances of NoLips can get arbitrarily close to pathological nonsmooth functions whose behaviors are captured by  $\overline{\text{PEP}}$ .



In the following sections, we show that problems [\(PEP\)](#) and [\(PEP\)](#) can be cast as semidefinite programs (SDP) [\[40\]](#) and solved numerically using standard packages [\[28, 25\]](#). The main ingredient consists in showing that interpolation constraints can actually be expressed using quadratic inequalities, as detailed in the next section.

#### 4.2 Interpolation involving differentiability and strict convexity

In this section, we show how to reformulate interpolation constraints for [\(PEP\)](#) and [\(PEP\)](#) as quadratic inequalities. We start by recalling interpolation conditions for the class of  $L$ -smooth and  $\mu$ -strongly convex functions.

**Theorem 3 (Smooth strongly convex interpolation, [\[37\]](#))** *Let  $I$  be a finite index set,  $\{(x_i, f_i, g_i)\}_{i \in I} \in (\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n)^{|I|}$  and  $0 \leq \mu \leq L \leq +\infty$ . The following statements are equivalent:*

(i) *There exists a proper closed convex function  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  such that  $f$  is  $\mu$ -strongly convex, has a  $L$ -Lipschitz continuous gradient and*

$$f_i = f(x_i), g_i \in \partial f(x_i) \quad \forall i \in I.$$

(ii) *For every  $i, j \in I$  we have*

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

In particular, when  $L = +\infty$  (meaning that we require no smoothness) and  $\mu = 0$ , those conditions reduce to the simpler *convex interpolation* conditions, reminiscent of subgradient inequalities:

$$f_i - f_j - \langle g_j, x_i - x_j \rangle \geq 0 \tag{18}$$

In our setting, we want to avoid working with smoothness and strong convexity, so we provide interpolation conditions for the class of differentiable strictly convex functions.

**Proposition 3 (Differentiable and strictly convex interpolation)** *Let  $I$  be a finite index set and  $\{(x_i, f_i, g_i)\}_{i \in I} \in (\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n)^{|I|}$ . The following statements are equivalent:*

(i) *There exists a convex function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  such that  $f$  is differentiable, strictly convex and*

$$f_i = f(x_i), g_i = \nabla f(x_i) \quad \forall i \in I.$$

(ii) *For every  $i, j \in I$  we have*

$$\begin{cases} f_i - f_j - \langle g_j, x_i - x_j \rangle > 0 & \text{if } x_i \neq x_j, \\ f_i = f_j \text{ and } g_i = g_j & \text{otherwise.} \end{cases} \tag{19}$$

*Proof* (i)  $\implies$  (ii). Assume that (i) holds, and choose such a function  $f$ . The first inequality of [\(19\)](#) follows from strict convexity of  $f$ , and the second line is a consequence of the fact that a differentiable convex function has a unique subgradient at each point [\[34, Thm 25.1\]](#).

(ii)  $\implies$  (i). Assume (ii). If for all  $i, j \in I$ , we have  $g_i = g_j$  and  $x_i = x_j$ , then there is only one point and one subgradient to be interpolated, and the result follows immediately from considering a well-chosen definite quadratic function. In the other case, define

$$\nu = \min_{\substack{i, j \in I \\ x_i \neq x_j}} f_i - f_j - \langle g_j, x_i - x_j \rangle.$$

Because of [\(19\)](#) and the finiteness of  $I$ , we have that  $\nu > 0$ . Now, define  $r$  as

$$r = \max_{i, j \in I} \|g_i - g_j\|^2 + \|x_i - x_j\|^2$$

so that  $r > 0$ . Condition [\(19\)](#) implies that for all  $i, j \in I$  we have

$$f_i - f_j - \langle g_j, x_i - x_j \rangle \geq \frac{\nu}{r} (\|g_i - g_j\|^2 + \|x_i - x_j\|^2). \tag{20}$$

Indeed, if  $x_i \neq x_j$ , this follows from the definition of  $\nu$  and  $r$ . If  $x_i = x_j$  both sides of the inequality are 0 because of the second line in (19). Let us choose two constants  $0 < \mu < L < +\infty$  such that

$$\frac{1}{L - \mu} \leq \frac{\nu}{r}, \quad \frac{\mu}{1 - \mu/L} \leq \frac{\nu}{r},$$

which is possible as it suffices to take  $L$  large enough and  $\mu$  small enough. We now proceed to show that the interpolation conditions of Theorem 3 hold with the constants  $\mu, L$  defined above. Using the inequality  $\|u - v\|^2 \leq 2\|u\|^2 + 2\|v\|^2$  and (20), we get that for all  $i, j$ ,

$$\begin{aligned} & \frac{1}{2L} \|g_i - g_j\|^2 + \frac{\mu}{2(1 - \mu/L)} \|x_i - x_j - \frac{1}{L}(g_i - g_j)\|^2 \\ & \leq \left( \frac{1}{2L} + \frac{\mu}{L(L - \mu)} \right) \|g_i - g_j\|^2 + \frac{\mu}{1 - \mu/L} \|x_i - x_j\|^2 \\ & \leq \left( \frac{1}{L} + \frac{\mu}{L(L - \mu)} \right) \|g_i - g_j\|^2 + \frac{\mu}{1 - \mu/L} \|x_i - x_j\|^2 \\ & = \frac{1}{L - \mu} \|g_i - g_j\|^2 + \frac{\mu}{1 - \mu/L} \|x_i - x_j\|^2 \\ & \leq \frac{\nu}{r} \|g_i - g_j\|^2 + \frac{\nu}{r} \|x_i - x_j\|^2 \\ & \leq f_i - f_j - \langle g_j, x_i - x_j \rangle. \end{aligned}$$

Under those conditions, Theorem 3 states that there exists a convex function  $f$  that interpolates  $\{(x_i, f_i, g_i)\}_{i \in I}$  which is  $\mu$ -strongly convex and has  $L$ -Lipschitz continuous gradients. A fortiori, since  $\mu > 0$  and  $L < \infty$ ,  $f$  is differentiable and strictly convex. Finally,  $f$  is finite on  $\mathbb{R}^n$  since it is  $L$ -smooth. ■

Using these results, we can now formulate interpolation conditions for the problems ( $\overline{\text{PEP}}$ ) and ( $\text{PEP}$ ) involving the classes  $\overline{\mathcal{B}}_L(\mathbb{R}^n)$  and  $\underline{\mathcal{B}}_L(\mathbb{R}^n)$  that were defined in Section 4.1.

**Corollary 1 (Interpolation conditions for ( $\overline{\text{PEP}}$ ))** *Let  $I$  be a finite index set and  $\{(x_i, f_i, g_i, h_i, s_i)\}_{i \in I} \in (\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n)^{|I|}$ . The following statements are equivalent.*

(i) *There exist functions  $(f, h) \in \overline{\mathcal{B}}_L(\mathbb{R}^n)$  such that*

$$\begin{aligned} f_i &= f(x_i), \quad g_i \in \partial f(x_i), \\ h_i &= h(x_i), \quad s_i \in \partial h(x_i), \\ Ls_i - g_i &\in \partial(Lh - f)(x_i). \end{aligned}$$

(ii) *For all  $i, j \in I$  such that  $i \neq j$ , we have*

$$\begin{aligned} f_i - f_j - \langle g_j, x_i - x_j \rangle &\geq 0, \\ (Lh_i - f_i) - (Lh_j - f_j) - \langle Ls_j - g_j, x_i - x_j \rangle &\geq 0. \end{aligned} \tag{21}$$

*Proof* (i)  $\implies$  (ii) follows immediately from the definition of a subgradient applied to convex functions  $f$  and  $Lh - f$ .

Assume that (ii) holds. By the specialization of (18) in Theorem 3, conditions (ii) imply that there exist two convex functions  $f, d : \mathbb{R}^n \rightarrow \mathbb{R}$  such that

$$\begin{aligned} f_i &= f(x_i), \quad g_i \in \partial f(x_i), \\ Lh_i - f_i &= d(x_i), \quad Ls_i - g_i \in \partial d(x_i). \end{aligned}$$

Defining the convex function  $h = (f + d)/L$ , we have that  $d = Lh - f$ , hence  $Ls_i - g_i \in \partial(Lh - f)(x_i)$ . We also get

$$h_i = h(x_i), \quad s_i \in \partial h(x_i),$$

where we used the fact that  $Ls_i \in \partial f(x_i) + \partial d(x_i) \subset \partial(f + d)(x_i) = L\partial h(x_i)$  (see [34, Thm 23.8] for the subdifferential of a sum of convex functions). Hence (i) holds. ■

**Corollary 2 (Interpolation conditions for (PEP))** Let  $I$  be a finite index set and  $\{(x_i, f_i, g_i, h_i, s_i)\}_{i \in I} \in (\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n)^{|I|}$ . Assume that  $x_i \neq x_j$  for every  $i \neq j \in I$ . The following statements are equivalent.

