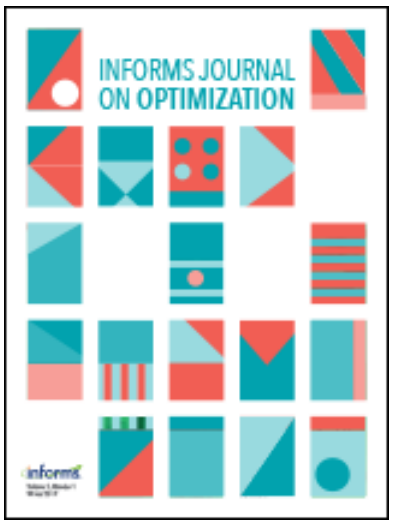

INFORMS Journal on Optimization

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

"Relative Continuity" for Non-Lipschitz Nonsmooth Convex Optimization Using Stochastic (or Deterministic) Mirror Descent

Haihao Lu

To cite this article:
Haihao Lu (2019) "Relative Continuity" for Non-Lipschitz Nonsmooth Convex Optimization Using Stochastic (or Deterministic) Mirror Descent. INFORMS Journal on Optimization 1(4):288-303. https://doi.org/10.1287/ijoo.2018.0008

http://pubsonline.informs.org/journal/ijoo

INFORMS JOURNAL ON OPTIMIZATION
Vol. 1, No. 4, Fall 2019, pp. 288–303
ISSN 2575-1484 (print), ISSN 2575-1492 (online)

"Relative Continuity" for Non-Lipschitz Nonsmooth Convex Optimization Using Stochastic (or Deterministic) Mirror Descent

Haihao Lu[a]

[a] Department of Mathematics and Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139
Contact: haihao@mit.edu, http://orcid.org/0000-0002-5217-1894 (HL)

Received: May 1, 2018
Revised: August 13, 2018
Accepted: September 14, 2018
Published Online in Articles in Advance: June 5, 2019

https://doi.org/10.1287/ijoo.2018.0008

Abstract. The usual approach to developing and analyzing first-order methods for nonsmooth (stochastic or deterministic) convex optimization assumes that the objective function is uniformly Lipschitz continuous with parameter M_f. However, in many settings, the nondifferentiable convex function f is not uniformly Lipschitz continuous—for example, (i) the classical support vector machine problem, (ii) the problem of minimizing the maximum of convex quadratic functions, and even (iii) the univariate setting with $f(x) := \max\{0, x\} + x^2$. Herein, we develop a notion of "relative continuity" that is determined relative to a user-specified "reference function" h (that should be computationally tractable for algorithms), and we show that many nondifferentiable convex functions are relatively continuous with respect to a correspondingly fairly simple reference function h. We also similarly develop a notion of "relative stochastic continuity" for the stochastic setting. We analyze two standard algorithms—the (deterministic) mirror descent algorithm and the stochastic mirror descent algorithm—for solving optimization problems in these new settings, providing the first computational guarantees for instances where the objective function is not uniformly Lipschitz continuous. This paper is a companion paper for nondifferentiable convex optimization to the recent paper by Lu et al. [Lu H, Freund RM, Nesterov Y (2018) Relatively smooth convex optimization by first-order methods, and applications. *SIAM J. Optim.* 28(1): 333–354.], which developed analogous results for differentiable convex optimization.

Funding: The research of H. Lu is supported by the Air Force Office of Scientific Research [Grant FA9550-15-1-0276] and the Massachusetts Institute of Technology–Belgium Université Catholique de Louvain Fund.

Keywords: non-Lipschitz optimization • mirror descent • stochastic optimization

1. Introduction

The usual approach to developing and analyzing first-order methods for nondifferentiable convex optimization (which we review shortly) assumes that the objective function is uniformly Lipschitz continuous in both deterministic and stochastic settings. However, in many settings, the objective function f is not uniformly Lipschitz continuous. For example, consider the support vector machine (SVM) problem for binary classification in machine learning with optimization formulation; that is,

$$\text{SVM:} \quad \min_x f(x) := \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i x^T w_i\} + \frac{\lambda}{2}\|x\|_2^2,$$

where w_i is the input feature vector of sample i and $y_i \in \{-1, 1\}$ is the label of sample i. Notice that f is not differentiable due to the presence of hinge loss terms in the summation, and f is also not Lipschitz continuous due to the presence of the ℓ_2-norm regularization term; thus, we cannot directly utilize typical subgradient or gradient schemes and their associated computational guarantees for SVM.

Another example is the problem of computing a point $x \in \mathbb{R}^m$ in the intersection of n ellipsoids, which can be tackled via the optimization problem (intersection of ellipsoids problem [IEP])

$$\text{IEP:} \quad f^* = \min_x f(x) := \max_{0 \le i \le n} \left\{\frac{1}{2} x^T A_i x + b_i^T x + c_i\right\},$$

where the ith ellipsoid is $\mathcal{Q}_i = \{x \in \mathbb{R}^m : 12 x^T A_i x + b_i^T x + c_i \le 0\}$ and $A_i \in \mathbb{R}^{m \times m}$ is a symmetric positive semidefinite matrix $i = 1, \ldots, n$. Observe that the objective function f is both nondifferentiable and non-Lipschitz, and therefore, it falls outside of the scope of standard classes of optimization problems for which first-order

methods are guaranteed to work. Nevertheless, using the machinery developed in this paper, we will show in Section 5.3 how to solve both of these problems using deterministic or stochastic mirror descent.

In this paper, we develop a general theory and algorithmic constructs that overcome the drawbacks in the usual analyses of first-order methods that are grounded on restricted notions of uniform Lipschitz continuity. Here, we develop a notion of "relative continuity" with respect to a given convex "reference function" h, a notion that does not require the specification of any particular norm—indeed, h does not need to be strongly (or even strictly) convex. Armed with relative continuity, we show the capability to solve a more general class of nondifferentiable convex optimization problems (without uniform Lipschitz continuity) in both deterministic and stochastic settings.

This paper is a companion for nondifferentiable convex optimization to our predecessor paper (Lu et al. 2018) for differentiable convex optimization. In Lu et al. (2018), with a very similar philosophy, we developed the notion of relative smoothness and relative strong convexity with respect to a given convex reference function. In that paper, we showed the capability to solve a more general class of differentiable convex optimization problems (without uniform Lipschitz continuous gradients), and we also showed linear convergence results for a primal gradient scheme when the objective function f is both smooth and strongly convex.

There are some concurrent works on smooth optimization sharing a similar spirit to Lu et al. (2018). Bauschke et al. (2016) presents a similar definition of relative smoothness as in Lu et al. (2018) and analyzes the convergence of mirror descent algorithm, although their algorithm and convergence complexity depend on a symmetry measure of the Bregman distance. Zhou et al. (2016) discuss a unified proof of mirror descent and the proximal point algorithm under a similar assumption of relative smoothness. Van Nguyen (2017) develops similar ideas on analyzing mirror descent in a Banach space. A more detailed discussion comparing these related works is also presented in Lu et al. (2018). More recently, Hanzely and Richtarik (2018) develop stochastic algorithms for the relatively smooth optimization setting.

In Section 2, we review the traditional setup for mirror descent in both the deterministic and stochastic settings. In Section 3, we introduce our notion of relative continuity in both the deterministic and stochastic settings together with some relevant properties. In Section 4, we prove computational guarantees associated with the mirror descent and stochastic mirror descent algorithms under relative continuity. In Section 5, we show constructively how our ideas apply to a large class of nondifferentiable and non-Lipschitz convex optimization problems that are not otherwise solvable by traditional first-order methods. Also, in Section 5, we analyze computational guarantees associated with mirror descent and stochastic mirror descent for the IEP and the SVM problem.

1.1. Notation

$\|\cdot\|$ denotes a given norm in $\mathbb{R}^n$ and $\|\cdot\|_*$ denotes the usual dual norm on the dual space. $\|x\|_2 := \sqrt{x^T x}$ denotes the Euclidean (inner product) norm, where x^T means the transpose of the vector x and $B_2(c,r) := \{x \in \mathbb{R}^n : \|x-c\|_2 \le r\}$. $\|A\|_2$ denotes the ℓ_2 (spectral) norm of a matrix A. The inner product $\langle\cdot,\cdot\rangle$ specifically denotes the dot inner product in the underlying vector space. For a conditional random variable $s(x)$ given x, $\mathbb{E}[s(x)|x]$ denotes the conditional expectation of $s(x)$ given x.

