Stochastic incremental mirror descent algorithms with Nesterov smoothing

Sandy Bitterlich, Sorin-Mihai Grad

Stochastic incremental mirror descent algorithms with Nesterov smoothing

Sandy Bitterlich * Sorin-Mihai Grad †

January 16, 2023

Abstract

For minimizing a sum of finitely many proper, convex and lower semicontinuous functions over a nonempty closed convex set in an Euclidean space we propose a stochastic incremental mirror descent algorithm constructed by means of the Nesterov smoothing. Further we modify the algorithm in order to minimize over a nonempty closed convex set in an Euclidean space a sum of finitely many proper, convex and lower semicontinuous functions composed with linear operators. Next a stochastic incremental mirror descent Bregman-proximal scheme with Nesterov smoothing is proposed in order to minimize over a nonempty closed convex set in an Euclidean space a sum of finitely many proper, convex and lower semicontinuous functions and a prox-friendly proper, convex and lower semicontinuous function. Different to the previous contributions from the literature on mirror descent methods for minimizing sums of functions, we do not require these to be (Lipschitz) continuous or differentiable. Applications in Logistics, Tomography and Machine Learning modelled as optimization problems illustrate the theoretical achievements.

Keywords. Mirror descent method, stochastic algorithm, Nesterov smoothing, incremental algorithm, proximal point algorithm, PET image reconstructions

1 Introduction

The original mirror descent method was introduced by Nemirovski in [38] (see also [39]) as a noneuclidean extension of the subgradient method for solving unconstrained convex optimization problems and since then it has been subject to various developments and employment in different areas (such as game theory [49], inverse problems [3], finance mathematics [28], machine learning [22, 29, 49], reinforcement learning [31] transport research [49], signal and image processing [3, 9, 23], compressed sensing [4], labeling and classification problems [29, 32, 37], location research [50], network optimization [24], system identification [15], optimal control [36], ranking problems [21], basis pursuit [23], metric learning [27], generative adversarial networks (GANs) [33], computer vision [30]), enjoying further an increasing popularity (proven, for instance, by the about sixty papers on this topic uploaded last year only on the preprint service arXiv). During these four decades it was noticed that it is connected to other iterative methods for solving various classes of optimization problems such as FTRL (follow the regularized leader) [32],

*Chemnitz University of Technology, Faculty of Mathematics, 09126 Chemnitz, Germany, e-mail: sandy.bitterlich@mathematik.tu-chemnitz.de.

†UMA, ENSTA Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France, e-mail: sorin-mihai.grad@ensta-paris.fr, ORCID-ID: 0000-0002-1139-7504. Corresponding author.

Thomson sampling and Information Radio [55], proximal gradient [53], Sinkhorn's algorithm [35], conditional gradient [5], AdaBoost [16] or dual averaging (also called *lazy mirror descent*) [25], being seen as a generalization of the proximal point algorithm with a nonlinear distance function (that could be a Bregman type one or, for instance, the Fenchel coupling [54]) and an optimal stepsize (see [30]) and as a dual approach to gradient descent (see [2]). Due to their convergence properties, mirror descent algorithms proved to be especially suitable for large-scale optimization problems. In [20] two main streams of current work on mirror descent methods are identified, namely accelerating deterministic mirror descent (see, for instance, [2,21,23,28]) and stochastic mirror descent with access to noised gradient oracle (like in [3, 4, 15, 19, 20, 29, 34, 37, 50, 54]). A further proof of the lively and continuous interest generated by the mirror descent type algorithms in the community is the fact that many of these articles were written during the last three years.

Algorithms of mirror descent type have been employed for solving not only unconstrained minimization problems (like in most of the cited references), but also constrained optimization problems [24,48,50], bilevel optimization problems [3,19], matrix optimization problems [27, 35], variational inequalities [25, 34, 49], online convex optimization problems [27, 32, 50], stochastic optimization problems [19, 51, 54], saddle point problems [33] and even multiobjective optimization problems [48]. Although in most cases the involved functions are convex and differentiable, there have been extensions of mirror descent towards nonsmooth optimization [14, 16, 21, 22, 37, 50] and even nonconvex optimization [21, 51]. Generalizations of mirror descent methods can be found, for instance, in [2,32], while for continuous versions (by means of dynamical systems) we refer to [34, 36]. The mirror descent type algorithms are usually employed for minimizing a single function, however in works like [5, 9, 10, 14, 15, 20, 21, 23, 31, 50, 51] such methods were used for minimizing sums of (convex) functions by considering splitting techniques, in order to solve problems arising from various applications from fields such as machine learning or imaging. A specific feature of mirror descent type algorithms is that the convergence statements are provided in terms of values of objective functions, however in papers like [14,37,51] the convergence of the generated iterative sequence is investigated, too.

In this paper we propose first a stochastic incremental mirror descent algorithm with Nesterov smoothing for minimizing a sum of finitely many proper, convex and lower semicontinuous functions over a given nonempty closed convex set in an Euclidean space, motivated by applications in fields like machine learning or image processing. Unlike the previous mirror descent methods for minimizing sums of functions, ours does not ask these to be Lipschitz continuous. Different to the few other contributions where mirror descent algorithms were introduced for minimizing functions lacking Lipschitz continuity like [29, 50], where a generalization of this property was considered, we employ smooth approximations (via the Nesterov smoothing from [41]) of the involved functions. To the best of our knowledge smoothing methods for the involved functions were considered in connection to mirror descent algorithms only in [22, 23] (see also [16] for objective functions somehow similar to the ones considered in our work) in contexts only vaguely related to our study. Then we show that the algorithm can be modified in order to minimize over a given nonempty closed convex set in an Euclidean space a sum of finitely many proper, convex and lower semicontinuous functions composed with linear operators mapping between two Euclidean spaces. Adding to the sum a further proper, convex and lower semicontinuous function that is prox-friendly requires modifications to the previously mentioned method. The resulting algorithm is a stochastic incremental mirror descent Bregman-proximal scheme with Nesterov smoothing, and this is further modified in order to minimize the sum over a given nonempty closed convex set in an