(i) There exist functions  $(f, h) \in \underline{\mathcal{B}}_L(\mathbb{R}^n)$  such that

$$\begin{aligned} f_i &= f(x_i), g_i = \nabla f(x_i), \\ h_i &= h(x_i), s_i = \nabla h(x_i). \end{aligned}$$

(ii) For all  $i, j \in I$  such that  $i \neq j$  we have

$$\begin{aligned} f_i - f_j - \langle g_j, x_i - x_j \rangle &> 0, \\ (Lh_i - f_i) - (Lh_j - f_j) - \langle Ls_j - g_j, x_i - x_j \rangle &> 0. \end{aligned} \tag{22}$$

*Proof* Note that since we assumed  $x_i \neq x_j$  for every  $i \neq j$ , interpolation conditions of Proposition 3 reduce to requiring a strict inequality in (19) for every  $i \neq j$ . As before, define  $d := Lh - f$ . Then since  $(f, h) \in \underline{\mathcal{B}}_L(\mathbb{R}^n)$  the functions  $f$  and  $d$  are differentiable strictly convex, hence (i)  $\implies$  (ii) follows simply from strict convexity of these functions.

Conversely, assume (ii). By using Proposition 3 again, we can interpolate differentiable strictly convex functions  $f$  and  $d$  and recover  $h$  with  $h = (f + d)/L$ , thus we have naturally  $Lh - f$  convex. The function  $h$  is thus also differentiable and strictly convex. Moreover, it can be seen from the proof of Proposition 3 that the interpolating functions can actually be chosen strongly convex, hence with this choice the well-posedness condition Assumption 1(iii) holds, and we can conclude that  $(f, h) \in \underline{\mathcal{B}}_L(\mathbb{R}^n)$ .  $\blacksquare$

#### 4.3 Semidefinite reformulations

Now that we established the interpolation conditions for (PEP) and ( $\overline{\text{PEP}}$ ), we may use them to obtain semidefinite performance estimation formulations as in [18, 37]. This is made possible by observing that interpolation conditions (21)-(22) are quadratic inequalities in the problem variables.

Let  $\{(x_i, f_i, g_i, h_i, s_i)\}_{i \in I}$  be a feasible point of one of the PEPs in dimension  $n$ . We write  $G \in \mathbf{S}_{3(N+2)}$  the Gram matrix that contains all dot products between  $x_i, g_i, s_i$  for  $i \in I$ , with

$$G = \begin{pmatrix} G^{xx} & G^{gx} & G^{sx} \\ G^{gx\top} & G^{gg} & G^{gs} \\ G^{sx\top} & G^{gs\top} & G^{ss} \end{pmatrix} \succeq 0$$

whose size is independent of the dimension  $n$ , where the blocks are defined as

$$G_{ij}^{xx} = \langle x_i, x_j \rangle, G_{ij}^{gx} = \langle g_i, x_j \rangle, G_{ij}^{gs} = \langle g_i, s_j \rangle, G_{ij}^{gg} = \langle g_i, g_j \rangle, G_{ij}^{sx} = \langle s_i, x_j \rangle, G_{ij}^{ss} = \langle s_i, s_j \rangle, \quad i, j \in I.$$

Denote by

$$F = (f_0, \dots, f_N, f_*) \in \mathbb{R}^{N+2}, \quad H = (h_0, \dots, h_N, h_*) \in \mathbb{R}^{N+2},$$

the vectors representing the function values of  $f, h$  at the iterates. Finally observe that all the constraints of (PEP) and ( $\overline{\text{PEP}}$ ) can be expressed using only  $G, F$  and  $H$ .

For instance, interpolation conditions (21) for  $\underline{\mathcal{B}}_L(\mathbb{R}^n)$  rewrite for all  $i, j \in I$  as

$$\begin{aligned} f_i - f_j - G_{ji}^{gx} + G_{jj}^{gx} &\geq 0, \\ (Lh_i - f_i) - (Lh_j - f_j) - L(G_{ji}^{sx} - G_{jj}^{sx}) + G_{ji}^{gx} - G_{jj}^{gx} &\geq 0. \end{aligned}$$

This allows us to reformulate the relaxation ( $\overline{\text{PEP}}$ ) as a semidefinite program, written

$$\begin{aligned} &\text{maximize } f_N - f_* \\ &\text{subject to } f_i - f_j - G_{ji}^{gx} + G_{jj}^{gx} \geq 0, \\ &\quad (Lh_i - f_i) - (Lh_j - f_j) - L(G_{ji}^{sx} - G_{jj}^{sx}) + G_{ji}^{gx} - G_{jj}^{gx} \geq 0 \quad \text{for } i, j \in I, \\ &\quad G_{**}^{gg} = 0, \\ &\quad G_{i+1, j}^{sx} = G_{ij}^{sx} - \lambda G_{ij}^{gx} \quad \text{for } i \in \{0 \dots N-1\}, j \in I, \\ &\quad h_* - h_0 - G_{0*}^{sx} + G_{00}^{sx} = 1, \\ &\quad G \succeq 0, \end{aligned} \tag{sdp- $\overline{\text{PEP}}$ }$$

in the variables  $G \in \mathbf{S}_{3(N+2)}$  and  $F, H \in \mathbb{R}^{N+2}$ .

Any feasible point of  $(\overline{\text{PEP}})$  can be cast into an admissible point of  $(\text{sdp-}\overline{\text{PEP}})$  by computing the semidefinite Gram matrix  $G$ . Conversely, if  $G, F, H$  is an admissible point of  $(\text{sdp-}\overline{\text{PEP}})$ , then the vectors  $\{(x_i, g_i, s_i)\}_{i \in I}$  can be recovered by performing, for instance, a Cholesky decomposition of  $G$ . Note that we expressed the algorithmic constraint  $s_{i+1} = s_i - \lambda g_i$  only through scalar products with the  $x_i$ 's in the SDP, since only the projection of the gradients on  $\text{Span}(\{x_i\}_{i \in I})$  is relevant in the PEPs. Because interpolation conditions from Corollary 1 are necessary and sufficient, we conclude that the problems are equivalent, that is

$$\text{val}(\text{sdp-}\overline{\text{PEP}}) = \text{val}(\overline{\text{PEP}}).$$

The rank of  $G$  determines the dimension of the interpolated problem. If we look instead for a solution that has a given dimension  $n$ , this would mean imposing a nonconvex rank constraint on  $G$ . Our formulation, on the other hand, is convex and finds the best convergence bound that is dimension-independent, which is a usual requirement for *large-scale settings*. In our setting, given the size of  $G$  and the algorithmic constraints, we can show that there exists worst-case instances of dimensions at most  $2N + 5$ . For NoLips, we show in the sequel that it is even possible to find simple worst-cases in a single dimension.

In the same way, the value of  $(\text{PEP})$  can be computed as

$$\begin{aligned} & \text{maximize } f_N - f_* \\ & \text{subject to } f_i - f_j - G_{ji}^{gx} + G_{jj}^{gx} > 0, \\ & \quad (Lh_i - f_i) - (Lh_j - f_j) - L(G_{ji}^{sx} - G_{jj}^{sx}) + G_{ji}^{gx} - G_{jj}^{gx} > 0 \quad \text{for } i \neq j \in I, \\ & \quad G_{**}^{gg} = 0, \\ & \quad G_{i+1,j}^{sx} = G_{ij}^{sx} - \lambda G_{ij}^{gx} \quad \text{for } i \in \{0 \dots N-1\}, j \in I, \\ & \quad h_* - h_0 - G_{0*}^{sx} + G_{00}^{sx} = 1, \\ & \quad G_{ii}^{xx} + G_{jj}^{xx} - 2G_{ij}^{xx} > 0 \quad \text{for } i \neq j \in I, \\ & \quad G \succeq 0, \end{aligned} \tag{sdp-PEP}$$

in the variables  $G \in \mathbf{S}_{3(N+2)}$  and  $F, H \in \mathbb{R}^{N+2}$ , where we used interpolation conditions for  $\mathcal{B}_{\mathcal{L}}(\mathbb{R}^n)$  from Corollary 2, since all points  $\{x_i\}_{i \in I}$  are constrained to be distinct. Therefore, as above we infer that

$$\text{val}(\text{sdp-PEP}) = \text{val}(\text{PEP}).$$

Recalling the hierarchy between the problems, we thus have

$$\text{val}(\text{sdp-PEP}) \leq \text{val}(\text{PEP}) \leq \text{val}(\text{sdp-}\overline{\text{PEP}}).$$

By comparing the two semidefinite programs stated above, one can notice that the only difference is that  $(\text{sdp-PEP})$  imposes some inequalities of  $(\text{sdp-}\overline{\text{PEP}})$  to be strict. In the next section, we use topological arguments to prove that the values of the two problems are actually equal. In fact, strict inequalities have little meaning in numerical optimization (the value of  $(\text{sdp-PEP})$  is actually a supremum and not a maximum); in our experiments, we focus on  $(\text{sdp-}\overline{\text{PEP}})$  as solvers usually admit only closed feasible sets.

