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

Haihao Lu*

October 12, 2017

Abstract

The usual approach to developing and analyzing first-order methods for non-smooth (stochastic or deterministic) convex optimization assumes that the objective function is uniformly Lipschitz continuous with parameter M_f. However, in many settings the non-differentiable convex function $f(\cdot)$ is not uniformly Lipschitz continuous – for example (i) the classical support vector machine (SVM) 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" $h(\cdot)$ (that should be computationally tractable for algorithms), and we show that many non-differentiable convex functions are relatively continuous with respect to a correspondingly fairly-simple reference function $h(\cdot)$. We also similarly develop a notion of "relative stochastic continuity" for the stochastic setting. We analysis two standard algorithms – the (deterministic) mirror descent algorithm and the stochastic mirror descent algorithm – for solving optimization problems in these two new settings, and we develop for the first time computational guarantees for instances where the objective function is not uniformly Lipschitz continuous. This paper is a companion paper for non-differentiable convex optimization to the recent paper by Lu, Freund, and Nesterov, which developed similar sorts of results for differentiable convex optimization.

1 Introduction and Motivation

The usual approach to developing and analyzing first-order methods for non-differentiable convex optimization (which we will review shortly) assumes that the objective function is uniformly Lipschitz continuous in both deterministic and stochastic settings. However, in many settings the non-differentiable convex function $f(\cdot)$ is not uniformly Lipschitz continuous. The following are two examples:

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

$$x \in \mathcal{Q} := \mathcal{Q}_1 \cap \mathcal{Q}_2 \cap \cdots \cap \mathcal{Q}_n \ , \tag{1}$$

*MIT Department of Mathematics and MIT Operations Research Center, 77 Massachusetts Avenue, Cambridge, MA 02139 (haihao@mit.edu). The author's research is supported by AFOSR Grant No. FA9550-15-1-0276 and the MIT-Belgium Université Catholique de Louvain Fund.

[1] This problem was suggested by Nesterov [12].

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

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

and notice that $f(x) \leq 0 \Leftrightarrow x \in \mathcal{Q}$, and $\mathcal{Q} \neq \emptyset \Leftrightarrow f^* \leq 0$. However, the objective function $f(\cdot)$ in (2) is both non-differentiable and non-Lipschitz, and so it falls outside of the scope of standard classes of optimization problems for which first-order methods are guaranteed to work. Using the machinery developed in this paper, we will show (Proposition 5.3) that in order to compute an ε-optimal solution, it suffices to use at most

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

iterations of the Mirror Descent Algorithm, where x^* is an optimal solution of (2), x^0 is the initial point, $\sigma := \max_{1\leq i\leq n} \|A_i\|_2^2$ where $\|A_i\|_2$ is the spectral radius of A_i, $\rho := 2\max_{1\leq i\leq n} \|A_i b_i\|_2$ and $\gamma := \max_{0\leq i\leq n} \|b_i\|_2^2$.

Support Vector Machine (SVM). The Support Vector Machine (SVM) is an important supervised learning model for binary classification in machine learning. The SVM optimization problem for binary classification is:

$$\text{SVM:} \quad f^* \ = \ \min_x f(x) := \tfrac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i x^T w_i\} + \tfrac{\lambda}{2}\|x\|_2^2 \ , \tag{3}$$

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 $f(\cdot)$ is not differentiable due to the presence of the hinge loss summation term, and $f(\cdot)$ 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 (3). Researchers have developed various approaches to overcome this limitation. For example, [6] introduced a splitting subgradient-type method, where the basic idea is to split the loss function and the regularization terms. [16] 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 lie 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 $f(\cdot)$ is Lipschitz continuous in the amended feasible region, see [13]. Indeed, one can show using quadratic optimization optimality conditions that it suffices to set $R = \min\{\frac{1}{\lambda}\left(\frac{1}{n}\sum_{i=1}^n \|w_i\|_2\right), \sqrt{2/\lambda}\}$ (see the Appendix 6) wherein the modulus of Lipschitz continuity in the amended feasible region is at most $M \leq \frac{1}{n}\sum_{i=1}^n \|w_i\|_2 + \min\{\sqrt{2\lambda}, \frac{1}{n}\sum_{i=1}^n \|w_i\|_2\}$. Furthermore, in [7] the authors show that if the initial point lies within a suitably chosen large ball, then Stochastic Subgradient Descent with a small step size will ensure in expectation that all iterates lie in the large ball, which then ensures that the norms of all subgradients are bounded in expectation. Using the constructs that we will

develop herein, we will show (Proposition 5.4) that a suitably designed version of Stochastic Mirror Descent – without any projection step to any large ball – will achieve ε-optimality within

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

iterations of Stochastic Mirror Descent, where x^* is an optimal solution of (3) and x^0 is the initial point.

We accomplish the above computational bounds as an application of a more general theory and algorithmic constructs developed herein to overcome the drawbacks in the standard analysis 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" $h(\cdot)$ and which does not require the specification of any particular norm – and indeed $h(\cdot)$ need not be strongly (or even strictly) convex. Armed with "relative continuity", we demonstrate the capability to solve a more general class of non-differentiable convex optimization problems (without uniform Lipschitz continuity) in both deterministic and stochastic settings.