Euclidean space of finitely many proper, convex and lower semicontinuous functions composed with linear operators, and the mentioned prox-friendly proper, convex and lower semicontinuous function. Different to the previous contributions from the literature on designing mirror descent methods for minimizing sums of functions mentioned above (in particular [10, 14, 20, 21, 23]), the functions we consider need not be (Lipschitz) continuous or differentiable. Moreover, our approach does not require knowledge of the Lipschitz constants or the subgradients of the involved functions, which sometimes can be computationally expensive to determine. We also show that in case some of the involved functions are Lipschitz continuous our methods can be easily combined with the ones proposed in [10]. In [11] one can find a variable smoothing approach to minimize convex optimization problems with stochastic gradients, so that large scale problems can be addressed, where, different to our work, the Moreau-envelope, a special case of Nesterov smoothing, is used. In order to illustrate our theoretical achievements we consider applications in Logistics (Location Optimization), Medical Imaging (Tomography) and Machine Learning (Support Vector Machines) modelled as optimization problems that are iteratively solved via the algorithms we propose in this work.

2 Preliminaries

In this section we give some basic definitions and notations, which we use in this paper.

In the following we assume that $\mathbb{R}^n$ is endowed with the Euclidean inner product $\langle \cdot, \cdot \rangle$ and associated norm $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. The closure of a set $U \subseteq \mathbb{R}^n$ is denoted by $\operatorname{cl} U$. For a convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ the *effective domain* is defined as $\operatorname{dom} f := \{x \in \mathbb{R}^n : f(x) < +\infty\}$ and we say that f is *proper*, if $f > -\infty$ and $\operatorname{dom} f \neq \emptyset$. The *subdifferential* of f at $x \in \mathbb{R}^n$ is given for $f(x) \in \mathbb{R}$ as

$$\partial f(x) := \{v \in \mathbb{R}^n : f(y) \geq f(x) + \langle v, y - x \rangle \forall y \in R^n\}$$

and otherwise as $\partial f(x) := \emptyset$. An element v of the subdifferential $\partial f(x)$ is called subgradient of f at x. For the gradient of a differentiable function f we write $\nabla f(x)$. The function f is said to be *strongly convex* if there exists $\beta \in]0, +\infty[$ such that for all $x, y \in \operatorname{dom} f$ and all $\lambda \in [0, 1]$ one has

$$f(\lambda y + (1 - \lambda)x) \leq \lambda f(y) + (1 - \lambda) f(x) - \lambda(1 - \lambda)\beta \|x - y\|^2/2.$$

Furthermore, we say that f is σ-*cocoercive* (or $(1/\sigma)$-*Lipschitz continuous*) for a $\sigma > 0$ if for every $x, y \in \mathbb{R}^n$ it holds

$$\sigma \|f(x) - f(y)\|^2 \leq \langle x - y, f(x) - f(y) \rangle.$$

The *(Fenchel) conjugate function* $f^* : \mathbb{R}^n \to \overline{\mathbb{R}}$ of a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is defined as

$$f^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle y, x \rangle - f(x)\} \quad (y \in \mathbb{R}^n)$$

and is a proper, convex and lower semicontinuous function. Note that f is proper, convex and lower semicontinuous if and only if $f^{**} = f$, where f^{**} is the conjugate function of f^*. The infimal convolution of two proper functions $f, g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is the function $f \square g : \mathbb{R}^n \to \overline{\mathbb{R}}$, defined by $(f \square g)(x) = \inf_{y \in \mathbb{R}^n} \{f(y) + g(x - y)\}$.

Definition 2.1. The *Moreau-envelope* of a proper, convex and lower semicontinuous function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ of coefficient $\gamma > 0$ is

$$\inf_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{1}{2\gamma} \|y - x\|^2 \right\},$$

and the *proximal point* of coefficient $\gamma > 0$ of f at $x \in \mathbb{R}^n$ is the unique optimal solution of the minimization problem

$$\mathrm{Prox}_{\gamma f}(x) = \arg\min_{y \in \mathbb{R}^n} \left\{ \gamma f(y) + \frac{1}{2}\|y - x\|^2 \right\}.$$

More generally, we call $\mathrm{Prox}_{\gamma f} : \mathbb{R}^n \to \mathbb{R}^n$ the *proximity operator* (or *proximal point mapping*) of f of coefficient γ.

Let $A : \mathbb{R}^n \to \mathbb{R}^p$ be a linear operator. Its *image* is denoted by $\mathrm{Im}\, A = \{Ax : x \in \mathbb{R}^n\}$. The operator $A^* : \mathbb{R}^p \to \mathbb{R}^n$, fulfilling $\langle A^* y, x\rangle = \langle y, Ax\rangle$ for all $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^p$, denotes the *adjoint operator* of A, while $\|A\| := \sup\{\|Ax\| : \|x\| \leq 1\}$ denotes the norm of A.

The mirror descent algorithm on which we build our study was considered in [40] for the problem of minimizing a proper convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ over a nonempty, convex and closed set $C \subseteq \mathbb{R}^n$, by involving a proper, lower semicontinuous and σ-strongly convex function (where $\sigma > 0$) $H : \mathbb{R}^n \to \overline{\mathbb{R}}$ such that $C = \mathrm{cl}(\mathrm{dom}\, H)$ and $\mathrm{Im}\, \nabla H^*$ is a subset of the interior of the domain of f, consisting in the following iterative scheme (where x_0 lies in the interior of the domain of f, $y_0 \in \mathbb{R}^n$ and $t_k > 0$, $k \geq 0$, are positive stepsizes)

$$(\forall k \geq 0) \begin{cases} y_{k+1} = y_k - t_k f'(x_k), \\ x_{k+1} = \nabla H^*(y_{k+1}). \end{cases}$$

As noted in [10], this scheme generalizes the classical subgradient method and is close to the subgradient projection algorithm.