#### 4.4 Tightness of the approach: nonsmooth limit behaviors

We are now ready to prove the main result of this section.

**Theorem 4** *The value of the performance estimation problem  $(\text{PEP})$  for NoLips is equal to the value of the nonsmooth relaxation  $(\overline{\text{PEP}})$ , which can be computed by solving the semidefinite program  $(\text{sdp-}\overline{\text{PEP}})$ .*

*Proof* We show that the closure of the feasible set of  $(\text{sdp-PEP})$  is the feasible set of  $(\text{sdp-}\overline{\text{PEP}})$ . We first need to prove that the strengthened problem  $(\overline{\text{PEP}})$  is feasible, by finding an instance of NoLips where  $f$  and  $Lh - f$  are strictly convex and such that all iterates are distinct. It suffices for instance to consider two one-dimensional quadratic functions. Define  $f, h : \mathbb{R} \rightarrow \mathbb{R}$  with

$$f(x) = \frac{\alpha}{2}x^2, \quad h(x) = \frac{1}{2}x^2 \quad \text{where } \alpha = \min\left(\frac{1}{2\lambda}, \frac{L}{2}\right).$$

Then  $f$  is strictly convex and so is  $Lh - f = \frac{L-\alpha}{2}x^2$  since  $L - \alpha \geq \frac{L}{2} > 0$ . The optimum is  $x_* = 0$ . Choose

$$x_0 = \sqrt{2}$$

for which we have  $D_h(x_*, x_0) = x_0^2/2 = 1$ . Then, Algorithm 1 is equivalent to gradient descent and the iterates satisfy

$$x_N = (1 - \lambda\alpha)^N x_0.$$

Since  $\alpha\lambda \leq 1/2 < 1$ , all the iterates are distinct and therefore we constructed a feasible point of (PEP). Let us therefore write  $(G, F, H)$  a corresponding feasible point of (sdp-PEP), and  $(\bar{G}, \bar{F}, \bar{H})$  a feasible point of (sdp- $\overline{\text{PEP}}$ ). Define the sequence  $\{(G^k, F^k, H^k)\}_{k \geq 1}$  as

$$\begin{aligned} G^k &= \frac{1}{k}G + (1 - \frac{1}{k})\bar{G}, \\ F^k &= \frac{1}{k}F + (1 - \frac{1}{k})\bar{F}, \\ H^k &= \frac{1}{k}H + (1 - \frac{1}{k})\bar{H}. \end{aligned}$$

Then, for every  $k \geq 1$ ,  $(G^k, F^k, H^k)$  is still a feasible point of (sdp-PEP), because of convexity of the constraints and the fact that adding a strict inequality to a weak inequality gives a strict inequality. Moreover, the sequence converges to the point  $(\bar{G}, \bar{F}, \bar{H})$  when  $k \rightarrow +\infty$ .

Hence we proved that for any feasible point of (sdp-PEP), there is a sequence of admissible points of (sdp-PEP) that converge to it. Since the objective is linear in the vector  $F$  therefore continuous, we deduce that the two problems have the same value:

$$\text{val}(\text{sdp-PEP}) = \text{val}(\text{sdp-}\overline{\text{PEP}}),$$

which means that  $\text{val}(\text{PEP}) = \text{val}(\overline{\text{PEP}})$ . As  $\text{val}(\text{PEP})$  lies in between these two values, we conclude that they are all equal. ■

Theorem 4 states that the value of the original problem (PEP) can be computed numerically with a semidefinite solver applied to (sdp- $\overline{\text{PEP}}$ ). The result itself also helps us gain some theoretical insight: it tells us that the worst-case for NoLips might be reached as  $(f, h)$  approach possibly pathological limiting nonsmooth functions in  $\overline{\mathcal{B}_L(\mathbb{R}^n)}$ .

Observe also that we focused on presenting the PEP for the class  $\mathcal{B}_L(\mathbb{R}^n)$  to avoid technicalities related to the domain of definition. However, we show in Appendix A that the exact same problem (sdp- $\overline{\text{PEP}}$ ) also solves the performance estimation problem for NoLips on the general class  $\mathcal{B}_L(C)$ , for any nonempty closed convex set  $C$ .

#### 4.5 Numerical evidence and computer-assisted proofs

We now provide several applications of the performance estimation framework that we developed for Bregman methods.

##### 4.5.1 Solving (PEP) for finding the exact worst-case convergence rate of NoLips

We first start by the most direct application, that is finding the exact worst-case performance of NoLips. Theorem 4 states that it can be computed by solving the semidefinite program (sdp- $\overline{\text{PEP}}$ ). The link to the MATLAB implementation is provided in Section 5.

To simplify our setting, note that we can assume without loss of generality that the relative smoothness constant  $L$  is 1, since we can replace  $h$  by a scaled version  $Lh$ . Recall that we know from Theorem 1, that

$$\text{val}(\text{PEP}) \leq \frac{1}{\lambda N}.$$

Table 1 shows the result of solving (sdp- $\overline{\text{PEP}}$ ) for several values of  $N$  up to 100, for a step size  $\lambda = 1$ . We observe that with high precision,  $\text{val}(\text{sdp-}\overline{\text{PEP}})$  is equal to the theoretical bound  $1/(\lambda N)$ .

**Table 1** Numerical value of the performance estimation problem (PEP) with  $\lambda = 1$ ,  $L = 1$ . *Rel. error* denotes the relative error between  $\text{val}(\text{PEP})$  and the theoretical bound of  $1/N$  given by Theorem 1. *Primal feasibility* corresponds to the maximal absolute value of constraint violation returned by the MOSEK solver.

N	val(PEP)	Rel. error	Primal feasibility
1	1.000	1.8e-11	4.3e-10
2	0.500	1.8e-8	2.8e-9
3	0.333	1.8e-8	2.8e-9
4	0.250	4.9e-8	2.3e-8
5	0.200	1.8e-10	6.4e-11
10	0.100	6.4e-11	1.3e-11
20	0.050	1.1e-8	1.9e-10
50	0.020	6.5e-6	5.0e-7
100	0.01	7.2e-5	1.6e-6

*Other values of  $\lambda$ .* One can wonder how the numerical value evolves when one varies the step size  $\lambda$ . Our experimental observations are as follows:

- For any  $\lambda \in (0, 1/L]$ ,  $\text{val}(\text{PEP})$  is exactly equal to the theoretical bound  $1/(\lambda N)$ .
- For any  $\lambda > 1/L$ ,  $\text{val}(\text{PEP}) = +\infty$ , hence Algorithm 1 does not converge in general with these step size values. This suggests that the maximal step size value allowed for NoLips is indeed  $1/L$ , unlike the Euclidean setting where gradient descent can be applied with a step size that goes up to  $2/L$ .

While results above suggest that  $1/(\lambda N)$  is the exact worst-case rate of NoLips, they provide only numerical evidence. We can however use them to deduce formal guarantees, both for proving an *upper bound* and a *lower bound*.

*Upper bound guarantee through duality.* As noticed in previous work on PEPs [18, 35], solving the dual of (sdp-PEP) can be used to deduce a proof. Indeed, the dual solution gives a combination of the constraints that, when transposed to analytical form, leads to a formal guarantee. This provides the following proof for the  $O(1/k)$  convergence rate of Theorem 1.