2. Traditional Mirror Descent

The optimization problem of interest is

$$P: \quad \min_x \quad f(x) \\ \text{s.t.} \quad x \in Q, \tag{1}$$

where $Q \subseteq \mathbb{R}^n$ is a closed convex set and $f : Q \to \mathbb{R}$ is a convex function that is not necessarily differentiable (and s.t. refers to such that). There are many deterministic and stochastic first-order methods for tackling (1) (see, for example, Nesterov 2003, Nedić and Lee 2014, Bubeck 2015, and the references therein). Virtually all such methods are designed to solve (1) when the objective function f satisfies a uniform Lipschitz continuity condition on Q, which in the deterministic setting, is (essentially) equivalent to the condition that there exists a constant $M_f < \infty$ for which

$$\|g(x)\|_* \le M_f \quad \text{for all } x \in Q \text{ and } g(x) \in \partial f(x), \tag{2}$$

where $\partial f(x)$ is the subdifferential of f at x (i.e., the collection of subgradients of f at x), $\|\cdot\|$ is a given norm on $\mathbb{R}^n$, and $\|\cdot\|_*$ denotes the usual dual norm.

Here, we use "$g(x)$" to denote an assignment of a subgradient (or an oracle call thereof) at x, and therefore, $g(x)$ is not a function or a point-to-set map.

Another useful functional notion is strong convexity: f is (uniformly) μ_f strongly convex for some $\mu_f > 0$ if

$$f(y) \geq f(x) + \langle g(x), y - x\rangle + \tfrac{\mu_f}{2}\|y - x\|^2 \quad \text{for all } x, y \in Q \text{ and } g(x) \in \partial f(x). \tag{3}$$

Nedić and Lee (2014), for example, obtain improved convergence guarantees for the stochastic mirror descent algorithm under strong convexity.

2.1. Deterministic Setting

Let us now recall the mirror decent algorithm (see Nemirovsky and Yudin 1983 and Beck and Teboulle 2003), which is also referred to as the prox subgradient method when interpreted in the space of primal variables. Mirror descent uses a differentiable convex "prox function" h to define a Bregman distance:

$$D_h(y, x) := h(y) - h(x) - \langle \nabla h(x), y - x\rangle \quad \text{for all } x, y \in Q. \tag{4}$$

The Bregman distance is used in the computation of the mirror descent update:

$$x^{i+1} \leftarrow \arg\min_{x \in Q} \left\{ f(x^i) + \langle g(x^i), x - x^i\rangle + \tfrac{1}{t_i} D_h(x, x^i) \right\},$$

where $\{t_i\}$ is the sequence of step sizes for the scheme. A formal statement of the mirror descent algorithm is presented in Algorithm 1. The traditional setup requires that h is one strongly convex with respect to the given norm $\|\cdot\|$, and in this setup, one can prove that, after k iterations, it holds for any $x \in Q$ that

$$\min_{0 \leq i \leq k} f(x^i) - f(x) \quad \leq \quad \frac{\frac{1}{2} M_f^2 \sum_{i=0}^k t_i^2 + D_h(x, x^0)}{\sum_{i=0}^k t_i}, \tag{5}$$

which leads to an $O(1/\sqrt{k})$ sublinear rate of convergence using an appropriately chosen step size sequence $\{t_i\}$ (see Beck and Teboulle 2003).

Algorithm 1 (Deterministic Mirror Descent Algorithm with Bregman Distance $D_h(\cdot, \cdot)$)

Initialize. Initialize with $x^0 \in Q$. Let h be a given convex function.

At iteration i:

Perform updates. Compute $g(x^i) \in \partial f(x^i)$, determine step size t_i, and compute update:

$$x^{i+1} \leftarrow \arg\min_{x \in Q} \{ f(x^i) + \langle g(x^i), x - x^i\rangle + \tfrac{1}{t_i} D_h(x, x^i) \}. \tag{6}$$

Notice in (6) by construction that the update requires the capability to solve instances of a "linearized subproblem" (LS) of the general form

$$\text{LS}: \quad x_{\text{new}} \leftarrow \arg\min_{x \in Q} \{\langle c, x\rangle + h(x)\} \tag{7}$$

for suitable iteration-specific values of c. Indeed, (6) is an instance of (7) with $c = t_i g(x^i) - \nabla h(x^i)$ at iteration i. It is especially important to note that the mirror descent update (6) is somewhat meaningless absent the capability to efficiently solve (7), a fact that we return to later. In the usual design and implementation of mirror descent for solving (1), one attempts to specify the norm $\|\cdot\|$ and the one-strongly convex prox function h in consideration of properties of the feasible domain Q while also ensuring that the LS subproblem (7) is efficiently solvable.

Notice that the mirror descent algorithm (Algorithm 1) itself does not require the traditional setup that h be one strongly convex for some particular norm; rather, this requirement is part of the traditional *analysis*. As we will see, we can instead analyze mirror descent by considering the intrinsic ways that f and $D_h(\cdot, \cdot)$ are related functionally in a manner that is constructive in terms of actual algorithm design and implementation. Furthermore, this is in the same spirit as was done in the predecessor paper (Lu et al. 2018).

2.2. Stochastic Setting

For some convex functions, computing an exact subgradient at $x \in Q$ may be expensive or even intractable, but sampling a random stochastic estimate of a subgradient at x, which we denote by $\tilde{g}(x)$, may be easy. We say that $\tilde{g}(x)$ is an unbiased stochastic subgradient if $\mathbb{E}[\tilde{g}(x)|x] \in \partial f(x)$. The usefulness of a stochastic subgradient methodology is easily seen in the context of machine and statistical learning problems. A prototypical learning problem is to compute an approximate solution of the following empirical loss minimization problem:

$$\begin{aligned} \min_x \quad & f(x) := \tfrac{1}{n}\sum_{j=1}^{n} f_j(x) \\ \text{s.t.} \quad & x \in Q, \end{aligned} \tag{8}$$

where f_j is a nondifferentiable convex loss function associated with sample j for $j = 1, \ldots, n$ data samples. When $n \gg 0$, the standard subgradient method needs to evaluate n subgradients to compute a subgradient of f, which can be prohibitively expensive. A typical alternative is to compute a stochastic subgradient. Letting x^i denote the ith iterate, at iteration i, a single sample index $\tilde{j}$ is drawn uniformly and independently on $\{1, \ldots, n\}$, and then a subgradient $\tilde{g} \in \partial f_{\tilde{j}}(x^i)$ is computed that is used to define $\tilde{g}(x^i) := \tilde{g}$. This stochastic subgradient is then used in place of a subgradient at iteration i. Notice that, by construction, $\tilde{g}(x^i)$ is a conditional random variable given x^i and that $\tilde{g}(x^i)$ is an unbiased stochastic subgradient, namely, $\mathbb{E}[\tilde{g}(x^i)|x^i] \in \partial f(x^i)$.

A stochastic version of mirror descent is presented in Algorithm 2. The structure of stochastic mirror descent is identical to that of mirror descent, the only difference being that the stochastic estimate of a subgradient $\tilde{g}(x^i)$ replaces the exact subgradient $g(x^i)$ in Algorithm 2.

Algorithm 2 (Stochastic Mirror Descent Algorithm with Bregman Distance $D_h(\cdot,\cdot)$)

Initialize. Initialize with $x^0 \in Q$. Let h be a given convex differentiable function.

At iteration i:

Perform updates. Compute a stochastic subgradient $\tilde{g}(x^i)$, determine step size t_i, and compute update:

$$x^{i+1} \leftarrow \arg\min_{x \in Q} \{ f(x^i) + \langle \tilde{g}(x^i), x - x^i \rangle + \tfrac{1}{t_i} D_h(x, x^i) \}.$$

A standard condition that is required in the traditional convergence analysis for stochastic mirror descent (as well as other stochastic first-order methods) is that there exists $G_f > 0$ for which

$$\mathbb{E}[\|\tilde{g}(x)\|_*^2 | x] \le G_f^2, \text{ for any } x \in Q. \tag{9}$$

For notational convenience, we say that f is G_f stochastically continuous if (9) holds. In Nedić and Lee (2014), they developed convergence results for stochastic mirror descent (Algorithm 2). Under the conditions that (i) f is G_f stochastically continuous, (ii) h is a differentiable and μ_h strongly convex function on Q, and (iii) Q is a closed bounded set, equation 27 of Nedić and Lee (2014) shows the following convergence result using step sizes $t_i = \sqrt{\frac{\mu_h D_{\max}}{G_f^2(i+1)}}$:

$$\mathbb{E}\left[f(\bar{x}^k) \right] - f^* \quad \le \quad \frac{3 G_f \sqrt{D_{\max}}}{2\sqrt{\mu_h (k+1)}}, \tag{10}$$

where $\bar{x}^k := \frac{1}{\sum_{i=0}^k t_i} \sum_{i=0}^k t_i x^i$ and $D_{\max} := \max_{x,y \in Q} D_h(x,y)$.