3 A stochastic incremental mirror descent algorithm with Nesterov smoothing

Problem 3.1. *We consider the convex optimization problem*

$$\min_{x \in C} \left\{ \sum_{i=1}^{m} f_i(x) \right\}, \tag{1}$$

where $C \subseteq \mathbb{R}^n$ is a nonempty, convex and closed set and for all $i = 1, \ldots, m$, $(m \in \mathbb{N})$ $f_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ fulfills

$$f_i(x) = \max_{u \in U_i} \{\langle A_i x, u\rangle - \phi_i(u)\},\ x \in \mathrm{dom}\, f_i, \tag{2}$$

where $U_i \subseteq \mathbb{R}^p$ is compact and convex, $A_i : \mathbb{R}^n \to \mathbb{R}^p$ is linear and $\phi_i : \mathbb{R}^p \to \overline{\mathbb{R}}$ a proper, lower semicontinuous and convex function. We assume that $C \cap (\cap_{i=1}^m \mathrm{dom}\, f_i) \neq \emptyset$. Furthermore, let $H : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper, lower semicontinuous and σ-strongly convex function (for $\sigma > 0$) such that $C = \mathrm{cl}(\mathrm{dom}\, H)$ and $\mathrm{Im}\, \nabla H^* \subseteq \cap_{i=1}^m \mathrm{dom}\, f_i$.

Due to the fact that H is a proper, lower semicontinuous and σ-strongly convex function, its conjugate function H^* is Fréchet differentiable and its gradient ∇H^* is σ-cocoercive. In the algorithms we propose in this paper we have the map ∇H^* as mirror map, which is induced by the function H. This map mirrors each iterate onto the feasible set C. So we can choose $H(x) = \frac{1}{2}\|x\|^2$, for $x \in C$ and $H(x) = +\infty$, otherwise, to obtain for the mirror map ∇H^* the orthogonal projection onto C. When $C = \mathbb{R}^n$ the map ∇H^* reduces to the identity operator, however one can choose other mirror maps as well, depending on the structure of C and the considered optimization problem. For further examples of mirror maps for corresponding sets C see the applications in section 6.

Remark 3.2. The construction (2) guarantees that the functions f_i, $i = 1, \dots, m$, are proper, convex and lower semicontinuous. Note that for every proper, lower semicontinuous and convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ one has for all $x \in \mathbb{R}^p$ that $f \circ A(x) = \sup_{u \in \operatorname{dom} f^*} \{\langle Ax, u\rangle - f^*(u)\}$, where $A : \mathbb{R}^p \to \mathbb{R}^n$ is a linear operator. This supremum is a maximum, for instance, when $\operatorname{dom} f^*$ is bounded, which happens when f is Lipschitz continuous (see [46]), while the opposite implication is not known to hold. In this case $f \circ A$ is of the form (2) with $f_i = f \circ A$, $A_i = A$, $U_i = \operatorname{dom} f^*$ and $\phi_i = f^*$. For deeper insights and examples of this construction, we refer the reader to [41,42], while in works like [8,52] it is employed for designing algorithms for solving various classes of optimization problems, some of which stemming from concrete applications.

To minimize the sum of the nonsmooth convex functions f_i $(i = 1, \dots, m)$ in Problem 3.1, at first we approximate them by smooth functions. For this we use the Nesterov smoothing technique (see [41], also employed in works like [45, 52]). One can of course discuss which class of (splitting proximal point type) algorithms delivers the desired results faster and cheaper, however we opt to employ a mirror descent type technique due to the known qualities of these methods. Furthermore, the considered functions are suitable for our approach and also in many applications only certain convergence properties of the values of the objective function, best obtained via mirror descent, are relevant.

Definition 3.3. For $i = 1, \dots, m$, and $\beta > 0$, a continuous and β-strongly convex function $b_{U_i} : \mathbb{R}^p \to \overline{\mathbb{R}}$ is called the *prox-function* of the set $U_i \subseteq \mathbb{R}^p$. Its *prox-center* is denoted $u_i^c = \arg\min_{u \in U_i} b_{U_i}(u)$ and its *prox-diameter* by $D_{U_i} = \sup_{u \in U_i} b_{U_i}(u)$.

Without loss of generality we set in the following $\beta = 1$ and assume that for all $i = 1, \dots, m$, $b_{U_i}(u_i^c) = 0$ and therefore $b_{U_i}(u) \geq 0$ for all $u \in U_i$.

Next we approximate the functions f_i $(i = 1, \dots, m)$ by the smooth functions $f_i^\gamma : \mathbb{R}^n \to \mathbb{R}$

$$f_i^\gamma(x) = \max_{u \in U_i} \{\langle A_i x, u\rangle - \phi_i(u) - \gamma b_{U_i}(u)\}, \tag{3}$$

where $\gamma > 0$ is the *smoothing parameter*. This procedure originates from [41] (see also [42]) and is called *Nesterov smoothing*. We define $u_i^\gamma(x) = \arg\max_{u \in U_i} \{\langle A_i x, u\rangle - \phi_i(u) - \gamma b_{U_i}(u)\}$. Furthermore, it holds

$$f_i^\gamma(x) \leq f_i(x) \leq f_i^\gamma(x) + \gamma D_{U_i} \ \forall x \in \operatorname{dom} f_i. \tag{4}$$

Lemma 3.4. *The functions f_i^γ, $i = 1, \dots, m$, defined as above are well defined, convex, and continuously differentiable at every $x \in U_i$. Furthermore, $\nabla f_i^\gamma = A_i^* u_i^\gamma$ which is $\|A_i\|^2/\gamma$-Lipschitz continuous, and it holds*

$$\|\nabla f_i^\gamma(x)\|^2 \leq 2\|A_i\|^2(2D_{U_i} + \|u_i^c\|^2) \ \forall x \in \mathbb{R}^n.$$