This paper is a companion for non-differentiable convex optimization to our predecessor paper [8] for differentiable convex optimization. In [8], 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 demonstrated the capability to solve a more general class of differentiable convex optimization problems (without uniform Lipschitz continuous gradients), and we also demonstrated linear convergence results for a Primal Gradient Scheme when the objective function $f(\cdot)$ is both smooth and strongly convex. The current paper is focused on the case of non-differentiable optimization, and uses some different and some similar ideas as compared to [8].

There are some concurrent work on smooth optimization sharing similar spirit to [8]. Bauschke, Bolte, and Teboulle [1] presents a similar definition of relative smoothness as in [8] 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. [17] discusses a unified proof of Mirror Descent and the Proximal Point Algorithm under a similar assumption of relative smoothness. Nguyen [15] develops similar ideas on analyzing Mirror Descent in a Banach space. A more detailed discussion comparing these related works is also presented in [8].

The structure of the current paper is as following: in Section 2 we review the traditional set-up 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 work out for a large class of non-differentiable 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 Intersection of Ellipsoids Problem (IEP) and also the Support Vector Machine (SVM) problem. In Section 6 we present numerical experiments that suggest the practical validity of our approach.

Notation. $\|\cdot\|$ denotes a given norm on the primal space $\mathbb{F}$ 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 $B_2(c,r) := \{x \in \mathbb{F} : \|x - c\|_2 \leq r\}$. $\|A\|_2$ 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

Our optimization problem of interest is:

$$\begin{array}{rrll} P: & f^* := & \text{minimum}_x & f(x) \\ & & \text{s.t.} & x \in Q \ , \end{array} \tag{4}$$

where $Q \subseteq \mathbb{F}$ is a closed convex set in the finite-dimensional vector space $\mathbb{F}$ and $f(\cdot) : Q \to \mathbb{R}$ is a convex function that is not necessarily differentiable. There are by now very many deterministic and stochastic first-order methods for tackling (4), see for example [4], [11], [9] and the references therein; virtually all such methods are designed to solve (4) when the objective function $f(\cdot)$ satisfies a uniform Lipschitz continuity condition on Q, which in the deterministic setting is equivalent to the condition that there exists a constant $M_f < \infty$ for which:

$$\|g(x)\|_* \leq M_f \quad \text{for all } x, y \in Q \text{ and } g(x) \in \partial f(x) \ , \tag{5}$$

where $\partial f(x)$ is the subdifferential of $f(\cdot)$ at x (i.e., the collection of subgradients of $f(\cdot)$ at x), $\|\cdot\|$ is a given norm on $\mathbb{F}$, and $\|\cdot\|_*$ denotes the usual dual norm. Note that we use the notation "$g(x)$" to denote an assignment of a subgradient (or an oracle call thereof) at x, and so $g(\cdot)$ is not a function nor is it a point-to-set map. Also recall the definition of strong convexity: $f(\cdot)$ is (uniformly) μ_f-strongly convex for some $\mu_f > 0$ if

$$f(y) \geq f(x) + \langle g(x), y - x \rangle + \tfrac{\mu_f}{2}\|y - x\|^2 \quad \text{for all } x, y \in Q \text{ and } g(x) \in \partial f(x) \ . \tag{6}$$

2.1 Deterministic Setting

Let us now recall the Mirror Decent Algorithm (see [10] and [2]), which is also referred to as the prox subgradient method in its interpretation in the space of primal variables. Mirror Descent employs a differentiable convex "prox function" $h(\cdot)$ to define a Bregman distance:

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

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 set-up requires that $h(\cdot)$ is 1-strongly convex

with respect to the given norm $\|\cdot\|$, and in this set-up one can prove that after k iterations it holds for any $x \in Q$ that:

$$\min_{0\leq i\leq k} f(x^i) - f(x) \quad \leq \quad \frac{\frac{1}{2}M_f^2 \sum_{i=0}^k t_i^2 + D_h(x, x^0)}{\sum_{i=0}^k t_i} \ , \tag{8}$$

which leads to an $O(1/\sqrt{k})$ sublinear rate of convergence using an appropriately chosen step-size sequence $\{t_i\}$, see [2].

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

Initialize. Initialize with $x^0 \in Q$. Let $h(\cdot)$ 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:

$$x^{i+1} \leftarrow \arg\min_{x\in Q}\{f(x^i) + \langle g(x^i), x - x^i\rangle + \tfrac{1}{t_i}D_h(x, x^i)\} \ . \tag{9}$$

Notice in (9) by construction that the update requires the capability to solve instances of a "linearized subproblem" (which we denote by "LS") of the general form:

$$\text{LS}: \qquad x_{\text{new}} \leftarrow \arg\min_{x\in Q}\{\langle c, x\rangle + h(x)\} \ , \tag{10}$$

for suitable iteration-specific values of c; indeed, (9) is an instance of the subproblem (10) 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 (9) is somewhat meaningless absent the capability to efficiently solve (10), a point which we will return to later on. In the usual design and implementation of Mirror Descent for solving (4), one attempts to specify the norm $\|\cdot\|$ and the 1-strongly convex prox function $h(\cdot)$ in consideration of the properties of the feasible domain Q while also ensuring that the LS subproblem (10) is efficiently solvable.