*Proof of Theorem 1* The proof relies on the fact that, since  $Lh - f$  is convex we have that  $\frac{1}{\lambda}h - f$  is convex for any  $\lambda \in (0, \frac{1}{L}]$ , and only consists in performing the following weighted sum of inequalities:

- convexity of  $f$ , between  $u$  and  $x_i$  ( $i = 0, \dots, k$ ) with weights  $\gamma_{*,i} = \frac{1}{k}$ :

$$f(u) \geq f(x_i) + \langle \nabla f(x_i), u - x_i \rangle,$$

- convexity of  $f$ , between  $x_i$  and  $x_{i+1}$  ( $i = 0, \dots, k-1$ ) with weights  $\gamma_{i,i+1} = \frac{i}{k}$ :

$$f(x_i) \geq f(x_{i+1}) + \langle \nabla f(x_{i+1}), x_i - x_{i+1} \rangle,$$

- convexity of  $\frac{1}{\lambda}h - f$ , between  $u$  and  $x_k$  with weight  $\mu_{*,k} = \frac{1}{k}$ :

$$\frac{1}{\lambda}h(u) - f(u) \geq \frac{1}{\lambda}h(x_k) - f(x_k) + \langle \frac{1}{\lambda}\nabla h(x_k) - \nabla f(x_k), u - x_k \rangle,$$

- convexity of  $\frac{1}{\lambda}h - f$ , between  $x_{i+1}$  and  $x_i$  ( $i = 0, \dots, k-1$ ) with weight  $\mu_{i+1,i} = \frac{i+1}{k}$

$$\frac{1}{\lambda}h(x_{i+1}) - f(x_{i+1}) \geq \frac{1}{\lambda}h(x_i) - f(x_i) + \langle \frac{1}{\lambda}\nabla h(x_i) - \nabla f(x_i), x_{i+1} - x_i \rangle,$$

- convexity of  $\frac{1}{\lambda}h - f$ , between  $x_i$  and  $x_{i+1}$  ( $i = 0, \dots, k-1$ ) with weight  $\mu_{i,i+1} = \frac{i}{k}$

$$\frac{1}{\lambda}h(x_i) - f(x_i) \geq \frac{1}{\lambda}h(x_{i+1}) - f(x_{i+1}) + \langle \frac{1}{\lambda}\nabla h(x_{i+1}) - \nabla f(x_{i+1}), x_i - x_{i+1} \rangle.$$

The weighted sum is written as

$$\begin{aligned}
0 &\geq \sum_{i=0}^k \gamma_{*,i} [f(x_i) - f(u) + \langle \nabla f(x_i), u - x_i \rangle] \\
&\quad + \sum_{i=0}^{k-1} \gamma_{i,i+1} [f(x_{i+1}) - f(x_i) + \langle \nabla f(x_{i+1}), x_i - x_{i+1} \rangle] \\
&\quad + \mu_{*,k} \left[ \frac{1}{\lambda} h(x_k) - f(x_k) - \left( \frac{1}{\lambda} h(u) - f(u) \right) + \langle \frac{1}{\lambda} \nabla h(x_k) - \nabla f(x_k), u - x_k \rangle \right] \\
&\quad + \sum_{i=0}^{k-1} \mu_{i+1,i} \left[ \frac{1}{\lambda} h(x_i) - f(x_i) - \left( \frac{1}{\lambda} h(x_{i+1}) - f(x_{i+1}) \right) + \langle \frac{1}{\lambda} \nabla h(x_i) - \nabla f(x_i), x_{i+1} - x_i \rangle \right] \\
&\quad + \sum_{i=0}^{k-1} \mu_{i,i+1} \left[ \frac{1}{\lambda} h(x_{i+1}) - f(x_{i+1}) - \left( \frac{1}{\lambda} h(x_i) - f(x_i) \right) + \langle \frac{1}{\lambda} \nabla h(x_{i+1}) - \nabla f(x_{i+1}), x_i - x_{i+1} \rangle \right].
\end{aligned}$$

By substitution of  $\nabla h(x_{i+1}) = \nabla h(x_i) - \lambda \nabla f(x_i)$  ( $i = 0, \dots, k-1$ ), one can reformulate the weighted sum exactly as (i.e., there is no residual):

$$0 \geq f(x_k) - f(u) - \frac{h(u) - h(x_0) - \langle \nabla h(x_0), u - x_0 \rangle}{\lambda k},$$

yielding the desired result. ■

*Lower bound through worst-case functions.* As (PEP) computes the *exact* worst-case performance of NoLips, experiments above suggest that  $1/(\lambda N)$  is also a lower bound, meaning that for every  $\epsilon > 0$ , there exist functions  $(f, h) \in \mathcal{B}_L$  such that the iterates of NoLips satisfy

$$f(x_N) - f_* \geq \frac{D_h(x_*, x_0)}{\lambda N} - \epsilon.$$

We detail here how such functions can be constructed from the solution of (sdp-PEP). The numerical solver allows us to find a maximizer  $\bar{G}, \bar{F}, \bar{H}$  (recall that only the relaxed problem has a maximizer as the feasible set is closed), and by factorizing the matrix  $G$  as  $P^T P$ , we can thus recover the corresponding discrete representation  $\{\bar{x}_i, \bar{g}_i, \bar{f}_i, \bar{h}_i, \bar{s}_i\}_{i \in I}$ . This discretization can in turn be interpolated to get the corresponding functions  $(\bar{f}, \bar{h}) \in \overline{\mathcal{B}_L}$ . There are multiple ways to perform this interpolation; see [37, Thm. 1] for a constructive approach.

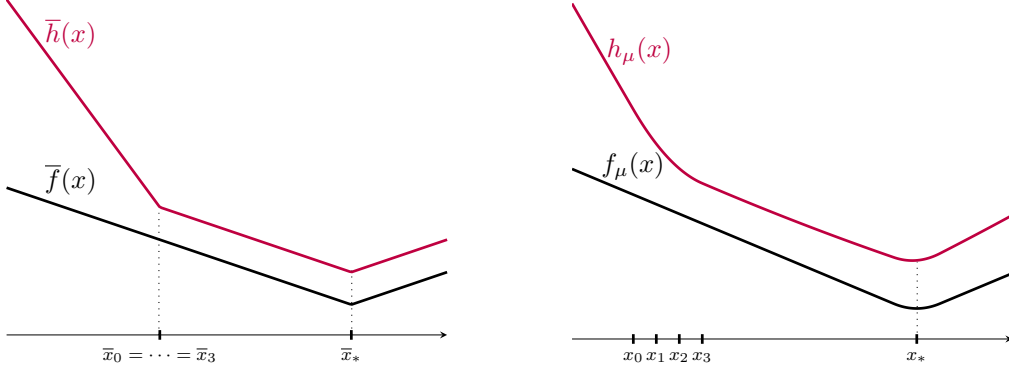
Recall that since functions  $(\bar{f}, \bar{h})$  yield a solution to (PEP), they belong to  $\overline{\mathcal{B}_L}$  and might thus form a *pathological* nonsmooth limiting worst-case. They can be approached by valid instances  $(f_\mu, h_\mu) \in \mathcal{B}_L$  by performing for instance smoothing through Moreau envelopes (as in Section 3.2) and adding a small quadratic to  $h$  to make it strictly convex.

There are however many possible maximizers of (sdp-PEP). If we seek a low-dimensional example that may be easily interpretable, we can search for a maximizer such that the Gram matrix  $G$  has minimal rank. Using rank minimization heuristics, we were able to find one-dimensional worst-case functions. Fix a number of iterations  $N \geq 1$ , assume  $\lambda = 1/L = 1$  and define  $\bar{f}, \bar{h} : \mathbb{R} \rightarrow \mathbb{R}$  as

$$\begin{aligned}
\bar{f}(x) &= |x - 1|, \\
\bar{h}(x) &= \bar{f}(x) + \max(-Nx, 0),
\end{aligned}$$

and set  $\bar{x}_0 = 0, \bar{x}_* = 1$ . Then clearly  $(\bar{f}, \bar{h}) \in \overline{\mathcal{B}_L}(\mathbb{R})$ . Figure 2 shows the functions  $\bar{f}, \bar{h}$  as well as their smoothed versions  $(f_\mu, h_\mu) \in \mathcal{B}_L(\mathbb{R})$ . Note that the pathological behavior also reflects in the iterates: in the limiting instance, all iterates  $\bar{x}_0, \dots, \bar{x}_N$  are equal. In the smoothed version, iterates are distinct (since  $h_\mu$  is strictly convex), but they get closer and closer as the smoothing parameter  $\mu$  goes to 0.

The smoothed function  $f_\mu$  is a Huber function, which is also the worst-case instance for Euclidean gradient descent on  $L$ -smooth functions described in [37]. This analysis could be formalized to prove the  $1/k$  lower bound for NoLips; however, this bound is just a particular case of the stronger result for general Bregman gradient methods derived in Section 3.2.