Proof. For the first part of the proof see [41, Theorem 1], where the continuity and finiteness of f_i, $i = 1, \dots, m$, imposed in the hypothesis, were not employed. It remains only the inequality to be shown.
For $i \in \{1, \dots, m\}$, and $x \in \mathbb{R}^n$ it holds

$$\|\nabla f_i^\gamma(x)\|^2 \leq \|A_i\|^2 \|u_i^\gamma(x)\|^2 \leq \|A_i\|^2 \left(2\|u_i^\gamma(x) - u_i^c\|^2 + 2\|u_i^c\|^2\right).$$

Due to the 1-strong convexity of b_{U_i} we have

$$\|u_i^\gamma(x) - u_i^c\|^2 \leq 2b_{U_i}(u_i^\gamma(x)) - 2b_{U_i}(u_i^c) - 2\nabla b_{U_i}(u_i^c)(u_i^\gamma(x) - u_i^c),$$

and, taking into consideration that $b_{U_i}(u_i^c) = 0$ and that $\nabla b_{U_i}(u_i^c) = 0$, it follows from this inequality that

$$\|u_i^\gamma(x) - u_i^c\|^2 \leq 2b_{U_i}(u_i^\gamma(x)) \leq 2D_{U_i}.$$

Hence

$$\|\nabla f_i^{\gamma}(x)\|^2 \leq 2\|A_i\|^2(2D_{U_i} + \|u_i^c\|^2).$$

□

Remark 3.5. Notice that for $i = 1, \ldots, m$, $g_i : \mathbb{R}^p \to \overline{\mathbb{R}}$, $\phi_i = g_i^*$, $b_{U_i} = (1/2)\|\cdot\|^2$ and $U_i = \operatorname{dom} g_i^*$ is compact and convex for a $\gamma > 0$ the function $f_i^{\gamma}(\cdot) = (g_i \square (1/(2\gamma))\|\cdot\|^2)(A_i\cdot)$ is the Moreau-envelope of $g_i \circ A_i$ and $\nabla f_i^{\gamma}(\cdot) = (1/\gamma)\left(\cdot - A_i^* \operatorname{Prox}_{\gamma g_i}(A_i\cdot)\right)$. In this case $\|u_i^c\| = 0$.

Remark 3.6. Other smoothing methods (like the general one presented in [8]) could be employed as well in the framework we consider in this paper as long as they guarantee the last result from Lemma 3.4, namely that the norms of the gradients of the smooth approximations of the considered functions are bounded.

Remark 3.7. An example for an application for an optimization problem in the form of Problem 3.1 is the continuous location problem considered for numerical experiments in section 6.1.

For the convergence analysis of the following algorithms we use two measures of distance in the sense of Bregman.

Definition 3.8. Let $H : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper and convex function. The *Bregman-distance-like function* of H is denoted as

$$d_H : \mathbb{R}^n \times \operatorname{dom} H \times \mathbb{R}^n \to \overline{\mathbb{R}}, \quad d_H(x, y, z) := H(x) - H(y) - \langle z, x - y\rangle.$$

Because of the subgradient inequality it holds that $d_H(x, y, z) \geq 0$ for every $(x, y) \in \mathbb{R}^n \times \operatorname{dom} H$ and all $z \in \partial H(y)$.

Definition 3.9. The *Bregman distance* associated to a proper and convex function $H : \mathbb{R}^n \to \overline{\mathbb{R}}$ fulfilling $\operatorname{dom} \nabla H := \{x \in \mathbb{R}^n : H \text{ is differentiable at } x\} \neq \emptyset$ is defined as

$$D_H : \mathbb{R}^n \times \operatorname{dom} \nabla H \to \overline{\mathbb{R}}, \quad D_H(x, y) := H(x) - H(y) - \langle \nabla H(y), x - y\rangle.$$

The following algorithm relies on the stochastic incremental mirror descent approach of [10, Algorithm 3.2], but instead of using subgradients of the functions f_i we smooth them by the Nesterov smoothing approach (3) and employ the gradients of the smooth functions, provided by Lemma (3.4).

Algorithm 3.10

Choose $x_0 \in \bigcap_{i=1}^{m} \operatorname{dom} f_i \cap C$, $y_{m,-1} \in \mathbb{R}^n$, the smoothing parameters $\gamma_k > 0$ and the stepsizes $t_k > 0$, $k \geq 0$:
for all $k \geq 0$ **do**
  $\psi_{0,k} := x_k$
  $y_{0,k} := y_{m,k-1}$
  for all $i := 1, \ldots, m$ **do**
    $y_{i,k} := y_{i-1,k} - \epsilon_{i,k} \frac{t_k}{p_i} \nabla f_i^{\gamma_k}(\psi_{i-1,k})$
    $\psi_{i,k} := \nabla H^*(y_{i,k})$
  end for
  $x_{k+1} := \psi_{m,k}$
end for,
where $\epsilon_{i,k} \in \{0, 1\}$ is a random variable independent of $\psi_{i-1,k}$ and $\mathbb{P}(\epsilon_{i,k} = 1) = p_i$ for all $1 \leq i \leq m$ and $k \geq 0$.

Remark 3.11. The hypothesis $\operatorname{Im}\nabla H^* \subseteq \cap_{i=1}^m \operatorname{dom} f_i$ guarantees that the sequence $\{x_k\}_k$ generated by Algorithm 3.10 contains only elements that lie in the intersection of the domains of the functions f_i, $i = 1, \ldots, m$.