Also notice in particular that the Mirror Descent Algorithm (Algorithm 1) itself does not require the traditional set-up that $h(\cdot)$ be 1-strongly convex for some particular norm; rather this requirement is part of the traditional analysis. As we will see in this paper, we can instead analyze Mirror Descent by evaluating the intrinsic ways that $f(\cdot)$ and $D_h(\cdot,\cdot)$ are related functionally, and in a way 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 [8].

2.2 Stochastic Setting

For some convex functions, computing an exact subgradient at $x \in Q$ may be very expensive or even intractable, but sampling a random stochastic estimate of a subgradient at x, which we denote by $\tilde{g}(x)$, may be very easy. We say that $\tilde{g}(x)$ is an unbiased stochastic subgradient if $\mathbb{E}\left[\tilde{g}(x)|x\right] \in \partial f(x)$. The usefulness of a stochastic subgradient methodology is easily seen in the context of machine and

statistical learning problems. A prototypical such learning problem is to compute an approximate solution of the following empirical loss minimization problem:

$$\begin{array}{rll} f^* := & \text{minimum}_x & f(x) := \frac{1}{n}\sum_{j=1}^n f_j(x) \\ & \text{s.t.} & x \in Q \ , \end{array} \tag{11}$$

where $f_j(\cdot)$ is a non-differentiable 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 in order to compute a subgradient of $f(\cdot)$, which can be prohibitively expensive. A typical alternative is to compute a stochastic subgradient. Let x^i denote the i-th 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 an unbiased conditional random variable given x^i, namely $\mathbb{E}[\tilde{g}(x^i)|x^i] \in \partial f(x^i)$. A stochastic version of Mirror Descent is presented here 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 $h(\cdot)$ 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:

$$\mathbb{E}[\|\tilde{g}(x)\|_*^2 | x] \leq G_f^2, \text{ for any } x \in Q \ . \tag{12}$$

For notational convenience, we will say that $f(\cdot)$ is G_f-stochastically continuous if (12) holds. In [9], Nedić and Lee developed convergence results for Stochastic Mirror Descent (Algorithm 2). Under the conditions that (i) $f(\cdot)$ is G_f-stochastically continuous, (ii) $h(\cdot)$ is a differentiable and μ_h-strongly convex function on Q, and (iii) Q is a closed bounded set, Nedić and Lee ([9] equation (27)) show the following convergence result using step-sizes $t_i = \sqrt{\frac{\mu_h D_{\max}}{G_f(i+1)}}$:

$$\mathbb{E}\left[f(\bar{x}^k)\right] - f* \ \leq \ \frac{3G_f\sqrt{D_{\max}}}{2\sqrt{\mu_h(k+1)}} \ , \tag{13}$$

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 (a) $f(\cdot)$ is μ_f-strongly convex, and (b) $h(\cdot)$ is L_h-smooth, Nedić and Lee ([9] Theorem 1) show that with step-sizes $t_i = \frac{2L_h}{\mu_f(i+1)}$ it holds that:

$$\mathbb{E}\left[f(\check{x}^k)\right] - f^* \ \leq \ \frac{2G_f^2 L_h}{\mu_f(k+1)\mu_h} \ , \tag{14}$$

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 $f(\cdot)$ – actually two different definitions – one for the deterministic and another for the stochastic setting, respectively. The starting point is a "reference function" $h(\cdot)$ which is a given differentiable convex function on Q that is used to construct the usual Bregman distance $D_h(\cdot,\cdot)$ (7), and that is used as part of the Mirror Descent update (9). However, we point out for emphasis that unlike the traditional set-up there are no presumptions/pre-conditions on $h(\cdot)$ such strong (or even strict) convexity.

3.1 Deterministic Setting

Consider the objective function $f(\cdot)$ of (4). We define "relative continuity" of $f(\cdot)$ relative to the reference function $h(\cdot)$ using the Bregman distance $D_h(\cdot,\cdot)$ of $h(\cdot)$ as follows.

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

$$\|g(x)\|_* \leq \frac{M\sqrt{2D_h(y,x)}}{\|y-x\|} \ . \tag{15}$$

(In the particular case when $h(x) = \frac{1}{2}\|x\|_2^2$, then note that the Bregman distance is $D_h(y,x) = \frac{1}{2}\|y-x\|_2^2$, and the relative continuity condition (15) becomes $\|g(x)\|_2 \leq M$, which in this case corresponds to the standard definition of Lipschitz continuity (5) for the ℓ_2-norm.)