Furthermore, if also (i) f is μ_f strongly convex and (ii) h is L_h smooth, theorem 1 of Nedić and Lee (2014) shows that, with step sizes $t_i = \frac{2L_h}{\mu_f (i+1)}$, it holds that

$$\mathbb{E}\left[f(\check{x}^k) \right] - f^* \quad \le \quad \frac{2 G_f^2 L_h}{\mu_f (k+1) \mu_h}, \tag{11}$$

where $\check{x}^k := \frac{2}{(i+1)(i+2)} \sum_{i=0}^k (i+1) x^i$.

3. Relative Continuity

In this section, we introduce our definition of relative continuity of a function f—actually two different definitions: one for the deterministic setting and another for the stochastic setting. The starting point is a reference function h, which is a given differentiable convex function on Q that is used to construct the usual

Bregman distance $D_h(\cdot,\cdot)$ (4) and used as part of the mirror descent update (6). However, we point out for emphasis that, unlike the traditional setup, there are no assumptions on h (such as strong or strict convexity).

3.1. Deterministic Setting

Consider the objective function f of (1). We define relative continuity of f relative to the reference function h using the Bregman distance $D_h(\cdot,\cdot)$ of h as follows.

Definition 3.1. f is M relative continuous with respect to the reference function h on Q if, for any $x, y \in Q$, $x \neq y$, and $g(x) \in \partial f(x)$, it holds that

$$\|g(x)\|_* \leq \frac{M\sqrt{2D_h(y,x)}}{\|y-x\|}. \tag{12}$$

(In the particular case when $h(x) = \frac{1}{2}\|x\|_2^2$, the Bregman distance is $D_h(y,x) = \frac{1}{2}\|y-x\|_2^2$, and the relative continuity condition (12) becomes $\|g(x)\|_2 \leq M$, which corresponds to the standard definition of Lipschitz continuity (2) for the ℓ_2 norm.)

We can rewrite (12) as

$$\|g(x)\|_*^2 \leq M^2 \frac{D_h(y,x)}{\frac{1}{2}\|y-x\|^2}, \tag{13}$$

which states that the square of the norm of any subgradient is bounded by the ratio of the Bregman distance $D_h(y,x)$ to $\frac{1}{2}\|y-x\|^2$.

The following proposition presents the "key property" of an M-relative continuous function that is used in the proofs of results to follow.

Proposition 3.1 (Key Property of M Relative Continuity). *If f is M relative continuous with respect to the reference function h, then for any $t>0$, it holds for all $x, y \in Q$ and $g(x) \in \partial f(x)$ that*

$$\tfrac{1}{t} D_h(y,x) + \langle g(x), y-x\rangle + \tfrac{1}{2} tM^2 \geq 0\,. \tag{14}$$

Proof. If f is M relative continuous with respect to h, then for any $t>0$, it follows that

$$-\langle g(x), y-x\rangle \leq \|g(x)\|_*\|y-x\| \leq M\sqrt{2D_h(y,x)} \leq \tfrac{1}{2}tM^2 + \frac{D_h(y,x)}{t},$$

where the last inequality is an application of the arithmetic-geometric mean inequality. The proof follows by rearranging terms. □

The key property (14) is what is used in the proofs of results herein, and therefore, we could define M relative continuity using (14) instead of (12). Furthermore, (14) is independent of any norm structure, and therefore, it is attractive for its generality. However, we use the definition (12), because it leads to easy verification of M relative continuity in practical instances, as is shown in Section 5.

The following proposition presents some scaling and additivity properties of relative continuity.

Proposition 3.2 (Additivity of Relative Continuity).

1. *If f is M relative continuous with respect to h, then for any $\alpha>0$, f is $\frac{M}{\alpha}$ relative continuous with respect to $\alpha^2 h$.*
2. *If f is M relative continuous with respect to h, then for any $\alpha>0$, αf is M relative continuous with respect to $\alpha^2 h$.*
3. *If f_j is M relative continuous with respect to h_j for $j=1,\ldots,n$, then $\sum_{j=1}^n f_j$ is $\sqrt{n}M$ relative continuous with respect to $\sum_{j=1}^n h_j$.*
4. *If f_j is M_j relative continuous with respect to h_j for $j=1,\ldots,n$, then for $\alpha_j>0$ and $M>0$, it holds that $\sum_{j=1}^n \alpha_j f_j$ is $\sqrt{n}M$ relative continuous with respect to $\sum_{j=1}^n \frac{\alpha_j^2}{\beta_j^2} h_j$ with $\beta_j := \frac{M}{M_j}$.*

Proof. Let $x, y \in Q$, $x \neq y$, and $g(x) \in \partial f(x)$.

1. It holds that

$$\|g(x)\|_* \leq \frac{M\sqrt{2D_h(y,x)}}{\|y-x\|} = \frac{\frac{M}{\alpha}\sqrt{2D_{\alpha^2 h}(y,x)}}{\|y-x\|},$$

which establishes the result.

2. Notice that $g(x)$ is a subgradient of $f(x)$ if and only if $\alpha g(x)$ is a subgradient of $\alpha f(x)$, whereby

$$\|\alpha g(x)\|_* = \alpha\|g(x)\|_* \le \frac{\alpha M\sqrt{2D_h(y,x)}}{\|y-x\|} = \frac{M\sqrt{2D_{\alpha^2 h}(y,x)}}{\|y-x\|},$$

which establishes the result.

3. Any subgradient of $\sum_{j=1}^n f_j$ at x can be written as $\sum_{j=1}^n g_j(x)$, where $g_j(x) \in \partial f_j(x)$ for $j = 1,\dots,n$ (see theorem B.21 of Bertsekas 1999). From the triangle inequality and the relative continuity of f_j, we have

$$\left\|\sum_{j=1}^n g_j(x)\right\|_* \le \sum_{j=1}^n \|g_j(x)\|_* \le \frac{M\left(\sum_{j=1}^n \sqrt{2D_{h_j}(y,x)}\right)}{\|y-x\|} \le \frac{\sqrt{n}M\left(\sqrt{2\sum_{j=1}^n D_{h_j}(y,x)}\right)}{\|y-x\|} = \frac{\sqrt{n}M\sqrt{2D_{h_1+\cdots+h_n}(y,x)}}{\|y-x\|},$$

where the third inequality is an application of the ℓ_1/ℓ_2-norm inequality applied to the n-tuple $(\sqrt{2D_{h_1}(y,x)},\dots,\sqrt{2D_{h_n}(y,x)})$.

4. It follows from part (2) that $\alpha_j f_j$ is M_j continuous relative to $\alpha_j^2 h_j$. Thus, $\alpha_j f_j$ is also M continuous relative to $\frac{\alpha_j^2}{\beta_j^2}h_j$ from part (1), whereby the proof is finished by utilizing part (3). □

We also make use of the notion of "relative strong convexity," which was introduced in Lu et al. (2018), and it is used here in some of the convergence guarantee analyses.

Definition 3.2. f is μ strongly convex relative to h on Q if there is a scalar $\mu \ge 0$ such that, for any $x, y \in \operatorname{int} Q$ and any $g(x) \in \partial f$, it holds that

$$f(y) \ge f(x) + \langle g(x), y-x\rangle + \mu D_h(y,x). \tag{15}$$

In Lu et al. (2018), it was shown that the notion of relative strong convexity embodied in Definition 3.2 is the natural way to define strong convexity in the context of mirror descent and similar algorithms and leads to linear convergence of mirror descent in the smooth setting. In the nonsmooth setting, we will show in Theorem 4.2 that relative strong convexity improves the convergence of mirror descent from $O(1/\sqrt{k})$ to $O(1/k)$.

3.2. Stochastic Setting

For $x \in Q$, let $\tilde{g}(x)$ denote a random stochastic estimate of a subgradient of f at x. Extending the definition of relative continuity from the deterministic setting, we define stochastic relative continuity as follows.

Definition 3.3. f is G stochastically relative continuous with respect to the reference function h on Q for some $G > 0$ if f together with the oracle to compute a stochastic subgradient satisfies the following.

1. Unbiasedness property: $\mathbb{E}[\tilde{g}(x)|x] \in \partial f(x)$.
2. Boundedness property: $\mathbb{E}[\|\tilde{g}(x)\|_*^2|x] \le G^2 \frac{D_h(y,x)}{\frac{1}{2}\|y-x\|^2}$ for all $x, y \in Q$ and $x \ne y$.

(In the particular case when $h(x) = \frac{1}{2}\|x\|_2^2$, the Bregman distance is $D_h(y,x) = \frac{1}{2}\|y-x\|_2^2$, whereby the stochastically relative continuity boundedness property becomes $\mathbb{E}[\|\tilde{g}(x)\|_2^2|x] \le G^2$ for all $x \in Q$, which corresponds to the standard condition (9) for the ℓ_2 norm.)