Theorem 3.12. *For Problem 3.1 let the sequence $\{x_k\}_k$ generated by Algorithm 3.10 and for a constant $\delta > 0$ take $\gamma_k := t_k\delta/\sigma$, $k \geq 0$. Then for all $N \geq 1$ and $y \in \mathbb{R}^n$ it holds*

$$\mathbb{E}\left(\min_{0\leq k\leq N-1}\sum_{i=1}^m f_i(x_k) - \sum_{i=1}^m f_i(y)\right) \leq$$
$$\frac{d_H(y, x_0, y_{0,0}) + \frac{1}{\sigma}\left(\delta\sum_{i=1}^m D_{U_i} + 2\left(\sum_{i=1}^m \|A_i\|\sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2\left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)\right)\sum_{k=0}^{N-1} t_k^2}{\sum_{k=0}^{N-1} t_k}.$$

Proof. The begin is as in the proof of [10, Theorem 3.3] (instead of f_i we have $f_i^{\gamma_k}$ for all $i = 1, \ldots, m$, and instead of $\|v_i(\psi_{i-1,k})\|^2 \leq L_{f_i}^2$ (where $v_i(\psi_{i-1,k})$ is a subgradient of f_i at $\psi_{i-1,k}$) we have from Lemma 3.4 $\|\nabla f_i^{\gamma_k}\|^2 \leq 2\|A_i\|^2(2D_{U_i} + \|u_i^c\|^2)$. So we can start from [(6), [10]] for every $k \geq 0$ with these modifications

$$\mathbb{E}(d_H(y, \psi_{m,k}, y_{m,k})) \leq \mathbb{E}(d_H(y, x_k, y_{0,k})) + t_k\mathbb{E}\left(\sum_{i=1}^m f_i^{\gamma_k}(y) - \sum_{i=1}^m f_i^{\gamma_k}(x_k)\right)$$
$$+\frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} - \mathbb{E}\left(\sum_{i=1}^m \frac{1}{2}d_H(\psi_{i,k}, \psi_{i-1,k}, y_{i-1,k})\right)$$
$$+\mathbb{E}\left(t_k\sum_{i=1}^m (f_i^{\gamma_k}(x_k) - f_i^{\gamma_k}(\psi_{i-1,k}))\right). \tag{5}$$

Using the Lipschitz continuity of ∇H^* and $\nabla f_i^{\gamma_k}$ and Lemma 3.4 it yields for every $k \geq 0$

$$\sum_{i=1}^m (f_i^{\gamma_k}(x_k) - f_i^{\gamma_k}(\psi_{i-1,k})) \leq \sum_{i=2}^m\sum_{j=1}^{i-1}(f_i^{\gamma_k}(\psi_{j-1,k}) - f_i^{\gamma_k}(\psi_{j,k}))$$
$$\leq \sum_{i=2}^m\sum_{j=1}^{i-1}\langle\nabla f_i^{\gamma_k}(\psi_{j-1,k}), \psi_{j-1,k} - \psi_{j,k}\rangle \leq \sum_{i=2}^m\sum_{j=1}^{i-1}\|\nabla f_i^{\gamma_k}(\psi_{j-1,k})\|\|\psi_{j-1,k} - \psi_{j,k}\| \leq$$
$$\sum_{i=2}^m\sum_{j=1}^{i-1}\|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\|\psi_{j-1,k} - \psi_{j,k}\| \leq \sum_{l=1}^m\|A_l\|\sqrt{4D_{U_l} + 2\|u_l^c\|^2}\sum_{i=2}^m\|\psi_{i-1,k} - \psi_{i,k}\|$$
$$\leq \sum_{l=1}^m\|A_l\|\sqrt{4D_{U_l} + \|u_l^c\|^2}\sum_{i=2}^m\|\nabla H^*(y_{i-1,k}) - \nabla H^*(y_{i,k})\|$$
$$\leq \frac{1}{\sigma}\sum_{l=1}^m\|A_l\|\sqrt{4D_{U_l} + 2\|u_l^c\|^2}\sum_{i=2}^m\|y_{i-1,k} - y_{i,k}\|$$
$$= \frac{1}{\sigma}\sum_{l=1}^m\|A_l\|\sqrt{4D_{U_l} + 2\|u_l^c\|^2}\sum_{i=2}^m\|\epsilon_{i,k}\frac{t_k}{p_i}\nabla f_i^{\gamma_k}(\psi_{i-1,k})\|$$
$$\leq \frac{1}{\sigma}t_k\sum_{l=1}^m\|A_l\|\sqrt{4D_{U_l} + 2\|u_l^c\|^2}\sum_{i=1}^m\frac{\epsilon_{i,k}}{p_i}\|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}.$$

So, for every $k \geq 0$ it holds

$$\begin{aligned}
&\mathbb{E}\left(t_k \sum_{i=1}^{m}(f_i^{\gamma_k}(x_k) - f_i^{\gamma_k}(\psi_{i-1,k}))\right) \\
&\leq \frac{1}{\sigma} t_k^2 \left(\sum_{l=1}^{m} \|A_l\| \sqrt{4D_{U_l} + 2\|u_l^c\|^2}\right) \mathbb{E}\left(\sum_{i=1}^{m} \frac{\epsilon_{i,k}}{p_i} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right) \\
&\leq \frac{1}{\sigma} t_k^2 \left(\sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 . \qquad (6)
\end{aligned}$$

Inequality (4) yields for every $k \geq 0$

$$t_k \mathbb{E}\left(\sum_{i=1}^{m} f_i^{\gamma_k}(y) - \sum_{i=1}^{m} f_i^{\gamma_k}(x_k)\right) \leq t_k \left(\mathbb{E}\left(\sum_{i=1}^{m} f_i(y) - \sum_{i=1}^{m} f_i(x_k)\right) + \gamma_k \sum_{i=1}^{m} D_{U_i}\right). \qquad (7)$$