We can rewrite (15) as

$$\|g(x)\|_*^2 \leq M^2 \frac{D_h(y,x)}{\frac{1}{2}\|y-x\|^2} \ , \tag{16}$$

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

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 $f(\cdot)$ is M-relative continuous with respect to the reference function $h(\cdot)$, then for any $t > 0$ it holds for all $x, y \in Q$ and $g(x) \in \partial f(x)$ that:*

$$\tfrac{1}{t} D_h(y,x) + \langle g(x), y-x \rangle + \tfrac{1}{2} t M^2 \geq 0 \ . \tag{17}$$

Proof: If $f(\cdot)$ is M-relative continuous with respect to $h(\cdot)$, then for any $t > 0$ it follows that

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

and the proof follows by rearranging terms. □

The "key property" (17) is what is used in the proofs of results herein, so we could define M-relative continuity using (17) instead of (15). Furthermore, (17) is independent of any norm structure, and so is attractive for this generality. However, we use the definition (15) because it leads to easy verification of M-relative continuity in practical instances as will be 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 $f(\cdot)$ is M-relative continuous with respect to $h(\cdot)$, then for any $\alpha > 0$, $f(\cdot)$ is $\frac{M}{\alpha}$-relative continuous with respect to $\alpha^2 h(\cdot)$.*

2. *If $f(\cdot)$ is M-relative continuous with respect to $h(\cdot)$, then for any $\alpha > 0$, $\alpha f(\cdot)$ is M-relative continuous with respect to $\alpha^2 h(\cdot)$.*

3. *If $f_j(\cdot)$ is M-relative continuous with respect to $h_j(\cdot)$ for $j = 1, \ldots, n$, then $\sum_{j=1}^n f_j(\cdot)$ is $\sqrt{n}M$-relative continuous with respect to $\sum_{j=1}^n h_j(\cdot)$.*

4. *If $f_j(\cdot)$ is M_j-relative continuous with respect to $h_j(\cdot)$ for $j = 1, \ldots, n$, then for $\alpha_j > 0$ and $M > 0$ it holds that $\sum_{j=1}^n \alpha_j f_j(\cdot)$ is $\sqrt{n}M$-relative continuous with respect to $\sum_{j=1}^n \frac{\alpha_j^2}{\beta_j^2} h_j(\cdot)$ 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

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

which establishes the proof.

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

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

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

4. It follows from part (2.) that $\alpha_j f_j(\cdot)$ is M_j-continuous relative to $\alpha_j^2 h_j(\cdot)$, thus $\alpha_j f_j(\cdot)$ is also M-continuous relative to $\frac{\alpha_j^2}{\beta_j^2} h_j(\cdot)$ from part (1.), whereby the proof is furnished directly by utilizing part (3.). $\square$

We also will make use of the notion of "relative strong convexity" which was introduced in [8], and will be used here in some of the convergence guarantee analyses.

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

$$f(y) \geq f(x) + \langle g(x), y - x \rangle + \mu D_h(y, x) \ . \tag{18}$$

3.2 Stochastic Setting

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

Definition 3.3. *$f(\cdot)$ is G-stochastically-relative continuous with respect to the reference function $h(\cdot)$ on Q for some $G > 0$ if $f(\cdot)$ together with the oracle to compute a stochastic subgradient satisfies:*

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

(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 $\mathbb{E}[\|\tilde{g}(x)\|_2^2 | x] \leq G^2$ for all $x \in Q$, which corresponds to the standard condition (12) for the ℓ_2-norm.)

For $x \in Q$, define

$$\tilde{M}(x) := \|\tilde{g}(x)\|_* \max_{y \in Q, y \neq x} \frac{\|y - x\|}{\sqrt{2D_h(y, x)}} \ . \tag{19}$$

Notice for a given x that $\max_{y \in Q, y \neq x} \frac{\|y-x\|}{\sqrt{2D_h(y,x)}}$ 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 $\tilde{g}(x)$. Meanwhile, if $f(\cdot)$ is G-stochastically-relative continuous, we have by the boundedness property that for any $x \in Q$

$$\mathbb{E}[\tilde{M}(x)^2 | x] = \mathbb{E}[\|\tilde{g}(x)\|_*^2 | x] \max_{y \in Q, y \neq x} \frac{\|y - x\|^2}{2D_h(y, x)} \leq G^2 \ . \tag{20}$$

Exactly as in the deterministic setting, we have:

Proposition 3.3. *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 furnished 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 $f(\cdot)$ that is G-stochastically-relative continuous with respect to a given reference function $h(\cdot)$. We also present computational guarantees for (deterministic) Mirror Descent (Algorithm 1) when $f(\cdot)$ is M-relative continuous with respect to the reference function $h(\cdot)$, which will follow 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 [14])) *Let $\phi(x)$ be a convex function, and let $D_h(\cdot,\cdot)$ be the Bregman distance for $h(\cdot)$. For a given vector z, let*

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

Then

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

Let us denote the (primitive) random variable at the i-th 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 which 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). If $f(\cdot)$ is G-stochastically-relative continuous with respect to $h(\cdot)$ for some $G > 0$, then the following inequality holds for all $k \geq 1$ and $x \in Q$:*

$$\mathbb{E}_{\xi_{k-1}}\left[f(\bar{x}^k)\right] - f(x) \quad \leq \quad \frac{\frac{1}{2}G^2 \sum_{i=0}^k t_i^2 + D_h(x,x^0)}{\sum_{i=0}^k t_i} \ , \tag{21}$$