**Fig. 2** Worst-case functions for NoLips in dimension 1 with  $N = 3$  iterations. The left figure shows the limiting instance  $(\bar{f}, \bar{h}) \in \overline{\mathcal{B}}_L(\mathbb{R})$ , while the right plot represents the smooth approximation by a valid instance  $(f_\mu, h_\mu) \in \mathcal{B}_L(\mathbb{R})$ , with smoothing parameter  $\mu = 0.1$ . As  $\mu$  goes to 0, functions  $f_\mu, h_\mu$  tend to a pathological behavior where all iterates are equal and for which we have exactly  $\bar{f}(\bar{x}_N) - f_* = D_{\bar{h}}(\bar{x}_*, \bar{x}_0)/N$ .

#### 4.5.2 Extension to other criteria

In our performance estimation problem, we focused on studying bounds of the form  $f(x_N) - f_* \leq \theta(N, L, \lambda)D_h(x_*, x_0)$ . However, we are not limited to this criterion, and different convergence measures might be considered by changing the objective and constraints in (PEP). For instance, another popular criterion is the stationarity measure  $D_h(x_k, x_{k+1})$ , which boils down to the squared gradient norm in the unconstrained Euclidean case. By adapting (PEP), we get the following new convergence result for NoLips.

**Proposition 4 (NoLips convergence rate, take II)** *Let  $L > 0$ ,  $C$  be a nonempty closed convex subset of  $\mathbb{R}^n$  and  $(f, h) \in \mathcal{B}_L(C)$  a relatively-smooth problem instance. Then the sequence  $\{x_k\}_{k \geq 0}$  generated by Algorithm 1 with constant step size  $\lambda \in (0, 1/L]$  satisfies for  $k \geq 2$*

$$\min_{1 \leq i \leq k} D_h(x_{i-1}, x_i) \leq \frac{2D_h(x_*, x_0)}{k(k-1)}$$

for every  $x_* \in \operatorname{argmin}_C f \cap \operatorname{dom} h$ .

*Proof* In the same way as before, the formal guarantee has been obtained by examining the dual of the corresponding PEP. The proof relies on the fact that  $\frac{1}{\lambda}h - f$  is convex for any  $\lambda \in (0, \frac{1}{L}]$ , and only consists in performing the following weighted sum of inequalities:

- convexity of  $f$ , between  $x_*$  and  $x_i$  ( $i = 0, \dots, k$ ) with weights  $\gamma_{*,i} = \frac{2\lambda}{k(k-1)}$ :

$$f(x_*) \geq f(x_i) + \langle \nabla f(x_i), x_* - x_i \rangle,$$

- optimality of  $x_*$  for each  $x_k$  with weight  $\gamma_{k,*} = \frac{2\lambda}{k-1}$ :

$$f(x_k) \geq f(x_*),$$

- convexity of  $\frac{1}{\lambda}h - f$ , between  $x_*$  and  $x_k$  with weight  $\mu_{*,k} = \frac{2\lambda}{k(k-1)}$ :

$$\frac{1}{\lambda}h(x_*) - f(x_*) \geq \frac{1}{\lambda}h(x_k) - f(x_k) + \langle \frac{1}{\lambda}\nabla h(x_k) - \nabla f(x_k), x_* - x_k \rangle,$$

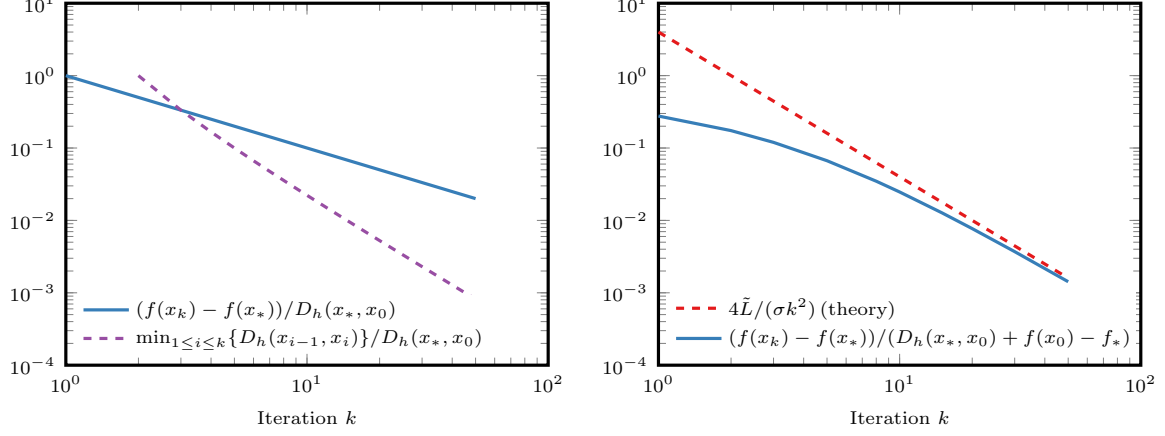
- convexity of  $\frac{1}{\lambda}h - f$ , between  $x_{i+1}$  and  $x_i$  ( $i = 0, \dots, k-1$ ) with weight  $\mu_{i+1,i} = \frac{2\lambda(i+1)}{k(k-1)}$

$$\frac{1}{\lambda}h(x_{i+1}) - f(x_{i+1}) \geq \frac{1}{\lambda}h(x_i) - f(x_i) + \langle \frac{1}{\lambda}\nabla h(x_i) - \nabla f(x_i), x_{i+1} - x_i \rangle,$$

- definition of smallest residual among the iterates ( $i = 1, \dots, k$ ) with weights  $\tau_i = \frac{2(i-1)}{k(k-1)}$ :

$$h(x_{i-1}) - h(x_i) - \langle \nabla h(x_i), x_{i-1} - x_i \rangle \geq \min_{1 \leq j \leq k} \{D_h(x_{j-1}, x_j)\}.$$





**Fig. 3** Numerical worst-case guarantees obtained from PEPs as functions of the iteration counter  $k$  (shown in log scale as rates are sublinear). **Left:** guarantees for NoLips (Algorithm 1) for two different convergence measures. Numerical values confirm exactly the theoretical rates of Theorem 1 and Proposition 4. **Right:** guarantees for IGA with no affine constraints (Algorithm 2) under the assumption that  $h$  is 1-strongly convex and  $f$  is 1-smooth, compared to the theoretical bound from [1]. Notice that the theoretical bound is not tight in this case, as it is obtained by making some approximations in the proof.

The weighted sum is written as

$$\begin{aligned}
0 &\geq \sum_{i=0}^k \gamma_{*,i} [f(x_i) - f(x_*) + \langle \nabla f(x_i), x_* - x_i \rangle] \\
&\quad + \gamma_{k,*} [f(x_*) - f(x_k)] \\
&\quad + \mu_{*,k} [\frac{1}{\lambda} h(x_k) - f(x_k) - (\frac{1}{\lambda} h(x_*) - f(x_*)) + \langle \frac{1}{\lambda} \nabla h(x_k) - \nabla f(x_k), x_* - x_k \rangle] \\
&\quad + \sum_{i=0}^{k-1} \mu_{i+1,i} [\frac{1}{\lambda} h(x_i) - f(x_i) - (\frac{1}{\lambda} h(x_{i+1}) - f(x_{i+1})) + \langle \frac{1}{\lambda} \nabla h(x_i) - \nabla f(x_i), x_{i+1} - x_i \rangle] \\
&\quad + \sum_{i=1}^k \tau_i [\min_{1 \leq j \leq k} \{D_h(x_{j-1}, x_j)\} - (h(x_{i-1}) - h(x_i) - \langle \nabla h(x_i), x_{i-1} - x_i \rangle)].
\end{aligned}$$

By substitution of  $\nabla h(x_{i+1}) = \nabla h(x_i) - \lambda \nabla f(x_i)$  ( $i = 0, \dots, k-1$ ), one can reformulate the weighted sum exactly as (i.e., there is no residual):

$$0 \geq \min_{1 \leq j \leq k} \{D_h(x_{j-1}, x_j)\} - 2 \frac{h(x_*) - h(x_0) - \langle \nabla h(x_0), x_* - x_0 \rangle}{k(k-1)},$$

yielding the desired result. ■

#### 4.5.3 Beyond NoLips: inertial Bregman algorithms

Our approach is not limited to the NoLips algorithm. For instance, we can also solve the performance estimation problem for the inertial Bregman algorithm proposed by Auslender and Teboulle [1], a.k.a. the Improved Interior Gradient Algorithm (IGA). We recall its simplified formulation in Algorithm 2, in the case where there are no affine constraints.