Combining (5) with (6) and (7) gives for every $k \geq 0$

$$\begin{aligned}
\mathbb{E}(d_H(y, \psi_{m,k}, y_{m,k})) \leq \mathbb{E}(d_H(y, x_k, y_{0,k})) + t_k \left(\mathbb{E}\left(\sum_{i=1}^{m} f_i(y) - \sum_{i=1}^{m} f_i(x_k)\right) + \gamma_k \sum_{i=1}^{m} D_{U_i}\right) \\
+ \frac{1}{\sigma} t_k^2 \left(\sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 \left(\sum_{i=1}^{m} \frac{1}{p_i^2}\right)^{\frac{1}{2}} - \mathbb{E}\left(\sum_{i=1}^{m} \frac{1}{2} d_H(\psi_{i,k}, \psi_{i-1,k}, y_{i-1,k})\right) \\
+ \frac{1}{\sigma} t_k^2 \left(\sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 .
\end{aligned}$$

Because $\psi_{m,k} = x_{k+1}, y_{m,k} = y_{0,k+1}$ and $d_H(\psi_{i,k}, \psi_{i-1,k}, y_{i-1,k}) \geq 0$ as $y_{i-1,k} \in \partial H(\psi_{i-1,k})$ it holds for every $k \geq 0$

$$\begin{aligned}
\mathbb{E}(d_H(y, x_{k+1}, y_{0,k+1})) \leq &\mathbb{E}(d_H(y, x_k, y_{0,k})) + t_k \mathbb{E}\left(\sum_{i=1}^{m} f_i(y) - \sum_{i=1}^{m} f_i(x_k)\right) + t_k \gamma_k \sum_{i=1}^{m} D_{U_i} \\
&+ \frac{1}{\sigma} t_k^2 \left(\sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=1}^{m} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right).
\end{aligned}$$

Summing up the inequality from $k = 0$ to $N - 1$, where $N \geq 1$, we obtain

$$\begin{aligned}
&\sum_{k=0}^{N-1} t_k \mathbb{E}\left(\sum_{i=1}^{m} f_i(x_k) - \sum_{i=1}^{m} f_i(y)\right) + \mathbb{E}(d_H(y, x_N, y_{0,N})) \leq \mathbb{E}(d_H(y, x_0, y_{0,0})) \\
&+ \sum_{i=1}^{m} D_{U_i} \sum_{k=0}^{N-1} \gamma_k t_k + \frac{1}{\sigma} \left(\sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=1}^{m} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right) \sum_{k=0}^{N-1} t_k^2 .
\end{aligned}$$

Since $d_H(y, x_N, y_{0,N}) \geq 0$, as $y_{0,N} \in \partial H(x_N)$, and $\gamma_k = \frac{t_k \delta}{\sigma}$, we get

$$\begin{aligned}
&\mathbb{E}\left(\min_{0 \leq k \leq N-1} \sum_{i=1}^{m} f_i(x_k) - \sum_{i=1}^{m} f_i(y)\right) \leq \\
&\frac{d_H(y, x_0, y_{0,0}) + \frac{1}{\sigma}\left(\delta \sum_{i=1}^{m} D_{U_i} + 2\left(\sum_{i=1}^{m} \|A_i\| \sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=1}^{m} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)\right) \sum_{k=0}^{N-1} t_k^2}{\sum_{k=0}^{N-1} t_k} .
\end{aligned}$$

□

In the following corollary we give the optimal stepsize choice for Algorithm 3.10, which follows from [7, Proposition 4.1] and delivers the lowest value of the expression in the right-hand side of the convergence statement in Theorem 3.12.

Corollary 3.13. *Let $x^* \in \operatorname{dom} H$ be an optimal solution to (1) and for a constant $\delta > 0$ let $\gamma_k := t_k\delta/\sigma$, $k \geq 0$. Then the optimal stepsize for the algorithm above is given by*

$$t_k := \sqrt{\frac{\sigma d_H(x^*, x_{0,0}, y_{0,0})}{\delta \sum_{i=1}^m D_{U_i} + 2\left(\sum_{i=1}^m \|A_i\|\sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)}} \frac{1}{\sqrt{k}}, \quad \forall k \geq 0$$

which yields for every $N \geq 1$

$$\mathbb{E}\left(\min_{0\leq k\leq N-1} \sum_{i=1}^m f_i(x_k) - \sum_{i=1}^m f_i(x^*)\right)$$
$$\leq \frac{2}{\sqrt{N}} \sqrt{\frac{d_H(x^*, x_{0,0}, y_{0,0})\left(\delta \sum_{i=1}^m D_{U_i} + 2\left(\sum_{i=1}^m \|A_i\|\sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)\right)}{\sigma}}.$$

Let us consider now the following optimization problem consisting in minimizing a sum of functions fulfilling (2) when composed with linear operators. Such problems can be seen both as special cases and generalizations of Problem 3.1 as mentioned in Remark 3.2. Taking into consideration this remark, only the maximum in the construction (2) needs to be attained in the case of such compositions when the involved functions are proper, convex and semicontinuous, and the operators linear, in which case we say that they fulfill the property $(2')$. Unlike the construction proposed in [10], our approach is flexible enough to allow modifying Algorithm 3.10 in order to solve such problems as well.

Problem 3.14. *We consider the convex optimization problem*

$$\min_{x\in C}\left\{\sum_{i=1}^m f_i(A_i x)\right\}, \tag{8}$$

where $C \subseteq \mathbb{R}^n$ is a nonempty, convex and closed set, $f_i : \mathbb{R}^p \to \overline{\mathbb{R}}$, $i = 1, \ldots, m$ are proper, convex and semicontinuous functions and $A_i : \mathbb{R}^n \to \mathbb{R}^p$ linear operators, such that $(2')$ holds for them and $C \cap (\cap_{i=1}^m \operatorname{dom}(f_i \circ A_i)) \neq \emptyset$.

For $i = 1, \ldots, m$, we smooth the functions $f_i \circ A_i$ via the Moreau-envelope which is a special case of Nesterov smoothing as mentioned above, obtaining $(f_i \circ A_i)^\gamma(\cdot) = (f_i \square (1/2\gamma)\|\cdot\|^2)(A_i\cdot)$ with the gradients $\nabla(f_i \circ A_i)^\gamma(\cdot) = (1/\gamma)(\cdot - A_i^* \operatorname{Prox}_{\gamma f_i}(A_i\cdot))$, where $\gamma > 0$.