For $x \in Q$, define

$$\tilde{M}(x) := \|\tilde{g}(x)\|_* \max_{y\in Q, y\ne x} \frac{\|y-x\|}{\sqrt{2D_h(y,x)}}. \tag{16}$$

Notice for a given x that $\max_{y\in Q, y\ne x} \frac{\|y-x\|}{\sqrt{2D_h(y,x)}}$ is a deterministic quantity, and therefore, $\tilde{M}(x)$ is a conditional random variable (given x) that is defined on the same probability space as $\tilde{g}(x)$. Clearly, if f is G stochastically relative continuous, we have by the boundedness property that, for any $x \in Q$,

$$\mathbb{E}[\tilde{M}(x)^2|x] = \mathbb{E}[\|\tilde{g}(x)\|_*^2|x] \max_{y\in Q, y\ne x} \frac{\|y-x\|^2}{2D_h(y,x)} \le G^2. \tag{17}$$

Exactly as in the deterministic setting, we have Proposition 3.3.

Proposition 3.3. *If f is G stochastically relative continuous with respect to the reference function h on Q, then for any $t>0$, it holds for all $x, y \in Q$ and any stochastic subgradient estimate $\tilde{g}(x)$ that*

$$\tfrac{1}{t} D_h(y,x) + \langle \tilde{g}(x), y - x\rangle + \tfrac{1}{2} t\tilde{M}^2(x) \geq 0.$$

Proof. For any $t>0$, we have

$$-\langle \tilde{g}(x), y-x\rangle \leq \|\tilde{g}(x)\|_*\|y-x\| \leq \tilde{M}(x)\sqrt{2D_h(y,x)} \leq \tfrac{1}{2}t\tilde{M}(x)^2 + \frac{D_h(y,x)}{t},$$

and the proof is finished by rearranging terms. □

4. Computational Analysis for Stochastic Mirror Descent and (Deterministic) Mirror Descent

In this section, we present computational guarantees for stochastic mirror descent (Algorithm 2) for minimizing a convex function f that is G stochastically relative continuous with respect to a given reference function h. We also present computational guarantees for (deterministic) mirror descent (Algorithm 1) when f is M relative continuous with respect to a reference function h, which follows as a special case of the stochastic setting.

We begin by recalling the standard three-point property for optimization using Bregman distances.

Lemma 4.1 (Three-Point Property (Tseng 2008)). *Let $\phi(x)$ be a convex function, and let $D_h(\cdot,\cdot)$ be the Bregman distance for h. For a given vector z, let*

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

Then,

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

Let us denote the (primitive) random variable at the ith iteration of the stochastic mirror descent algorithm (Algorithm 2) by γ_i (i.e., γ_i is the random variable that determines the [stochastic] subgradient $\tilde{g}(x^i)$ at iterate x^i in the stochastic mirror descent algorithm). Then, x^{i+1} is computed according to the update of the stochastic mirror descent algorithm, whereby x^{i+1} is a random variable that depends on all previous values $\gamma_0, \ldots, \gamma_i$, and we denote this string of random variables by

$$\xi_i := \{\gamma_0, \ldots, \gamma_i\}.$$

The following theorem states convergence guarantees for the stochastic mirror descent algorithm in terms of expectation.

Theorem 4.1 (Convergence Bound for Stochastic Mirror Descent Algorithm). *Consider the stochastic mirror descent algorithm (Algorithm 2) with given step size sequence $\{t_i\}$. If f is G stochastically relative continuous with respect to h for some $G>0$, then the following inequality holds for all $k \geq 1$ and $x \in Q$:*

$$\mathbb{E}_{\xi_{k-1}}\left[f(\bar{x}^k)\right] - f(x) \quad \leq \quad \frac{\frac{1}{2}G^2\sum_{i=0}^k t_i^2 + D_h(x,x^0)}{\sum_{i=0}^k t_i}, \tag{18}$$

where $\bar{x}^k := \frac{1}{\sum_{i=0}^k t_i}\sum_{i=0}^k t_i x^i$.

Proof. First notice that

$$\begin{aligned}
f(x^i) + \left\langle g(x^i), x - x^i\right\rangle &= f(x^i) + \left\langle \mathbb{E}_{\gamma_i}[\tilde{g}(x^i)|x^i], x - x^i\right\rangle \\
&= f(x^i) + \mathbb{E}_{\gamma_i}\left[\left\langle \tilde{g}(x^i), x - x^i\right\rangle | x^i\right] \\
&\geq f(x^i) + \mathbb{E}_{\gamma_i}\left[\left\langle \tilde{g}(x^i), x^{i+1} - x^i\right\rangle + \tfrac{1}{t_i}D_h(x^{i+1},x^i) + \tfrac{1}{t_i}D_h(x,x^{i+1}) - \tfrac{1}{t_i}D_h(x,x^i)|x^i\right] \\
&\geq f(x^i) + \mathbb{E}_{\gamma_i}\left[-\tfrac{1}{2}\tilde{M}(x^i)^2 t_i + \tfrac{1}{t_i}D_h(x,x^{i+1}) - \tfrac{1}{t_i}D_h(x,x^i)|x^i\right] \\
&\geq f(x^i) - \tfrac{1}{2}G^2 t_i + \tfrac{1}{t_i}\mathbb{E}_{\gamma_i}[D_h(x,x^{i+1})|x^i] - \tfrac{1}{t_i}D_h(x,x^i),
\end{aligned} \tag{19}$$

where the first equality uses the unbiasedness of $\tilde{g}(x)$, the second equality is because of linearity, the first inequality is from the three-point property with $\phi(x) = t_i\langle \tilde{g}(x^i), x - x^i\rangle$, the second inequality uses Proposition 3.3,

and the last inequality uses (17). Because $f(x) \geq f(x^i) + \langle g(x^i), x - x^i \rangle$ from the definition of a subgradient, we have from (19)

$$f(x) \geq f(x^i) + \langle g(x^i), x - x^i \rangle \geq f(x^i) - \tfrac{1}{2}G^2 t_i + \tfrac{1}{t_i}\mathbb{E}_{\gamma_i}[D_h(x, x^{i+1})|x^i] - \tfrac{1}{t_i}D_h(x, x^i).$$

Taking expectation with respect to ξ_i on both sides of the above inequality yields

$$f(x) \geq \mathbb{E}_{\xi_{i-1}}[f(x^i)] - \tfrac{1}{2}G^2 t_i + \tfrac{1}{t_i}\mathbb{E}_{\xi_i}[D_h(x, x^{i+1})] - \tfrac{1}{t_i}\mathbb{E}_{\xi_{i-1}}[D_h(x, x^i)]. \tag{20}$$

Now rearrange and multiply through by t_i to yield

$$t_i\mathbb{E}_{\xi_{i-1}}[f(x^i) - f(x)] \leq \tfrac{1}{2}G^2 t_i^2 + \mathbb{E}_{\xi_{i-1}}[D_h(x, x^i)] - \mathbb{E}_{\xi_i}[D_h(x, x^{i+1})].$$

Summing up the above inequality over i and noting that $D_h(x, x^{k+1}) \geq 0$, we arrive at

$$\begin{aligned}\tfrac{1}{2}G^2 \sum_{i=0}^{k} t_i^2 + D_h(x, x^0) \geq \sum_{i=0}^{k} t_i\mathbb{E}_{\xi_{i-1}}[f(x^i) - f(x)] &= \mathbb{E}_{\xi_{k-1}} \sum_{i=0}^{k} t_i[f(x^i) - f(x)] \\ &\geq \left(\sum_{i=0}^{k} t_i\right)\mathbb{E}_{\xi_{k-1}}\left[f(\bar{x}^k) - f(x)\right],\end{aligned} \tag{21}$$

where the last inequality uses the convexity of f. Dividing by $\sum_{i=0}^{k} t_i$ completes the proof. □

Remark 4.1. As a direct consequence of (21), we obtain the following result, which is similar to the deterministic setting (5):

$$\mathbb{E}_{\xi_{k-1}}\left[\min_{0\leq i\leq k} f(x^i)\right] - f(x) \quad \leq \quad \frac{\frac{1}{2}G^2 \sum_{i=0}^{k} t_i^2 + D_h(x, x^0)}{\sum_{i=0}^{k} t_i}.$$

Theorem 4.1 implies the following high-probability result using a simple Markov bound.