---

**Algorithm 2** Improved Interior Gradient Algorithm (IGA) [1]

---

**Input:** Functions  $f, h$ , initial point  $x_0 \in \text{int dom } h$ , step size  $\lambda$ .

Set  $z_0 = x_0$  and  $t_0 = 1$ .

**for**  $k = 0, 1, \dots$  **do**

$$y_k = (1 - \frac{1}{t_k})x_k + \frac{1}{t_k}z_k$$

$$z_{k+1} = \text{argmin} \{ \langle \nabla f(y_k), u - y_k \rangle + \frac{1}{t_k \lambda} D_h(u, z_k) \mid u \in \mathbb{R}^n \}$$

$$x_{k+1} = (1 - \frac{1}{t_k})x_k + \frac{1}{t_k}z_{k+1}$$

$$t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2.$$

**end for**

---

In the setting where  $f$  has  $\tilde{L}$ -Lipschitz continuous gradients and  $h$  is a  $\sigma$ -strongly convex kernel function, IGA with step size  $\lambda = \sigma/\tilde{L}$  enjoys the following convergence rate [1, Thm. 5.2]:

$$f(x_N) - f_* \leq \frac{4\tilde{L}}{\sigma N^2} (D_h(x_*, x_0) + f(x_0) - f_*). \quad (23)$$

Our PEP framework can also be applied to this algorithm, in order to find the smallest value of  $\theta(N, \tilde{L}, \sigma, \lambda)$  which satisfies

$$f(x_N) - f_* \leq \theta(N, \tilde{L}, \sigma, \lambda) (D_h(x_*, x_0) + f(x_0) - f_*)$$

for every instance of IGA with the supplementary assumptions made above. In this case, we use the standard interpolation conditions of Theorem 3 for  $L$ -smooth and strongly convex functions. Results are shown in Figure 3. The exact numerical worst-case performance of IGA is slightly below the theoretical bound above, since the proof in [1] makes some approximations.

*IGA in the general relatively-smooth case: failure of inertia.* We pointed out in Section 2 that the setting in which  $f$  is  $\tilde{L}$ -smooth and  $h$  is  $\sigma$ -strongly convex is a particular case of relative smoothness with constant  $L = \tilde{L}/\sigma$ . The natural question that was also raised in [39, Section 6] is therefore: does IGA converge for the general class  $\mathcal{B}_L(C)$ ? Solving the corresponding PEP yields the following results. For Algorithm 2 with the setting that  $(f, h) \in \mathcal{B}_L(C)$  and several choices of step size in  $(0, 1/L]$ , the solver states that the value of the corresponding performance estimation problem **is unbounded**, i.e., there does not exist any  $\theta$  such that the bound (23) holds for every instance  $(f, h) \in \mathcal{B}_L$ .

One could legitimately wonder whether there exist other sequences  $\{t_k\}_{k \geq 0}$  with  $t_k > 1$ , perhaps *less aggressive* than the one in Algorithm 2, such that the method converges (note that choosing  $t_k = 1 \ \forall k \geq 0$  would yield the standard NoLips scheme). After solving the PEP with several choices of such sequences and observing that it is unbounded, we formulate the following conjecture: for *any* sequence  $\{t_k\}_{k \geq 0}$ , in IGA, such that  $t_{k_0} > 1$  for some  $k_0$ , it is not possible to bound  $f(x_N) - f_*$  in general. Of course, this constitutes numerical evidence and not a formal proof. The conjecture could be proved by constructing worst-case functions in the same spirit as in Section 3, with some pathological lack of smoothness that would cause the iterates to diverge when taking a step size larger than  $1/L$ .

These experiments lead us to believe that inertial methods with non-adaptive coefficients fail to converge in the general relatively-smooth setting.

#### 4.5.4 From worst-case functions for NoLips to a lower bound for general Bregman methods

We briefly explain how, with the PEP methodology, the worst case functions from Section 3.2 were discovered.

We described in Section 4.5.1 how a one-dimensional worst-case instance  $(\bar{f}, \bar{h})$  for NoLips was discovered from low-rank solutions of ( $\text{sdp-PEP}$ ). However, this instance may not be difficult enough for a more generic Bregman algorithm that can use arbitrary linear combinations of gradients (as in Definition 4, our definition of the *Bregman gradient algorithm*), and thus cannot be used to prove a general lower bound.

Our objective now is to find worst-case instances that are difficult for **any** Bregman gradient algorithm. A desirable property would be that these instances allow to explore only *one dimension* per oracle call, so that the function *hides information* in the unexplored dimensions. This is similar in spirit to the so-called “worst function in the world” of Nesterov [32]. In order to achieve this goal, we propose to search for functions  $f$  for which all gradients  $\nabla f(x_i)$  are orthogonal, guaranteeing that one new dimension is explored at each step. Note that a similar approach has been used in some previous work on PEPs to find lower bounds or optimal methods e.g., in [15, 17]. This amounts to adding some orthogonality constraints to (PEP) and solving

$$\begin{aligned} & \text{maximize} && (f(x_N) - f(x_*)) / D_h(x_*, x_0) \\ & \text{subject to} && (f, h) \in \mathcal{B}_L(\mathbb{R}^n), \\ & && x_* \text{ is a minimizer of } f, \\ & && x_1, \dots, x_N \text{ are generated from } x_0 \text{ by Algorithm 1 with step size } \lambda, \\ & && \langle \nabla f(x_i), \nabla f(x_j) \rangle = 0 \text{ for } i \neq j \in I, \end{aligned} \tag{PEP-orth}$$

in the variables  $f, h, x_0, \dots, x_N, x_*, n$ .

In the same spirit as before, we were able to find a dimension- $N$  solution of (PEP-orth). This allows us to interpolate the following worst-case pathological instance:

$$\begin{aligned} \bar{f}(x) &= \|x - (1, \dots, 1)\|_\infty, \\ \bar{h}(x) &= \bar{f}(x) + \sum_{i=2}^N \max(-x^{(i)}, 0). \end{aligned}$$

Again, these are nonsmooth functions and, as such, they do not form valid instances for NoLips. However, they can be approached by a sequence of such functions, for instance by applying smoothing with the Moreau envelope, and adding a small quadratic term to make  $h$  strictly convex. Along with a few tweaks, this is how we found the example that was used to prove the general lower bound for  $\mathcal{B}_L$  in Section 3.2.

## 5 Conclusion

Our paper has two main contributions: proving optimality of NoLips for the general relatively-smooth setting, and developing numerical performance estimation techniques for Bregman gradient algorithms. We presented the performance estimation problem on the basic NoLips algorithm for simplicity, but our approach can be applied to different settings and various algorithms involving Bregman distances. We provided several applications illustrating how the PEP methodology is an efficient tool for conjecturing and analyzing the worst-case behavior of Bregman algorithms.

There is a fundamental concept linking the two parts of the paper, which is that of *limiting nonsmooth pathological behavior*. When looking for worst-case guarantees over a class of functions that is open such as the class of differentiable convex functions, the performance estimation problem is a *supremum* and the worst-case maximizing *sequence* might approach some function that is not in this class, e.g., one that is nonsmooth in our case. This idea, observed by analyzing the equivalence between (PEP) and the nonsmooth relaxation ( $\overline{\text{PEP}}$ ), was used in the proof of the lower bound in Section 3.2. Moreover, the worst-case sequence of functions was directly inspired by examining particular solutions of ( $\overline{\text{PEP}}$ ).

Our result also shows that additional assumptions on functions  $f$  and  $h$  are needed in order to prove better bounds or devise faster algorithms than NoLips. If the usual properties of  $L$ -smoothness and strong convexity are too restrictive and do not hold in many applications, the future challenge is to find weaker assumptions, that define a larger class of functions where improved rates can be obtained. One other possible approach would be to find algorithms that do not fit in Definition 4, for instance by including second-order oracles of  $h$ , in the case when  $h$  is simple enough.

**Code.** Experiments have been run in MATLAB, using the semidefinite solver MOSEK [28] as well as the modeling toolbox YALMIP [25]. The support for Bregman methods has been added to the Performance Estimation Toolbox (PESTO, [36]) for which we provide some examples. The code can be downloaded from <https://github.com/RaduAlexandruDragomir/BregmanPerformanceEstimation>.