Remark 3.15. We will look at an optimization problem for reconstructing images in PET in section 6.2, which is in the setting of Problem 3.14.

We obtain the following mirror descent proximal point algorithm.

Algorithm 3.16

Choose $x_0 \in \bigcap_{i=1}^{m} \operatorname{dom}(f_i \circ A_i) \cap C$, $y_{m,-1} \in \mathbb{R}^n$, the smoothing parameters $\gamma_k > 0$ and the stepsizes $t_k > 0$, $k \geq 0$:
for all $k \geq 0$ **do**
  $\psi_{0,k} := x_k$
  $y_{0,k} := y_{m,k-1}$
  for all $i := 1, \ldots, m$ **do**
    $y_{i,k} := y_{i-1,k} - \epsilon_{i,k} \frac{t_k}{\gamma_k p_i} \left(\psi_{i-1,k} - A_i^* \operatorname{Prox}_{\gamma_k f_i}(A_i \psi_{i-1,k})\right)$
    $\psi_{i,k} := \nabla H^*(y_{i,k})$
  end for
  $x_{k+1} := \psi_{m,k}$
end for,
where $\epsilon_{i,k} \in \{0,1\}$ is a random variable independent of $\psi_{i-1,k}$ and $\mathbb{P}(\epsilon_{i,k} = 1) = p_i$ for all $1 \leq i \leq m$ and $k \geq 0$.

Because this algorithm is derived from Algorithm 3.10 the convergence result of Theorem 3.12 is also valid, where $D_{U_i} = D_{\operatorname{dom} f_i^*} = \sup_{u \in \operatorname{dom} f_i^*} (1/2)\|u\|^2$ and $\|u_i^c\| = 0$.

Theorem 3.17. *For Problem 3.14 let the sequence $\{x_k\}_k$ generated by Algorithm 3.16 and for a constant $\delta > 0$ take $\gamma_k := t_k \delta / \sigma$, $k \geq 0$. Then for all $N \geq 1$ and $y \in \mathbb{R}^n$ it holds*

$$\mathbb{E}\left(\min_{0 \leq k \leq N-1} \sum_{i=1}^{m} f_i(x_k) - \sum_{i=1}^{m} f_i(y)\right) \leq$$
$$\frac{d_H(y, x_0, y_{0,0}) + \frac{1}{\sigma}\left(\delta \sum_{i=1}^{m} D_{dom f_i^*} + 4\left(\sum_{i=1}^{m} \|A_i\| \sqrt{D_{dom f_i^*}}\right)^2 \left(\left(\sum_{i=1}^{m} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)\right) \sum_{k=0}^{N-1} t_k^2}{\sum_{k=0}^{N-1} t_k}.$$

Analogously, the optimal stepsize choice for Algorithm 3.16 is given in the following corollary.

Corollary 3.18. *Let $x^* \in \operatorname{dom} H$ be an optimal solution to (8) and for a constant $\delta > 0$ let $\gamma_k := t_k \delta / \sigma$, $k \geq 0$. Then the optimal stepsize for Algorithm 3.16 above is given by*

$$t_k := \sqrt{\frac{\sigma d_H(x^*, x_0, y_{0,0})}{\delta \sum_{i=1}^{m} D_{dom f_i^*} + 4\left(\sum_{i=1}^{m} \|A_i\| \sqrt{D_{dom f_i^*}}\right)^2 \left(\left(\sum_{i=1}^{m} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)}} \frac{1}{\sqrt{k}}, \quad \forall k \geq 0$$

which yields for every $N \geq 1$

$$\mathbb{E}\left(\min_{0 \leq k \leq N-1} \sum_{i=1}^{m} f_i(A_i x_k) - \sum_{i=1}^{m} f_i(A_i x^*)\right)$$
$$\leq 2\sqrt{\frac{d_H(x^*, x_0, y_{0,0})\left(\delta \sum_{i=1}^{m} D_{dom f_i^*} + 4\left(\sum_{i=1}^{m} \|A_i\| \sqrt{D_{dom f_i^*}}\right)^2 \left(\left(\sum_{i=1}^{m} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)\right)}{\sigma}} \frac{1}{\sqrt{N}}.$$

Remark 3.19. The difference between Algorithm 3.10 and its counterpart in the Lipschitzian case [10, Algorithm 2.2] is that we do not need to know the Lipschitz constants

or the subgradients of the functions f_i ($i = 1, \ldots, m$), which sometimes can be computationally expensive to determine (cf. [1, 12, 17, 44]), but (in particular for its special case Algorithm 3.16) their proximal point mappings which, for many functions, including the ones which usually occur when modelling applications in fields like image deblurring and denoising or machine learning, are already known. A further advantage of our method is that we do not need to impose the Lipschitz continuity of the gradients of the objective functions, as the gradients of their Nesterov smooth approximations satisfy this hypothesis by construction. Instead we ask the weaker condition of closedness of the domains of their conjugates. Note also that by employing the parameters $\gamma_k > 0$, $k \geq 0$, Algorithm 3.10 presents additional flexibility when compared with its mentioned counterpart.

Remark 3.20. Additionally assuming the functions f_i, $i = 1, \ldots, m$, Lipschitz continuous does not make Algorithm 3.10 collapse to [10, Algorithm 3.2] and also the assertion of Theorem 3.12 does not rediscover its counterpart [10, Theorem 3.3] because of the different constructions. This has motivated us to include in our study the results in Subsection 5 where combinations of these algorithms are proposed.

4 Incremental mirror descent Bregman-prox-scheme with Nesterov smoothing

In this section we consider an extension of the optimization problem (1) by adding another nonsmooth function to its objective function. The iterative scheme we propose for solving it is an extension of Algorithm 3.10, but instead of smoothing the new function, we evaluate it by a proximal step of Bregman type. For this we need additional differentiability assumptions on the function which induces the mirror map.