Corollary 4.1. *Let x^* be an optimal solution of (1). Under the conditions of Theorem 4.1, for any $\delta > 0$, it holds that*

$$\mathbb{P}\left[f(\bar{x}^k) - f^* \geq \delta\right] \quad \leq \quad \frac{\frac{1}{2}G^2 \sum_{i=0}^{k} t_i^2 + D_h(x^*, x^0)}{\delta \sum_{i=0}^{k} t_i}.$$

Proof. Using the Markov inequality, we have

$$\mathbb{P}\left[f(\bar{x}^k) - f^* \geq \delta\right] \leq \frac{\mathbb{E}\left[f(\bar{x}^k) - f^*\right]}{\delta} \leq \frac{\frac{1}{2}G^2 \sum_{i=0}^{k} t_i^2 + D_h(x^*, x^0)}{\delta \sum_{i=0}^{k} t_i}. \quad \square$$

Similar to the case of traditional analysis of stochastic mirror descent, the stochastic mirror descent algorithm (Algorithm 2) leads to an $O(\frac{1}{\varepsilon^2})$ convergence guarantee (in expectation) by using an appropriate step size sequence $\{t_i\}$ as the next corollary shows.

Corollary 4.2. *Under the conditions of Theorem 4.1, for a given $\varepsilon > 0$, suppose that the step sizes are set to*

$$t_i := \frac{\varepsilon}{G^2}$$

for all i. Then, within

$$k := \left\lceil \frac{2G^2 D_h(x^*, x^0)}{\varepsilon^2} \right\rceil - 1$$

iterations of the stochastic mirror descent algorithm, it holds that

$$\mathbb{E}\left[f(\bar{x}^k)\right] - f^* \leq \varepsilon,$$

where x^* is any optimal solution of (1).

Proof. Substituting the values of t_i in (18) yields the result directly. □

Remark 4.2. Similar to the standard stochastic gradient descent scheme, the step size rule in Corollary 4.2 leads to the optimal rate of convergence provided by the bound in Theorem 4.1.

Remark 4.3. Let us now compare these results with related results of Nedić and Lee (2014). To attain an ε optimality gap, Nedić and Lee (2014) proved a bound of $\left\lceil \frac{9G_f^2 D_{\max}}{4\mu_h \varepsilon^2} \right\rceil$ iterations, which follows by rearranging (10). In addition to not requiring Lipschitz continuity of f, our bound does not require that h be strongly convex. We also do not require boundedness of the feasible region, and in most settings, $D_h(x^*, x^0) \ll D_{\max}$ even when $D_{\max} < +\infty$. Furthermore, even in the setting of (10), it holds that

$$G^2 = \mathbb{E}\big[\tilde{g}(x)^2 | x\big] \max_{y \in Q, y \neq x} \frac{\|y - x\|^2}{2D_h(y,x)} \leq \mathbb{E}\big[\|\tilde{g}(x)\|_*^2 | x\big] \tfrac{1}{\mu_h} \leq \tfrac{G_f^2}{\mu_h},$$

where the first inequality utilizes the strong convexity (in the standard sense) of h, and the second inequality is due to the assumption that f is G_f stochastically continuous (in the standard sense). Thus, we see that the bound in Corollary 4.2 improves on the bound in Nedić and Lee (2014).

In the case when f is also μ_f strongly convex relative to h (see Definition 3.2), we obtain an $O(\frac{1}{k})$ convergence guarantee in expectation, which is also similar to the traditional case of stochastic gradient descent. This is shown in the next result.

Theorem 4.2 (Convergence Bound for Stochastic Mirror Descent Algorithm Under Strong Convexity Relative to h). *Consider the stochastic mirror descent algorithm (Algorithm 2). If f is G stochastically relative continuous with respect to h for some $G > 0$ and f is μ strongly convex relative to h for some $\mu > 0$ and if the step sizes are chosen as $t_i = \frac{2}{\mu(i+1)}$, then the following inequality holds for all $k \geq 1$:*

$$\mathbb{E}_{\xi_{k-1}}\big[f(\hat{x}^k)\big] - f^* \leq \frac{2G^2}{\mu(k+1)},$$

where $\hat{x}^k := \frac{2}{k(k+1)} \sum_{i=0}^{k} i \cdot x^i$.

Proof. For any $x \in Q$, it follows from the definition of μ strong convexity (15) that

$$f(x) \geq f(x^i) + \big\langle g(x^i), x - x^i \big\rangle + \mu D_h(x, x^i).$$

Combining the above inequality with (19) yields

$$f(x) \geq f(x^i) - \tfrac{1}{2}G^2 t_i + \tfrac{1}{t_i}\mathbb{E}_{\gamma_i}[D_h(x, x^{i+1}) | x^i] - (\tfrac{1}{t_i} - \mu) D_h(x, x^i).$$

Substituting $t_i = \frac{2}{\mu(i+1)}$ and multiplying by i in the above inequality yields

$$\begin{aligned} i\big(f(x^i) - f(x)\big) &\leq \tfrac{G^2 i}{\mu(i+1)} + \tfrac{\mu}{2}\big(i(i-1)D_h(x, x^i) - i(i+1)\mathbb{E}_{\gamma_i}[D_h(x, x^{i+1}) | x^i]\big) \\ &\leq \tfrac{G^2}{\mu} + \tfrac{\mu}{2}\big(i(i-1)D_h(x, x^i) - i(i+1)\mathbb{E}_{\gamma_i}[D_h(x, x^{i+1}) | x^i]\big). \end{aligned}$$

Taking expectation over ξ_{i-1} and summing up the above inequality over i then yields

$$\left(\sum_{i=1}^{k} i\right)\mathbb{E}_{\xi_{k-1}}[f(\hat{x}^k) - f(x)] \leq \sum_{i=1}^{k} i\big(\mathbb{E}_{\xi_{i-1}}[f(x^i)] - f(x)\big) \leq \frac{kG^2}{\mu} - k(k+1)\left(\frac{\mu}{2}\right)\mathbb{E}_{\xi_k}[D_h(x, x^{k+1})] \leq \frac{kG^2}{\mu},$$

where the first inequality uses the convexity of f and the observation that $\mathbb{E}_{\xi_{i-1}} f(x^i) = \mathbb{E}_{\xi_{k-1}} f(x^i)$ for $i \leq k$. Taking $x = x^*$, where x^* is an optimal solution of (1), the proof is completed by noticing that $\sum_{i=1}^{k} i = \frac{k(k+1)}{2}$. □

Remark 4.4. It may not be easy to find cases when the objective function is both G stochastically relative continuous and μ relatively strongly convex relative to h. However, as long as these properties are satisfied along the path of iterates or around the minimum, one can achieve the faster convergence of Theorem 4.2.

Remark 4.5. Let us also compare the computational guarantee of Theorem 4.2 with the results in Nedić and Lee (2014). To attain an ε optimality gap, Nedić and Lee (2014) proved the bound (11). Notice that we do not require that f is uniformly Lipschitz continuous, that h is strongly convex in the traditional sense, or that h is uniformly smooth. However, even if these requirements hold, it follows from Remark 4.3 that $G^2 \leq \frac{G_f^2}{\mu_h}$, and it also holds that

$$D_f(x,y) \geq \tfrac{\mu_f}{2}\|x-y\|^2 \geq \tfrac{\mu_f}{L_h} D_h(x,y),$$

where the first inequality utilizes that f is μ_f strongly convex and the second inequality h is L_h smooth in the standard sense. Thus, f is at least $\mu = \frac{\mu_f}{L_h}$ strongly convex relative to h (this follows by applying proposition 1.1 in Lu et al. 2018). Therefore, even under the stronger requirements of Nedić and Lee (2014), Theorem 4.2 improves on the corresponding result in Nedić and Lee (2014).

We end this section with a discussion of the deterministic setting, namely, the (deterministic) mirror descent algorithm (Algorithm 1). Suppose that there is no stochasticity in the computation of subgradients. We can cast this as an instance of the stochastic mirror descent algorithm (Algorithm 2), wherein $\tilde{g}(x) = g(x) \in \partial f(x)$ for all $x \in Q$. In this case, relative stochastic continuity (Definition 3.3) is equivalent to relative continuity (Definition 3.1) with the same constant. Thus, deterministic mirror descent is a special case of stochastic mirror descent, and we have the following computational guarantees as special cases of the stochastic case.

Theorem 4.3 (Convergence Bound for Deterministic Mirror Descent Algorithm). *Consider the (deterministic) mirror descent algorithm (Algorithm 1). If f is M relative continuous with respect to h for some $M>0$, then for all $k \geq 1$ and $x \in Q$, the following inequality holds:*

$$f(\bar{x}^k) - f(x) \quad \leq \quad \frac{\frac{1}{2}M^2 \sum_{i=0}^k t_i^2 + D_h(x,x^0)}{\sum_{i=0}^k t_i},$$

where $\bar{x}^k := \frac{1}{\sum_{i=0}^k t_i} \sum_{i=0}^k t_i x^i$.