**Acknowledgements.** The authors would like to thank the anonymous reviewers for constructive suggestions as well as Dmitrii Ostrovskii and Edouard Pauwels for useful comments. RD acknowledges support from an AMX fellowship. AT acknowledges support from the European Research Council (grant SEQUOIA 724063). AA is at CNRS, and CS Department, Ecole Normale Supérieure, PSL Research University, 45 rue d’Ulm, 75005, Paris. AA would like to acknowledge support from the *ML and Optimisation* joint research initiative with the *fonds AXA pour la recherche* and Kamet Ventures, a Google focused award, as well as funding 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). JB acknowledges the support of ANR-3IA ANITI, ANR Chess, Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant numbers FA9550-19-1-7026, FA9550-18-1-0226. JB acknowledges financial support of the research foundation TSE-Partnership.

## A Extension of performance analysis to the case when $C$ is a general closed convex subset of $\mathbb{R}^n$

For simplicity of the presentation, we left out in Section 4 the case when the domain  $C$  is a proper subset of  $\mathbb{R}^n$ . We show in this section that it actually corresponds to the same minimization problem ( $\overline{\text{sdp-PEP}}$ ).

Let us formulate the performance estimation problem for Algorithm 1 in the general case. Recall that we denote  $\mathcal{B}_L$  the union of  $\mathcal{B}_L(C)$  for all closed convex subsets of  $\mathbb{R}^n$  and for every  $n \geq 1$ . The performance estimation problem writes

$$\begin{aligned} & \text{maximize} && (f(x_N) - f(x_*))/D_h(x_*, x_0) \\ & \text{subject to} && (f, h) \in \mathcal{B}_L, \\ & && x_* \text{ is a minimizer of } f \text{ on } \overline{\text{dom } h} \text{ such that } x \in \text{dom } h, \\ & && x_1, \dots, x_N \text{ are generated from } x_0 \text{ by Algorithm 1 with step size } \lambda, \end{aligned} \tag{PEP-C}$$

in the variables  $f, h, x_0, \dots, x_N, x_*, n$ . Now, as (PEP-C) is a problem that includes (PEP) in the special case where  $C = \mathbb{R}^n$ , its value is larger:

$$\text{val}(\text{PEP}) \leq \text{val}(\text{PEP-C})$$

Let us show that  $\text{val}(\text{PEP-C})$  is upper bounded by the same relaxation  $\text{val}(\overline{\text{PEP}})$ , which allows to conclude that the values are equal. We recall that the problem ( $\overline{\text{PEP}}$ ) can be written, using interpolation conditions of Corollary 1, as

$$\begin{aligned} & \text{maximize} && f_N - f_* \\ & \text{subject to} && f_i - f_j - \langle g_j, x_i - x_j \rangle \geq 0, \\ & && (Lh_i - f_i) - (Lh_j - f_j) - \langle Ls_j - g_j, x_i - x_j \rangle \geq 0 \quad \text{for } i, j \in I, \\ & && g_* = 0, \\ & && s_{i+1} = s_i - \lambda g_i \quad \text{for } i \in \{1, \dots, N-1\}, \\ & && h_* - h_0 - \langle s_0, x_* - x_0 \rangle = 1, \end{aligned} \tag{\overline{PEP}}$$

in the variables  $n, \{(x_i, f_i, g_i, h_i, s_i)\}_{i \in I}$ . We show that every admissible point of (PEP-C) can be cast into an admissible point of ( $\overline{\text{sdp-PEP}}$ ). This actually amounts to show that, from the point of view of performance estimation, an instance  $(f, h) \in \mathcal{B}_L(C)$  is actually equivalent to some instance in  $\mathcal{B}_L(\mathbb{R}^n)$ .

Let  $f, h, x_0, \dots, x_N, x_*$  be a feasible point of (PEP-C). We distinguish two cases.

*Case 1:*  $x_* \in \text{int dom } h$ . This is the simplest case, as the necessary conditions are the same as in the situation where  $C = \mathbb{R}^n$ . Indeed, then we have  $x_0, \dots, x_N, x_* \in \text{int dom } h$ , since  $x_0$  is constrained to be in the interior and the next iterates are in  $\text{int dom } h$  by Assumption 1. Since  $f$  and  $h$  are differentiable on  $\text{int dom } h$ , convexity of  $f$  and  $Lh - f$  imply that the first two constraints of ( $\overline{\text{PEP}}$ ) hold for all  $i, j \in I$ . Finally,  $g_* = 0$  follows from the fact that  $x_*$  minimizes  $f$  and that it lies on the interior of the domain. Hence the discrete representation satisfies the constraints of ( $\overline{\text{sdp-PEP}}$ ).

*Case 2:*  $x_* \in \partial \text{dom } h$ . In this case,  $f$  and  $h$  are not necessarily differentiable at  $x_*$ , but are still differentiable still at  $x_0, \dots, x_N$  for the same reasons. But we can still, with a small modification at  $x_*$ , derive a discrete representation that fits the constraints of ( $\overline{\text{PEP}}$ ) and whose objective is the same. Indeed, define

$$\begin{aligned} (g_i, f_i, s_i, h_i) &= (\nabla f(x_i), f(x_i), \nabla h(x_i), h(x_i)) \quad \text{for } i = 0, \dots, N, \\ (g_*, f_*, s_*, h_*) &= (0, f(x_*), v, h(x_*)), \end{aligned}$$

where  $v \in \mathbb{R}^n$  is a vector that are specified later. Then, for  $i \in I$  and  $j \in \{0 \dots N\}$ , convexity of  $f$  and  $Lh - f$  imply that the constraints

$$\begin{aligned} f_i - f_j - \langle g_j, x_i - x_j \rangle &\geq 0 \\ (Lh_i - f_i) - (Lh_j - f_j) - \langle Ls_j - g_j, x_i - x_j \rangle &\geq 0 \end{aligned}$$

hold. It remains to verify them for  $i \in \{0 \dots N\}$  and  $j = *$ . The first one holds because  $x_*$  minimizes  $f$  on  $\text{dom } h$ , so with  $g_* = 0$  we have  $f_i - f_* \geq 0$ . We now show that the second one is satisfied, i.e., that we can choose  $v \in \mathbb{R}^n$  so that

$$(Lh_i - f_i) - (Lh_* - f_*) - \langle Lv, x_i - x_* \rangle \geq 0 \quad \forall i \in \{0 \dots N\}.$$

To this extent, we use the fact that  $x_* \in \partial \text{dom } h$  and that  $x_i \in \text{int dom } h$  for  $i = 0 \dots N$ . This means that  $\{x_*\} \cap \text{int dom } h = \emptyset$ , and therefore by the hyperplane separation theorem [34, Thm 11.3], there exists a hyperplane that separates the convex sets  $\{x_*\}$  and  $\text{int dom } h$  *properly*, meaning that there exists a vector  $u \in \mathbb{R}^n$  such that

$$\langle x_i - x_*, u \rangle < 0 \quad \forall i \in \{0, \dots, N\}.$$

Set

$$\begin{aligned} \alpha &= \min_{i=0 \dots N} (Lh_i - f_i) - (Lh_* - f_*), \\ \beta &= \min_{i=0, \dots, N} -\langle x_i - x_*, u \rangle > 0, \end{aligned}$$

where  $\beta > 0$  because of the separation result. Choose  $s_* = v$  as  $v = \frac{|\alpha|}{L\beta}u$ . Then we have

$$\begin{aligned} (Lh_i - f_i) - (Lh_* - f_*) - \langle Ls_*, x_i - x_* \rangle &\geq \alpha + L \frac{|\alpha|}{L\beta} \beta \\ &\geq \alpha + |\alpha| \\ &\geq 0. \end{aligned}$$

This eventually provides an instance  $\{(x_i, g_i, f_i, h_i, s_i)\}_{i \in I}$  that is admissible for  $(\overline{\text{PEP}})$ .

To conclude, we proved that in both cases, an admissible point of  $(\text{PEP-C})$  can be turned into an admissible point of  $(\text{sdp-}\overline{\text{PEP}})$  with the same objective value. Hence we have

$$\text{val}(\text{PEP-C}) \leq \text{val}(\text{sdp-}\overline{\text{PEP}}).$$

Recalling that  $\text{val}(\text{PEP}) \leq \text{val}(\text{PEP-C})$  and that  $\text{val}(\text{sdp-}\overline{\text{PEP}}) = \text{val}(\text{PEP})$  by Theorem 4, we get

$$\text{val}(\text{PEP-C}) = \text{val}(\text{PEP}).$$

In other words, solving the performance estimation problem  $(\text{PEP-C})$  for functions with any closed convex domain is equivalent to solving the performance estimation problem  $(\text{PEP})$  restricted to functions that have full domain.