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

Proof: First notice that

$$\begin{aligned} f(x^i) + \left\langle g(x^i), x - x^i \right\rangle &= f(x^i) + \left\langle \mathbb{E}_{\gamma_i}[\tilde{g}(x^i)|x^i], x - x^i \right\rangle \\ &= f(x^i) + \mathbb{E}_{\gamma_i}\left[\left\langle \tilde{g}(x^i), x - x^i \right\rangle \middle| x^i\right] \\ &\geq f(x^i) + \mathbb{E}_{\gamma_i}\left[\left\langle \tilde{g}(x^i), x^{i+1} - x^i \right\rangle + \tfrac{1}{t_i} D_h(x^{i+1}, x^i) + \tfrac{1}{t_i} D_h(x, x^{i+1}) - \tfrac{1}{t_i} D_h(x, x^i) | x^i\right] \\ &\geq f(x^i) + \mathbb{E}_{\gamma_i}\left[-\tfrac{1}{2}\tilde{M}(x^i)^2 t_i + \tfrac{1}{t_i} D_h(x, x^{i+1}) - \tfrac{1}{t_i} D_h(x, x^i) | x^i\right] \\ &\geq f(x^i) - \tfrac{1}{2} G^2 t_i + \tfrac{1}{t_i}\mathbb{E}_{\gamma_i}[D_h(x, x^{i+1})|x^i] - \tfrac{1}{t_i} D_h(x, x^i) \ , \end{aligned} \tag{22}$$

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 (20). Since also $f(x) \geq f(x^i) + \left\langle g(x^i), x - x^i \right\rangle$ from the definition of a subgradient, we have from (22):

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

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

$$f(x) \geq \mathbb{E}_{\xi_{i-1}}[f(x^i)] - \tfrac{1}{2} G^2 t_i + \tfrac{1}{t_i}\mathbb{E}_{\xi_i}[D_h(x, x^{i+1})] - \tfrac{1}{t_i}\mathbb{E}_{\xi_{i-1}}[D_h(x, x^i)] \tag{23}$$

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:

$$\begin{aligned} \tfrac{1}{2} G^2 \textstyle\sum_{i=0}^k t_i^2 + D_h(x, x^0) \geq \sum_{i=0}^k t_i \mathbb{E}_{\xi_{i-1}}[f(x^i) - f(x)] &= \mathbb{E}_{\xi_{k-1}} \textstyle\sum_{i=0}^k t_i [f(x^i) - f(x)] \\ &\geq \left(\textstyle\sum_{i=0}^k t_i\right) \mathbb{E}_{\xi_{k-1}}\left[f(\bar{x}^k) - f(x)\right] \ , \end{aligned} \tag{24}$$

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

Remark 4.1. *By slightly modifying the logic in* (24)*, we can obtain the following result which is similar to the deterministic setting* (8)*:*

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

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

Corollary 4.1. *Let x^* be an optimal solution of* (4)*. Under the conditions of Theorem 4.1, for any $\delta > 0$ it holds that:*

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

Proof: Using the Markov inequality, we have:

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

□

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 (4).

Proof: Substituting the values of t_i in (21) yields the result directly. □

Remark 4.2. *Let us now compare these results to related results of Nedić and Lee [9]. In order to attain an ε-optimality gap, [9] proved a bound of $\left\lceil \frac{9G_f^2 D_{\max}}{4\mu_h \varepsilon^2} \right\rceil$ iterations (this follows by rearranging (13)). In addition to not requiring Lischitz continuity of $f(\cdot)$, our bound does not require that $h(\cdot)$ 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 (13), it holds that:*

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

where the first inequality utilize the strong convexity (in the standard sense) of $h(\cdot)$, and the second inequality is due to the assumption that $f(\cdot)$ is G_f-stochasticlly continuous (in the standard sense). Thus we see that the bound in Corollary 4.2 improves on the bound in [9].

In the case when $f(\cdot)$ is also μ_f-strongly convex relative to $h(\cdot)$ (see Definition 3.2), we can 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 $h(\cdot)$) *Consider the Stochastic Mirror Descent Algorithm (Algorithm 2). If $f(\cdot)$ is G-stochastically-relative continuous with respect to $h(\cdot)$ for some $G > 0$ and $f(\cdot)$ is μ-strongly convex relative to $h(\cdot)$ 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 $k \geq 1$:*

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

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 relative convexity (18) that

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

Combining the above inequality with (22) 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:

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

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

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

where the first inequality uses the convexity of $f(\cdot)$ 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 (4), the proof is completed by noticing $\sum_{i=1}^{k} i = \frac{k(k+1)}{2}$. □

Remark 4.3. *Let us also compare the computational guarantee of Theorem 4.2 to the results in Nedić and Lee [9]. In order to attain an ε-optimality gap, [9] proved the bound* (14)*. First notice that we do not require either that $f(\cdot)$ is uniformly Lipschitz continuous nor that $h(\cdot)$ is strongly convex or smooth. However, even if these requirements hold, it follows from Remark 4.2 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 $f(\cdot)$ is μ_f strongly convex and the second inequality $h(\cdot)$ is L_h smooth in the standard sense. Thus $f(\cdot)$ is at least $\mu = \frac{\mu_f}{L_h}$-strongly convex relative to $h(\cdot)$ by using Proposition 1.1 in [8]. Therefore, even under the stronger requirements of [9], Theorem 4.2 improves on the corresponding result in [9].