Corollary 4.3. *Under the conditions of Theorem 4.3, for a given $\varepsilon > 0$, suppose that the step sizes are set to*

$$t_i := \frac{\varepsilon}{M^2}$$

for all i. Then, within

$$k := \left\lceil \frac{2M^2 D_h(x^*,x^0)}{\varepsilon^2} \right\rceil - 1$$

iterations of deterministic mirror descent, it holds that

$$f(\bar{x}^k) - f^* \leq \varepsilon,$$

where $\bar{x}^k := \frac{1}{\sum_{i=0}^k t_i} \sum_{i=0}^k t_i x^i$ *and x^* is any optimal solution of (1).*

Theorem 4.4 (Convergence Bounds for Deterministic Mirror Descent with Strong Relative Convexity). *Consider the deterministic mirror descent algorithm (Algorithm 1). If f is M relative continuous with respect to h for some $M>0$ and f is μ strongly convex relative to h for some $\mu>0$ and if the step sizes are chosen as $t_i = \frac{2}{\mu(i+1)}$, then the following inequality holds for all $k \geq 1$:*

$$f(\hat{x}^k) - f^* \leq \frac{2M^2}{\mu(k+1)},$$

where $\hat{x}^k := \frac{2}{k(k+1)} \sum_{i=0}^k i \cdot x^i$.

5. Specifying a Reference Function h with Relative Continuity for Mirror Descent

Let us discuss using either deterministic or stochastic mirror descent (Algorithm 1 or Algorithm 2) for solving the optimization problem (1) with objective function f that is M relative continuous or G stochastically relative continuous (respectively) with respect to the reference function h. To efficiently execute the update step in Algorithm 1 and/or Algorithm 2, we need h to be such that the LS (7) is efficiently solvable for any given c. Therefore, to execute mirror descent for solving (1) using Algorithm 1 or Algorithm 2, we need to specify a differentiable convex reference function h that has the following two properties.

1. f is M relative continuous (or G stochastically relative continuous) with respect to h on Q for M (or G) that is easy to determine.
2. The LS (7) has a solution, and the solution is efficiently computable.

We now discuss quite broadly how to construct such a reference function h with these two properties when $\|g(x)\|_*^2$ is bounded by a polynomial in $\|x\|_2$.

5.1. Deterministic Setting

Suppose that $\|g(x)\|_*^2 \le p_r(\|x\|_2)$ for all $x \in Q$ and all $g(x) \in \partial f(x)$, where $p_r(\alpha) = \sum_{i=0}^r a_i \alpha^i$ is an r-degree polynomial of α with coefficients $\{a_i\}$ that are nonnegative. Let

$$h(x) := \sum_{i=0}^{r} \tfrac{a_i}{i+2}\|x\|_2^{i+2}.$$

Then, the following proposition states that f is one relative continuous with respect to h. This implies that, no matter how fast the subgradient of f grows polynomially as $\|x\|_2 \to \infty$, f is relatively continuous with respect to the simple reference function h, although f does not exhibit uniform Lipschitz continuity.

Proposition 5.1. *f is one continuous relative to $h(x) = \sum_{i=0}^r \frac{a_i}{i+2}\|x\|_2^{i+2}$.*

Proof. Let $h_i(x) = \frac{1}{i+2}\|x\|_2^{i+2}$; then, $h(x) = \sum_{i=0}^r a_i h_i(x)$, and by the definition of Bregman distance, we have

$$\begin{aligned} D_{h_i}(y,x) &= \tfrac{1}{i+2}\|y\|_2^{i+2} - \tfrac{1}{i+2}\|x\|_2^{i+2} - \langle \|x\|_2^i x, y - x\rangle \\ &= \tfrac{1}{i+2}\big(\|y\|_2^{i+2} + (i+1)\|x\|_2^{i+2} - (i+2)\|x\|_2^i\langle x, y\rangle\big). \end{aligned}$$

Notice that

$$\begin{aligned} &\|y\|_2^{i+2} + (i+1)\|x\|_2^{i+2} - (i+2)\|x\|_2^i\langle x, y\rangle \\ &= \big(\|y\|_2^{i+2} + \frac{i}{2}\|x\|_2^{i+2} - \tfrac{i+2}{2}\|x\|_2^i\|y\|_2^2\big) + \tfrac{i+2}{2}\|x\|_2^i\big(\|x\|_2^2 + \|y\|_2^2 - 2\langle x, y\rangle\big) \\ &\ge 0 + \tfrac{i+2}{2}\|x\|_2^i\big(\|x\|_2^2 + \|y\|_2^2 - 2\langle x, y\rangle\big) \\ &= \tfrac{i+2}{2}\|x\|_2^i\|y - x\|_2^2, \end{aligned}$$

where the inequality above is an application of arithmetic-geometric mean inequality $a^\lambda b^{1-\lambda} \le \lambda a + (1-\lambda)b$ with $a = \|x\|_2^{i+2}$, $b = \|y\|_2^{i+2}$, and $\lambda = \frac{i}{i+2}$. Thus, we have

$$D_h(y,x) = \sum_{i=0}^{r} a_i D_{h_i}(y,x) \ge \tfrac{1}{2}\|y - x\|_2^2\left(\sum_{i=0}^{r} a_i\|x\|_2^i\right). \tag{22}$$

Therefore,

$$\|g(x)\|_*^2 \le p_r(\|x\|_2) = \sum_{i=0}^{r} a_i\|x\|_2^i \le \frac{D_h(y,x)}{\frac{1}{2}\|y - x\|_2^2},$$

which shows that f is one relative continuous with respect to h. □

Solving the Linearization Subproblem (7). Let us see how we can solve the linearization subproblem (7) for this class of optimization problems. The linearization subproblem (7) can be written as

$$\text{LS}: \quad \min_{x \in \mathbb{R}^n} \ \langle c, x\rangle + \sum_{i=0}^{r} \tfrac{a_i}{i+2}\|x\|_2^{i+2}, \tag{23}$$

and the first-order optimality condition is simply

$$c + \left(\sum_{i=0}^{r} a_i\|x\|_2^i\right)x = 0, \tag{24}$$

whereby $x = -\theta c$ for some scalar $\theta \geq 0$, and it remains to simply determine the value of the nonnegative scalar θ. In the case when $c = 0$, we have that $x = 0$ satisfies (24), and therefore, let us examine the case when $c \neq 0$, in which case from (24) θ must satisfy

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

which implies that θ is the unique positive root of a univariate polynomial monotone in $\theta \geq 0$. For $r \in \{0, 1, 2, 3\}$, this root can be computed in closed form. Otherwise, the root can be computed efficiently (up to machine precision) using any suitable root-finding method.

Remark 5.1. We can incorporate a simple set constraint $x \in Q$ in problem (23) provided that we can easily compute the Euclidean projection on Q. In this case, the linearization subproblem (7) can be converted to a one-dimensional convex optimization problem (see appendix A.1 of Lu et al. 2018 for details).

5.2. Stochastic Setting

In the stochastic setting, the stochastic subgradient $\tilde{g}(x)$ is a conditional random variable for a given x. Suppose that $\mathbb{E}\big[\|\tilde{g}(x)\|_*^2 | x\big] \leq p_r(\|x\|_2)$ for all $x \in Q$, where $p_r(\alpha) = \sum_{i=0}^{r} a_i \alpha^i$ is an r-degree polynomial with coefficients $\{a_i\}$ that are nonnegative. Let

$$h(x) := \sum_{i=0}^{r} \tfrac{a_i}{i+2} \|x\|_2^{i+2},$$

and similar to the deterministic case, we have Proposition 5.2.

Proposition 5.2. *f is one stochastically continuous relative to $h(x) = \sum_{i=0}^{r} \frac{a_i}{i+2} \|x\|_2^{i+2}$.*

Proof. For any $x, y \in Q$ with $x \neq y$, we have

$$\mathbb{E}\big\|\tilde{g}(x)\big\|_*^2 | x\big] \leq p_r(\|x\|_2) = \sum_{i=0}^{r} a_i \|x\|_2^i \leq \frac{2D_h(y, x)}{\|y - x\|_2^2}, \tag{25}$$

where the last inequality follows from (22), and thus, f is one stochastically continuous relative to h. □

Solving the Linearization Subproblem (7). The linear optimization subproblem is identical in structure to that in the deterministic case and therefore, can be solved as discussed at the end of Section 5.1.

5.3. Relative Continuity for Instances of SVM and IEP

Here, we examine in detail the two motivating examples stated in Section 1, namely, the SVM problem and the IEP. We first prove the following lemma, which presents upper bounds on the Bregman distances $D_h(\cdot, \cdot)$ for $h(x) = \frac{1}{3}\|x\|_2^3$ and $h(x) = \frac{1}{4}\|x\|_2^4$.

Lemma 5.1.

1. *Let $h(x) := \frac{1}{3}\|x\|_2^3$. Then, $D_h(y, x) \leq \frac{1}{3}\|y - x\|_2^2\big(\|y\|_2 + 2\|x\|_2\big)$.*
2. *Let $h(x) := \frac{1}{4}\|x\|_2^4$. Then, $D_h(y, x) \leq \frac{1}{4}\|y - x\|_2^2\big(\|y + x\|_2^2 + 2\|x\|_2^2\big)$.*

Proof.