## References

1. Auslender, A., Teboulle, M.: Interior Gradient and Proximal Methods for Convex and Conic Optimization. *SIAM Journal on Optimization* **16**(3), 697–725 (2006) [2](#), [3](#), [5](#), [7](#), [8](#), [14](#), [25](#), [26](#)
2. Bach, F.: Duality Between Subgradient and Conditional Gradient Methods. *SIAM Journal on Imaging Sciences* **25**(1), 115–129 (2015) [14](#)
3. Bauschke, H.H., Bolte, J., Chen, J., Teboulle, M., Wang, X.: On Linear Convergence of Non-Euclidean Gradient Methods without Strong Convexity and Lipschitz Gradient Continuity. *Journal of Optimization Theory and Applications* **182**(3), 1068–1087 (2019) [3](#)
4. Bauschke, H.H., Bolte, J., Teboulle, M.: A Descent Lemma Beyond Lipschitz Gradient Continuity: First-Order Methods Revisited and Applications. *Mathematics of Operations Research* **42**(2), 330–348 (2017) [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
5. Bauschke, H.H., Combettes, P.L.: *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer Publishing Company, Inc. (2011) [4](#)
6. Beck, A., Teboulle, M.: Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization. *Operations Research Letters* **31**(3), 167–175 (2003) [3](#)
7. Beck, A., Teboulle, M.: A Fast Iterative Shrinkage-Thresholding Algorithm. *SIAM Journal on Imaging Sciences* **2**(1), 183–202 (2009) [2](#)
8. Ben-tal, A., Margalit, T., Nemirovski, A.: The Ordered Subsets Mirror Descent Optimization Method with Applications to Tomography. *SIAM Journal on Optimization* **12**(1), 79–108 (2001) [3](#)
9. Bertero, M., Boccaci, P., Desidera, G., Vicidomini, G.: Image Deblurring with Poisson Data: From Cells to Galaxies. *Inverse Problems* (2009) [6](#)
10. Bolte, J., Sabach, S., Teboulle, M., Vaisbourd, Y.: First Order Methods Beyond Convexity and Lipschitz Gradient Continuity with Applications to Quadratic Inverse Problems. *SIAM Journal on Optimization* **28**(3), 2131–2151 (2018) [2](#), [3](#), [4](#), [6](#)
11. Bubeck, S.: *Introduction to online optimization*. Lecture Notes (2011) [3](#)

12. Bùì, M.N., Combettes, P.L.: Bregman Forward-Backward Operator Splitting. arXiv preprint arXiv:1908.03878 (2019) [3](#)
13. Censor, Y., Zenios, S.A.: Proximal Minimization Algorithm with D-functions. *Journal of Optimization Theory and Applications* **73**(3), 451–464 (1992) [3](#)
14. Dragomir, R.A., D’Aspremont, A., Bolte, J.: Quartic First-Order Methods for Low Rank Minimization. arXiv preprint arXiv:1901.10791v2. To appear in *Journal of Optimization Theory and Applications* (2020) [2](#), [6](#)
15. Drori, Y.: The Exact Information-Based Complexity of Smooth Convex Minimization. *Journal of Complexity* **39**, 1–16 (2017) [3](#), [7](#), [27](#)
16. Drori, Y., Shamir, O.: The complexity of finding stationary points with stochastic gradient descent. In: *Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research*, vol. 119, pp. 2658–2667. PMLR (2020) [3](#)
17. Drori, Y., Taylor, A.B.: Efficient First-Order Methods for Convex Minimization: a Constructive Approach. *Mathematical Programming* **184**, 183–220 (2020) [3](#), [27](#)
18. Drori, Y., Teboulle, M.: Performance of First-Order Methods for Smooth Convex Minimization: A Novel Approach. *Mathematical Programming* **145**(1-2), 451–482 (2014) [2](#), [3](#), [14](#), [15](#), [19](#), [22](#)
19. Drori, Y., Teboulle, M.: An Optimal Variant of Kelley’s Cutting-Plane Method. *Mathematical Programming* **160**(1-2), 321–351 (2016) [3](#)
20. Eckstein, J.: Nonlinear Proximal Point Algorithms Using Bregman Functions, with Applications to Convex Programming. *Mathematics of Operations Research* **18**(1), 202–226 (1993) [3](#), [14](#)
21. Guzmán, C., Nemirovski, A.: On Lower Complexity Bounds for Large-Scale Smooth Convex Optimization. *Journal of Complexity* **31**(1), 1–14 (2015) [3](#), [7](#), [14](#)
22. Hanzely, F., Richtarik, P., Xiao, L.: Accelerated Bregman Proximal Gradient Methods for Relatively Smooth Convex Optimization. ArXiv preprint arXiv:1808.03045v1 (2018) [2](#), [3](#), [7](#), [8](#)
23. Juditsky, A., Nemirovski, A.: First Order Methods for Nonsmooth Convex Large-Scale Optimization , I : General Purpose Methods. In: S.S. Wright, S. Nowozin, S. J. (eds.) *Optimization for Machine Learning*, pp. 121–147. MIT Press (2010) [3](#)
24. Kim, D., Fessler, J.A.: Optimized First-Order Methods for Smooth Convex Minimization. *Mathematical Programming* **159**(1-2), 81–107 (2016) [3](#)
25. Lofberg, J.: YALMIP : A Toolbox for Modeling and Optimization in MATLAB. In: *In Proceedings of the CACSD Conference* (2004) [17](#), [27](#)
26. Lu, H., Freund, R.M., Nesterov, Y.: Relatively-Smooth Convex Optimization by First-Order Methods, and Applications. *SIAM Journal on Optimization* **28**(1), 333–354 (2018) [2](#), [3](#), [4](#), [5](#), [8](#)
27. Moreau, J.J.: Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. Fr.* **93**(2), 273–299 (1965) [9](#)
28. Mosek, A.: The MOSEK optimization toolbox for MATLAB manual. Version 9.0. (2019). URL <http://docs.mosek.com/9.0/toolbox/index.html> [17](#), [27](#)
29. Muckkamala, M.C., Ochs, P., Pock, T., Sabach, S.: Convex-Concave Backtracking for Inertial Bregman Proximal Gradient Algorithms in Non-Convex Optimization. arXiv preprint arXiv:1904.03537 (2019) [3](#)
30. Nemirovski, A., Yudin, D.B.: *Problem Complexity and Method Efficiency in Optimization* (1983) [2](#), [3](#), [7](#)
31. Nesterov, Y.: A Method for Solving a Convex Programming Problem with Convergence Rate  $O(1/k^2)$  **27**(2), 367–372 (1983) [2](#), [8](#)
32. Nesterov, Y.: *Introductory Lectures on Convex Optimization: A Basic Course*, 1 edn. Springer Publishing Company, Inc (2003) [3](#), [7](#), [8](#), [9](#), [27](#)
33. Nesterov, Y.: Implementable Tensor Methods in Unconstrained Convex Optimization. CORE Discussion Paper (2018) [2](#)
34. Rockafellar, R.T.: *Convex Analysis* (1970) [4](#), [17](#), [18](#), [29](#)
35. Taylor, A., Hendrickx, J., Glineur, F.: Exact Worst-Case Performance of First-Order Methods for Composite Convex Optimization. *SIAM Journal on Optimization* **27**, 1283–1313 (2017) [14](#), [22](#)
36. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Performance Estimation Toolbox (PESTO): Automated Worst-Case Analysis of First-Order Optimization Methods. In: *IEEE 56th Annual Conference on Decision and Control (CDC 2017)*, pp. 1278–1283 (2017) [27](#)
37. Taylor, A.B., Hendrickx, J.M., Glineur, F.: Smooth Strongly Convex Interpolation and Exact Worst-Case Performance of First-Order Methods. *Mathematical Programming* **161**(1-2), 307–345 (2017)

- [3](#), [11](#), [15](#), [17](#), [19](#), [23](#)
38. Teboulle, M.: Entropic Proximal Mappings with Applications to Nonlinear Programming. *Mathematics of Operations Research* **17**(3), 670–690 (1992) [3](#)
  39. Teboulle, M.: A simplified view of first order methods for optimization. *Mathematical Programming* **170**(1), 67–96 (2018) [2](#), [26](#)
  40. Vandenberghe, L., Boyd, S.: Semidefinite Programming. *SIAM Review* **38**(1), 49–45 (1996) [17](#)
  41. Walid, K., Bayen, A., Bartlett, P.L.: Accelerated Mirror Descent In Continuous and Discrete Time. In: *Advances in Neural Information Processing Systems* 28, pp. 2845—2853 (2015) [2](#), [5](#)
  42. Woodworth, B., Srebro, N.: Lower Bound for Randomized First Order Convex Optimization. arXiv preprint arXiv:1709.03594 (2017) [3](#), [7](#)