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 $f(\cdot)$ is M-relative continuous with respect to $h(\cdot)$ 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 x^* is any optimal solution of (4).

□

Theorem 4.4. (Convergence Bounds for Deterministic Mirror Descent with Strong Relative Convexity) *Consider the Deterministic Mirror Descent Algorithm (Algorithm 1). If $f(\cdot)$ is M-relative continuous with respect to $h(\cdot)$ for some $M > 0$ and $f(\cdot)$ is μ-strongly convex relative to $h(\cdot)$ 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 $h(\cdot)$ 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 (4) with objective function $f(\cdot)$ that is M-relative continuous or G-stochastically-relative continuous (respectively) with respect to the reference function $h(\cdot)$. In order to efficiently execute the update step in Algorithm 1 and/or Algorithm 2 we need $h(\cdot)$ to be such that the linearization subproblem LS (10) is efficiently solvable for any given c. Therefore, in order execute Mirror Descent for solving (4) using Algorithm 1 or Algorithm 2, we need to specify a differentiable convex reference function $h(\cdot)$ that has the following two properties:

(i) $f(\cdot)$ is M-relative continuous (or G-stochastically-relative continuous) with respect to $h(\cdot)$ on Q for M (or G) that is easy to determine, and

(ii) the linearization subproblem LS (10) has a solution, and the solution is efficiently computable.

We now discuss quite broadly how to construct such a reference function $h(\cdot)$ 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 whose coefficients $\{a_i\}$ are nonnegative. Let

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

Then the following proposition states that $f(\cdot)$ is 1-relative continuous with respect to $h(\cdot)$. This implies that no matter how fast the subgradient of $f(\cdot)$ grows polynomially as $\|x\|_2 \to \infty$, $f(\cdot)$ can still be relatively continuous with respect to the simple reference function $h(\cdot)$, even though $f(\cdot)$ does not exhibit uniform Lipschitz continuity.

Proposition 5.1. *$f(\cdot)$ is 1-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

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

Notice that

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

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

$$D_h(y,x) = \sum_{i=0}^{r} a_i D_{h_i}(y,x) \ge \tfrac{1}{2}\|y-x\|_2^2 \left(\sum_{i=0}^{r} a_i \|x\|_2^i \right) \ . \tag{25}$$

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 $f(\cdot)$ is 1-relative continuous with respect to $h(\cdot)$. $\square$

Solving the linearization subproblem (10)**.** Let us see how we can solve the linearization subproblem (10) for this class of optimization problems. The linearization subproblem (10) can be written as

$$\text{LS}: \quad \min_{x \in \mathbb{R}^n} \ \langle c, x \rangle + \sum_{i=0}^{r} \tfrac{a_i}{i+2} \|x\|_2^{i+2} \ , \tag{26}$$

and the first-order optimality condition is simply:

$$c + \left(\sum_{i=0}^{r} a_i \|x\|_2^i \right) x = 0 \ , \tag{27}$$

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 $x = 0$ satisfies (27), so let us examine the case when $c \neq 0$, in which case from (27) 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 monotone univariate polynomial 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* (26) *provided that we can easily compute the Euclidean projection on Q. In this case, the linearization subproblem* (10) *can be converted to a* 1*-dimensional convex optimization problem, see Appendix A.1 of [8] 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 $\mathbb{E}\left[\|\tilde{g}(x)\|_*^2 | x\right] \leq p_r(\|x\|_2)$ for all $x \in Q$, where $p_r(\alpha) = \sum_{i=0}^{r} a_i \alpha^i$ is an r-degree polynomial whose coefficients $\{a_i\}$ 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. *$f(\cdot)$ is* 1*-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:

$$\mathbb{E}\left\|\tilde{g}(x)\right\|_*^2 | x\right] \leq p_r(\|x\|_2) = \sum_{i=0}^{r} a_i \|x\|_2^i \leq \frac{2D_h(y,x)}{\|y - x\|_2^2} \ , \tag{28}$$

where the last inequality follows from (25), and thus $f(\cdot)$ is 1-stochastically continuous relative to $h(\cdot)$.

Solving the linearization subproblem (10). The linear optimization subproblem is identical in structure to that in the deterministic case and so can be solved as discussed at the end of Section 5.1.

5.3 Relative Continuity for IEP and SVM

Here we re-examine the two motivating examples stated at the beginning of the paper, namely the Intersection of Ellipsoids Problem (IEP) and the Support Vector Machine model (SVM). We first prove the following lemma, which presents an upper bound on the Bregman distance $D_h(\cdot,\cdot)$ 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* $D_h(y,x) \leq \frac{1}{3}\|y-x\|_2^2\,(\|y\|_2 + 2\|x\|_2)$.
2. *Let* $h(x) := \frac{1}{4}\|x\|_2^4$. *Then* $D_h(y,x) \leq \frac{1}{4}\|y-x\|_2^2\left(\|y+x\|_2^2 + 2\|x\|_2^2\right)$.

Proof:

1.

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

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.