1.

$$\begin{aligned} D_h(y, x) &= \tfrac{1}{3}\big(\|y\|_2^3 + 2\|x\|_2^3 - 3\|x\|_2\langle x, y\rangle\big) \\ &\leq \tfrac{1}{3}\big(\|y\|_2^3 + 2\|x\|_2^3 - 3\|x\|_2\langle x, y\rangle - 2\|y\|_2\langle y, x\rangle + 2\|y\|_2^2\|x\|_2 - \|x\|_2\langle x, y\rangle + \|y\|_2\|x\|_2^2\big) \\ &= \tfrac{1}{3}\|y - x\|_2^2\big(\|y\|_2 + 2\|x\|_2\big), \end{aligned}$$

where the first equality follows from simplifying and combining terms, the inequality follows from applying the Cauchy–Schwarz inequality twice, and the final equality is from simplifying and combining terms.

2.

$$\begin{aligned} D_h(y,x) &= \tfrac{1}{4}\big(\|y\|_2^4 + 3\|x\|_2^4 - 4\|x\|_2^2\langle x,y\rangle\big) \\ &\le \tfrac{1}{4}\big(\|y\|_2^4 + 3\|x\|_2^4 - 4\|x\|_2^2\langle x,y\rangle + 4\|x\|_2^2\|y\|_2^2 - 4\langle x,y\rangle^2\big) \\ &= \tfrac{1}{4}\|y-x\|_2^2\big(\|y+x\|_2^2 + 2\|x\|_2^2\big), \end{aligned}$$

where the first equality follows from simplifying and combining terms, the inequality follows from applying the Cauchy–Schwarz inequality once, and the final equality is from simplifying and combining terms. □

SVM. The SVM is an important supervised learning model for binary classification in machine learning. The SVM optimization problem for binary classification is

$$SVM: \quad \min_x f(x) := \tfrac{1}{n}\sum_{i=1}^n \max\{0, 1 - y_i x^T w_i\} + \tfrac{\lambda}{2}\|x\|_2^2, \tag{26}$$

where w_i is the input feature vector of sample i and $y_i \in \{-1,1\}$ is the label of sample i. Notice that f is not differentiable due to the presence of the hinge loss terms in the summation, and f is also not Lipschitz continuous due to the presence of the ℓ_2-norm regularization term; thus, we cannot directly utilize typical subgradient or gradient schemes and their associated computational guarantees in the analysis of (26). Researchers have developed various approaches to overcome this limitation. For example, Duchi and Singer (2009) introduced a splitting subgradient-type method, where the basic idea is to split the loss function and the regularization terms. Yu et al. (2010) introduced a quasi-Newton method, where they do not need to worry about the unbounded subgradient. Another approach is to a priori constrain x to be in an ℓ_2 ball of radius R for R sufficiently large so that the ball contains the optimal solution and to project onto this ball at each iteration; in this approach, f is Lipschitz continuous in the amended feasible region (see Shalev-Shwartz et al. 2011). Indeed, one can show using quadratic optimization optimality conditions that it suffices to set $R = \min\{\frac{1}{\lambda}(\frac{1}{n}\sum_{i=1}^n \|w_i\|_2), \sqrt{2/\lambda}\}$ (see Section 5.3), wherein the modulus of Lipschitz continuity in the amended feasible region is at most $M \le \frac{1}{n}\sum_{i=1}^n \|w_i\|_2 + \min\{\sqrt{2\lambda}, \frac{1}{n}\sum_{i=1}^n \|w_i\|_2\}$. Furthermore, in Lacoste-Julien et al. (2012), the authors show that, if the initial point is within a suitably chosen large ball, then stochastic subgradient descent with a small step size ensures in expectation that all iterates are in the large ball, which then ensures that the norms of all subgradients are bounded in expectation.

Let us see how we can directly use the constructs of relative continuity to tackle the SVM problem with a suitably designed version of stochastic mirror descent—without any projection step to a ball. We can rewrite the objective function of (26) as

$$f(x) = \tfrac{1}{n}\sum_{j=1}^n f_j(x),$$

where $f_j(x) = \max\{0, 1 - y_j x^T w_j\} + \frac{\lambda}{2}\|x\|_2^2$. We consider computing a stochastic estimate of the subgradient of f by using a single sample index $\tilde{j}$ drawn randomly from $\{1,\dots,n\}$, namely, $\tilde{g}(x) \in \partial f_{\tilde{j}}(x)$, where $\tilde{j}$ is drawn uniformly at random from $\{1,\dots,n\}$. Then, $\|\tilde{g}(x)\|_2^2 \le (\lambda\|x\|_2 + \|w_{\tilde{j}}\|_2)^2$, whereby

$$\mathbb{E}[\|\tilde{g}(x)\|_2^2 | x] \le \lambda^2\|x\|_2^2 + \frac{2\lambda}{n}\left(\sum_{i=1}^n \|w_i\|_2\right)\|x\|_2 + \frac{1}{n}\sum_{i=1}^n \|w_i\|_2^2,$$

and notice that the right-hand side is a polynomial in $\|x\|_2$ of degree $r = 2$. If we choose the reference function h as

$$h(x) := \frac{\lambda^2}{4}\|x\|_2^4 + \frac{2\lambda}{3n}\left(\sum_{i=1}^n \|w_i\|_2\right)\|x\|_2^3 + \frac{1}{2n}\left(\sum_{i=1}^n \|w_i\|_2^2\right)\|x\|_2^2, \tag{27}$$

it follows from Proposition 5.2 that f is one stochastically continuous relative to $h(x)$.

Proposition 5.3 (Computational Guarantees for Stochastic Mirror Descent for the SVM Problem (26)). *Consider applying the stochastic mirror descent algorithm (Algorithm 2) to the support vector machine problem (26) using the reference function (27). For an absolute optimality tolerance value $\varepsilon > 0$ and using the constant step sizes $t_i := \varepsilon$, let the algorithm be run for*

$$k := \left\lceil \frac{\|x^* - x^0\|^2\left(3\lambda^2\left(\|x^* + x^0\|_2^2 + 2\|x^0\|_2^2\right) + \frac{8\lambda}{n}\left(\sum_{i=1}^n \|w_i\|_2\right)\left(\|x^*\|_2 + 2\|x^0\|_2\right) + \frac{6}{n}\left(\sum_{i=1}^n \|w_i\|_2^2\right)\right)}{6\varepsilon^2} \right\rceil - 1$$

iterations, where x^* is the optimal solution of (26). Then, it holds that

$$\mathbb{E}\left[f(\bar{x}^k) - f^*\right] \le \varepsilon,$$

where $\bar{x}^k := \frac{1}{k+1}\sum_{i=0}^k x^i$.

Proof. We showed above (using Proposition 5.2) that f is one stochastically continuous relative to h defined in (27). Furthermore, applying Lemma 5.1, it follows that

$$D_h(x^*, x^0) \le \frac{\lambda^2}{4}\|x^* - x^0\|_2^2\left(\|x^* + x^0\|_2^2 + 2\|x^0\|_2^2\right) + \frac{2\lambda}{3n}\left(\sum_{i=1}^n \|w_i\|_2\right)\left(\|x^*\|_2 + 2\|x^0\|_2\right) + \frac{1}{2n}\left(\sum_{i=1}^n \|w_i\|_2^2\right)\|x^* - x^0\|_2^2.$$

The proof is finished by substituting these values into the computational guarantee of Corollary 4.2. □