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

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. □

Intersection of Ellipsoids Problem (IEP). Recall from the Introduction that we can write the IEP problem as:

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

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 $f_i(\cdot)$ is the i-th term in the objective function of (29). Since $g(x) \in \partial f(x)$ 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 [3]), 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 $f(\cdot)$ is 1-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 5.3. (Computational Guarantees for Deterministic Mirror Descent for the IEP problem (2)). *Consider applying the Deterministic Mirror Descent algorithm (Algorithm 1) to the Ellipsoid Intersection Problem (2) using the reference function (30), where $\sigma := \max_{0\le i\le n} \|A_i\|_2^2$ where $\|A_i\|_2$ 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 (2). 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 $f(\cdot)$ is 1-continuous relative to $h(x) = \frac{\sigma}{4}\|x\|_2^4 + \frac{\rho}{3}\|x\|_2^3 + \frac{\gamma}{2}\|x\|_2^2$. Furthermore, it follows from applying Lemma 5.1 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 furnished by substituting these values into the computational guarantee of Corollary 4.3. □

Support Vector Machine (SVM). Recall the support vector machine problem:

$$\text{SVM:} \quad f^* \ = \ \min_x f(x) := \tfrac{1}{n}\sum_{i=1}^n \max\{0, 1 - y_i x^T w_i\} + \tfrac{\lambda}{2}\|x\|_2^2 \ . \tag{31}$$

We can rewrite the objective function of (31) as

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

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 $f(\cdot)$ by using a single sample index $\tilde{j}$ drawn randomly from $\{1,\ldots,n\}$, namely $\tilde{g}(x) \in$

$\partial f_{\tilde{j}}(x)$ where $\tilde{j}$ is drawn uniformly at random from $\{1,\ldots,n\}$. Then $\|\tilde{g}(x)\|_2^2 \leq (\lambda\|x\|_2 + \|w_{\tilde{j}}\|_2)^2$, whereby

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

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

$$h(x) := \frac{\lambda^2}{4}\|x\|_2^4 + \frac{2\lambda}{3n}\left(\sum_{i=1}^n \|w_i\|_2\right)\|x\|_2^3 + \frac{1}{2n}\left(\sum_{i=1}^n \|w_i\|_2^2\right)\|x\|_2^2 \ , \tag{32}$$

it follows from the Proposition 5.2 that $f(\cdot)$ is 1-stochastically continuous relative to $h(x)$.

Proposition 5.4. (Computational Guarantees for Stochastic Mirror Descent for the SVM problem (3)**.)** *Consider applying the Stochastic Mirror Descent algorithm (Algorithm 2) to the Support Vector Machine problem* (3) *using the reference function* (32)*. 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 (3)*. Then it holds that*

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

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

Proof: We showed above (using Proposition 5.2) that $f(\cdot)$ is 1-stochastically continuous relative to $h(\cdot)$ defined in (32). Furthermore, it follows from applying Lemma 5.1 that

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

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

6 Numerical Experiment

In this section we present the results of a numerical experiment where we solve a large-scale instance of the Support Vector Machine problem and we compare the performance of (i) Algorithm 2 using the reference function (32) with (ii) the traditional stochastic subgradient descent method (SGD). The dataset we use is the training set of KDD04 Physics dataset [5], which contains $n = 50,000$ sample observations and 65 features (after removing redundant features and features with missing data). We set the regularization parameter to $\lambda = 0.0001$, and implemented both methods using the initial point $x^0 := 0$. For each method we analyzed three different step-size sequences, namely $\{\frac{c}{k}\}_k$, $\{\frac{c}{\sqrt{k}}\}_k$, and a constant step-size sequence $\{c\}_k$. In all cases the constant c was heuristically

optimized to yield the lowest average optimality gap with $500,000$ iterations. We performed 10 independent tests, and computed the average of optimality gaps and over the tests, for 10 epochs ($500,000$ iterations). Figure 1 presents a log-log plot of the average optimality gaps versus the number of iterations. We see from Figure 1 that Algorithm 2 and SGD exhibit very similar behavior. This is to be expected in this case since the regularization parameter λ is very small, and as a result the reference function (32) is nearly identical to a constant times $\frac{1}{2}\|x\|_2^2$. With the step-size sequence $\{\frac{c}{k}\}_k$, both methods exhibit an $O(1/k)$ convergence rate, this being the case since the objective function is locally strongly convex. With the step-size sequences $\{\frac{c}{\sqrt{k}}\}_k$ and $\{c\}_k$, both methods exhibit $O(1/\sqrt{k})$ convergence rates, which is consistent with the theoretical guarantees. Notice that Algorithm 2 delivers a slightly smaller optimality gap than SGD after $500,000$ iterations for each of the three step-size sequences.