IEP.[1] Consider the problem of computing a point $x \in \mathbb{R}^m$ in the intersection of n ellipsoids, namely,

$$x \in \mathcal{Q} := \mathcal{Q}_1 \cap \mathcal{Q}_2 \cap \cdots \cap \mathcal{Q}_n, \tag{28}$$

where $\mathcal{Q}_i = \{x \in \mathbb{R}^m : \frac{1}{2}x^T A_i x + b_i x + c_i \le 0\}$ and $A_i \in \mathbb{R}^{m\times m}$ is a given symmetric positive semidefinite matrix $i = 1, \ldots, n$. This problem can be cast as a second-order cone optimization problem, and hence, it can be tackled using interior-point methods. However, interior-point methods are typically only effective when the dimensions m and/or n are of moderate size. However, another way to tackle the problem is to use a first-order method to solve the unconstrained problem

$$\text{IEP:} \quad f^* = \min_x f(x) := \max_{0\le i\le n}\{\tfrac{1}{2}x^T A_i x + b_i^T x + c_i\}, \tag{29}$$

and notice that $f(x) \le 0 \Leftrightarrow x \in \mathcal{Q}$ and $\mathcal{Q} \ne \emptyset \Leftrightarrow f^* \le 0$. However, the objective function f in (29) is both nondifferentiable and non-Lipschitz, and therefore, it falls outside of the scope of optimization problems for which traditional first-order methods are applicable. Let us see how we can use the machinery of relative continuity to tackle this problem. Let $\sigma := \max_{0\le i\le n}\|A_i\|_2^2$, where $\|A_i\|_2$ is the spectral radius of A_i; let $\rho := 2\max_{0\le i\le n}\|A_i b_i\|_2$, and let $\gamma := \max_{0\le i\le n}\|b_i\|_2^2$. Notice that, for any x and $i = 1, \ldots, n$, we have $g_i(x) := A_i x + b_i = \nabla f_i(x)$, where $f_i(x)$ is the ith term in the objective function of (29). Because $g(x) \in \partial f(x)$ if and only if $g(x)$ is a convex combination of the active gradients $\nabla f_i(x)$ (see Danskin's Theorem, proposition B.22 in Bertsekas 1999), it follows for any $g(x) \in \partial f(x)$ that

$$\|g(x)\|_2^2 \le \max_{0\le i\le n}\|A_i x + b_i\|_2^2 \le \max_{0\le i\le n}\|A_i\|_2^2\|x\|_2^2 + 2\|b_i^T A_i\|_2\|x\|_2 + \|b_i\|_2^2 \le \sigma\|x\|_2^2 + \rho\|x\|_2 + \gamma.$$

Therefore, we have $\|g(x)\|_2^2 \le p_2(\|x\|_2)$, where $p_1(\alpha) = \sigma\alpha^2 + \rho\alpha + \gamma$ is a quadratic function of α, which is a polynomial in α of degree $r = 2$. It follows from Proposition 5.1 that f is one continuous relative to the reference function

$$h(x) := \tfrac{\sigma}{4}\|x\|_2^4 + \tfrac{\rho}{3}\|x\|_2^3 + \tfrac{\gamma}{2}\|x\|_2^2. \tag{30}$$

Proposition 5.4 (Computational Guarantees for Deterministic Mirror Descent for the IEP (29)). *Consider applying the deterministic mirror descent algorithm (Algorithm 1) to the ellipsoid intersection problem (29) using the reference function (30), where $\sigma := \max_{0\le i\le n}\|A_i\|_2^2$, $\|A_i\|_2$ is the spectral radius of A_i, $\rho := 2\max_{0\le i\le n}\|A_i b_i\|_2$, and $\gamma := \max_{0\le i\le n}\|b_i\|_2^2$. For an absolute optimality tolerance value $\varepsilon > 0$ and using the constant step sizes $t_i := \varepsilon$, let the algorithm be run for*

$$k := \left\lceil \frac{\|x^* - x^0\|^2\left(3\sigma\left(\|x^* + x^0\|_2^2 + 2\|x^0\|_2^2\right) + 4\rho\left(\|x^*\|_2 + 2\|x^0\|_2\right) + 6\gamma\right)}{6\varepsilon^2} \right\rceil - 1$$

iterations, where x^* is any optimal solution of (29). Then, it holds that

$$f(\bar{x}^k) - f^* \le \varepsilon,$$

where $\bar{x}^k := \frac{1}{k+1}\sum_{i=0}^{k} x^i$.

Proof. We showed above (using Proposition 5.1) that f is one continuous relative to $h(x) = \frac{\sigma}{4}\|x\|_2^4 + \frac{\rho}{3}\|x\|_2^3 + \frac{\gamma}{2}\|x\|_2^2$. Furthermore, applying Lemma 5.1, it follows that $D_h(x^*, x^0) \le \frac{\sigma}{4}\|x^* - x^0\|_2^2\left(\|x^* + x_0\|_2^2 + 2\|x_0\|_2^2\right) + \frac{\rho}{3}\|x^* - x^0\|_2^2(\|x^*\|_2 + 2\|x^0\|_2) + \frac{\gamma}{2}\|x^* - x^0\|_2^2$. The proof is finished by substituting these values into the computational guarantee of Corollary 4.3. □

Acknowledgments
The author expresses his gratitude to Robert M. Freund for thoughtful discussions that helped motivate this work, commenting on earlier drafts of this paper, and advising on the presentation and positioning of this paper. The author also thanks Yurii Nesterov for encouraging the author's work on this topic and pointing out the application of the intersection of ellipsoids problem.

Appendix. Finite Radius Bound for SVM
Here, we derive an upper bound on the norm of an optimal solution of the SVM problem (26).

Proposition 5.5. *The optimal solution to the SVM problem (26) is in the ball $B_2(0, R)$ for $R = \min\{\frac{1}{n\lambda}\sum_{i=1}^{n}\|w_i\|_2, \sqrt{2/\lambda}\}$.*

Proof. For convenience, define $A_i := y_i w_i$ for $i = 1, \ldots, n$. Then, we can rewrite the SVM problem as the following constrained optimization problem:

$$\begin{aligned} \min_{s,x} \quad & \tfrac{1}{n} e^T s + \tfrac{\lambda}{2}\|x\|_2^2 \\ \text{s.t.} \quad & s + Ax \ge e, \\ & s \ge 0. \end{aligned}$$

Let π and β be the multipliers on the inequality constraints above. Then, the Karush-Kuhn-Tucker conditions imply, among other things, that the optimal solution x^* must satisfy

$$\pi^* + \beta^* = \frac{1}{n} e,$$

$$\lambda x^* = A^T \pi^*,$$

where $\pi^* \ge 0$ and $\beta^* \ge 0$. Define $\bar{\pi}^* = n\pi^*$. Then, $0 \le \bar{\pi}^* \le e$ and

$$\lambda\|x^*\|_2 = \|A^T\pi^*\|_2 = \tfrac{1}{n}\|A^T\bar{\pi}^*\|_2 \le \tfrac{1}{n}\sum_{i=1}^{n}\|A_i\|_2 = \tfrac{1}{n}\sum_{i=1}^{n}\|w_i\|_2,$$

which proves the first term in the definition of R. Also, we have $\frac{\lambda}{2}\|x^*\|_2^2 \le f(x^*) \le f(0) = 1$, and thus, $\|x^*\|_2 \le \sqrt{2/\lambda}$. Therefore, $\|x^*\|_2 \le \min\{\frac{1}{n\lambda}\sum_{i=1}^{n}\|w_i\|_2, \sqrt{2/\lambda}\}$, which finishes the proof. □

Endnote
[1] This problem was suggested by Nesterov (2016).

References
Bauschke HH, Bolte J, Teboulle M (2016) A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Math. Oper. Res.* 42(2):330–348.
Beck A, Teboulle M (2003) Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.* 31(3):167–175.
Bertsekas D (1999) *Nonlinear Programming* (Athena Scientific, Belmont, MA).
Bubeck S (2015) Convex optimization: Algorithms and complexity. *Foundations Trends Machine Learn.* 8(3/4):231–357.
Duchi J, Singer Y (2009) Efficient online and batch learning using forward backward splitting. *J. Machine Learn. Res.* 10(December):2899–2934.

Hanzely F, Richtarik P (2018) Fastest rates for stochastic mirror descent methods. Working paper, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.

Lacoste-Julien S, Schmidt M, Bach F (2012) A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method. Working paper, Cornell University, Ithaca, NY.

Lu H, Freund RM, Nesterov Y (2018) Relatively smooth convex optimization by first-order methods, and applications. *SIAM J. Optim.* 28(1): 333–354.

Nedić A, Lee S (2014) On stochastic subgradient mirror-descent algorithm with weighted averaging. *SIAM J. Optim.* 24(1):84–107.

Nemirovsky AS, Yudin DB (1983) *Problem Complexity and Method Efficiency in Optimization* (Wiley, New York).

Nesterov Y (2003) *Introductory Lectures on Convex Optimization: A Basic Course* (Kluwer Academic Publishers, Boston).

Nesterov Y (2016) Personal communication with the author during visit to Massachusetts Institute of Technology, Cambridge, MA, May.

Shalev-Shwartz S, Singer Y, Srebro N (2011) Pegasos: Primal estimated sub-gradient solver for SVM. *Math. Programming* 127(1):3–30.

Tseng P (2008) On accelerated proximal gradient methods for convex-concave optimization. Technical report, MIT, Cambridge, MA.

Van Nguyen Q (2017) Forward-backward splitting with Bregman distances. *Vietnam J. Math.* 45(3):519–539.

Yu J, Vishwanathan SVN, Günter S, Schraudolph NN (2010) A quasi-Newton approach to nonsmooth convex optimization problems in machine learning. *J. Machine Learn. Res.* 11(March):1145–1200.

Zhou Y, Liang Y, Shen L (2016) A unified approach to proximal algorithms using Bregman distance. Technical report, Syracuse University, Syracuse, NY.