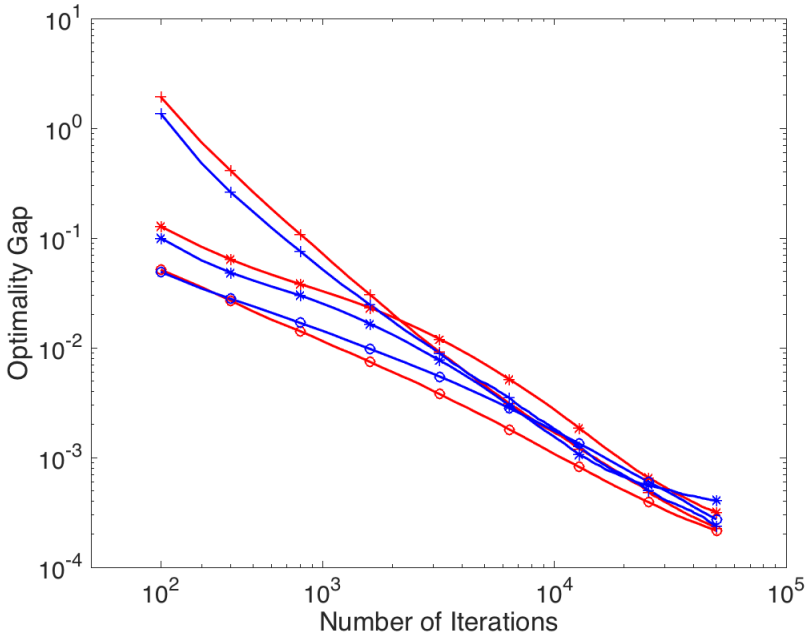

Figure 1: Comparison of Algorithm 2 (red lines) and traditional Stochastic Gradient Descent (blue lines) for solving the SVM problem with the KDD04 Physics dataset [5]. The lines labelled "$+$" use a step-size $\frac{c}{k}$, the lines labelled "$*$" use a step-size $\frac{c}{\sqrt{k}}$ and the lines labelled "$\circ$" use a constant step-size c. In each of the six cases the constant c is heuristically optimized to yield the lowest average optimality gap with $500,000$ iterations.

Acknowledgement

The author would like to express his gratitude to Robert M. Freund for numerous thoughtful discussions helping motivate the work, for reading the draft carefully, and for advising on the presentation and positioning of this paper. The author also wish to thank Yurii Nesterov for encouraging the author's work on this topic, and for pointing out the application of IEP.

Appendix: Finite Radius Bound for SVM

Here we derive an upper bound on the norm of an optimal solution of the SVM problem (3).

Proposition 6.1. *The optimal solution to the SVM problem* (3) *lies in the ball* $B_2(0, R)$ *for* $R = \min\left\{\frac{1}{n\lambda}\sum_{i=1}^n \|w_i\|_2, \sqrt{2/\lambda}\right\}$.

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

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

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

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

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

where $\pi^* \geq 0$ and $\beta^* \geq 0$. Define $\bar{\pi}^* = n\pi^*$. Then $0 \leq \bar{\pi}^* \leq e$ and

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

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

References

[1] H.H. Bauschke, J. Bolte, and M. Teboulle, *A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications*, Mathematics of Operations Research **42** (2016), no. 2, 330–348.

[2] A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters **31** (2003), no. 3, 167–175.

[3] D. Bertsekas, *Nonlinear programming*, Athena Scientific, Belmont, MA, 1999.

[4] S. Bubeck, *Convex optimization: Algorithms and complexity*, Foundations and Trends® in Machine Learning **8** (2015), no. 3-4, 231–357.

[5] R. Caruana, T. Joachims, and L. Backstrom, *KDD-cup 2004: results and analysis*, ACM SIGKDD Explorations Newsletter **6** (2004), no. 2, 95–108.

[6] J. Duchi and Y. Singer, *Efficient online and batch learning using forward backward splitting*, Journal of Machine Learning Research **10** (2009), no. Dec, 2899–2934.

[7] S. Lacoste-Julien, M. Schmidt, and F. Bach, *A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method*, arXiv preprint arXiv:1212.2002 (2012).

[8] H. Lu, R.M. Freund, and Y. Nesterov, *Relatively-smooth convex optimization by first-order methods, and applications*, arXiv preprint arXiv:1610.05708 (2016).

[9] A. Nedić and S. Lee, *On stochastic subgradient mirror-descent algorithm with weighted averaging*, SIAM Journal on Optimization **24** (2014), no. 1, 84–107.

[10] A. S. Nemirovsky and D. B. Yudin, *Problem complexity and method efficiency in optimization*, Wiley, New York, 1983.

[11] Y. Nesterov, *Introductory lectures on convex optimization: a basic course*, Kluwer Academic Publishers, Boston, 2003.

[12] ———, *private communication*, (2016).

[13] S. Shalev-Shwartz, Y. Singer, and N. Srebro, *Pegasos: Primal estimated sub-gradient solver for SVM*, Proceedings of the 24th International Conference on Machine learning, ACM, 2007, pp. 807–814.

[14] P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization*, Tech. report, May 21, 2008.

[15] Q. Van Nguyen, *Forward-backward splitting with Bregman distances*, Vietnam Journal of Mathematics **45** (2017), no. 3, 519–539.

[16] J. Yu, S.V.N. Vishwanathan, S. Günter, and N.N. Schraudolph, *A quasi-Newton approach to nonsmooth convex optimization problems in machine learning*, Journal of Machine Learning Research **11** (2010), no. Mar, 1145–1200.

[17] Y. Zhou, Y. Liang, and L. Shen, *A unified approach to proximal algorithms using Bregman distance*, Tech. report, 2016.