Problem 4.1. *We consider the convex optimization problem*

$$\min_{x \in C} \left\{ \sum_{i=1}^{m} f_i(x) + g(x) \right\}, \tag{9}$$

where $C \subseteq \mathbb{R}^n$ is a nonempty, convex and closed set, for $i = 1, \ldots, m$, the functions $f_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ are defined like in Problem 3.1 and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper, convex and lower semicontinuous function such that $C \cap (\cap_{i=1}^m \operatorname{dom} f_i \cap \operatorname{dom} g) \neq \emptyset$. Furthermore, let $H : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper, lower semicontinuous and σ-strongly convex function (for $\sigma > 0$) such that $C = \operatorname{cl}(\operatorname{dom} H)$, let H be continuously differentiable on $\operatorname{int}(\operatorname{dom} H)$, $\operatorname{Im} \nabla H^* \subseteq (\cap_{i=1}^m \operatorname{dom} f_i) \cap \operatorname{int}(\operatorname{dom} H)$ and $\operatorname{int}(\operatorname{dom} H) \cap \operatorname{dom} g \neq \emptyset$.

Definition 4.2. Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper, convex, lower semicontinuous function. The *Bregman-proximal operator* of h with respect to the proper, lower semicontinuous and σ-strongly convex function H is defined as

$$\operatorname{Prox}_h^H : \operatorname{dom} \nabla H \to \mathbb{R}^n, \ \operatorname{Prox}_h^H(x) := \arg\min_{u \in \mathbb{R}^n} \{h(u) + D_H(u, x)\}.$$

Because H is σ-strongly convex, the Bregman-proximal operator is well defined. For $H = (1/2)\|\cdot\|^2$ the Bregman-proximal operator is the classical proximity operator.

We propose the following algorithm for solving the optimization problem (9).

Algorithm 4.3

Choose $x_0 \in \operatorname{Im} \nabla H^* \cap C$, the smoothing parameters $\gamma_k > 0$ and the stepsizes $t_k > 0$, $k \geq 0$:
for all $k \geq 0$ **do**
  $\psi_{0,k} := x_k$
  for all $i := 1, \ldots, m$ **do**
    $\psi_{i,k} := \nabla H^*(\nabla H(\psi_{i-1,k}) - \epsilon_{i,k} \frac{t_k}{p_i} \nabla f_i^{\gamma_k}(\psi_{i-1,k}))$
  end for
  $x_{k+1} := \operatorname{Prox}_{t_k g}^H(\psi_{m,k})$
end for,
where $\epsilon_{i,k} \in \{0,1\}$ is a random variable independent of $\psi_{i-1,k}$ and $\mathbb{P}(\epsilon_{i,k} = 1) = p_i$ for all $1 \leq i \leq m$ and $k \geq 0$.

Remark 4.4. Note that when $g = 0$ Algorithm 4.3 corresponds essentially to Algorithm 3.10. But even for this case the constants obtained in the convergence result given below and in Theorem 3.12 are not the same due to the construction of the algorithms (note, for instance, that Algorithm 3.10 requires an additional starting point) and therefore to some main differences in the proofs.

Theorem 4.5. *Let the sequence $\{x_k\}_k$ generated by Algorithm 4.3 and for a constant $\delta > 0$ let $\gamma_k := t_k \delta / \sigma$. Then for all $N \geq 1$ and all $y \in \mathbb{R}^n$ one has*

$$\mathbb{E}\left(\min_{0 \leq k \leq N-1} \left(\sum_{i=1}^m f_i + g\right)(x_{k+1}) - \left(\sum_{i=1}^m f_i + g\right)(y)\right) \leq$$

$$\frac{D_H(y, x_0) + \frac{1}{\sigma}\left(\delta \sum\limits_{i=1}^m D_{U_i} + 2\left(\sum\limits_{i=1}^m \|A_i\| \sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum\limits_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + \frac{3}{2} + m\right)\right) \sum\limits_{k=0}^{N-1} t_k^2}{\sum\limits_{k=0}^{N-1} t_k}.$$

Proof. When $y \notin \cap_{i=1}^m \operatorname{dom} f_i \cap \operatorname{dom} g$ the assertion follows automatically, so we consider further $y \in \cap_{i=1}^m \operatorname{dom} f_i \cap \operatorname{dom} g$. We start the proof with inequalities (5) and (6) from Theorem 3.12 and use instead of the Bregman distance like functions the Bregman distance to obtain

$$\mathbb{E}(D_H(y, \psi_{m,k})) \leq \mathbb{E}(D_H(y, x_k)) + t_k \mathbb{E}\left(\sum_{i=1}^m f_i^{\gamma_k}(y) - \sum_{i=1}^m f_i^{\gamma_k}(x_k)\right)$$

$$+ \frac{1}{\sigma} t_k^2 \left(\sum_{i=1}^m \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right) - \mathbb{E}\left(\sum_{i=1}^m \frac{1}{2} D_H(\psi_{i,k}, \psi_{i-1,k})\right). \tag{10}$$

Like in [(12), [10]] we get for every $k \geq 0$

$$t_k \mathbb{E}((g(x_{k+1}) - g(y))) + \mathbb{E}(D_H(y, x_{k+1})) \leq \mathbb{E}(D_H(y, \psi_{m,k})) - \mathbb{E}(D_H(x_{k+1}, \psi_{m,k})). \tag{11}$$

By combining (10) and (11) we obtain for every $k \geq 0$

$$t_k\mathbb{E}((g(x_{k+1}) - g(y))) + t_k\mathbb{E}\left(\sum_{i=1}^m f_i^{\gamma_k}(x_k) - \sum_{i=1}^m f_i^{\gamma_k}(y)\right) + \mathbb{E}(D_H(y, x_{k+1}))$$
$$\leq \mathbb{E}(D_H(y, x_k)) + \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2\left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)$$
$$- \mathbb{E}(D_H(x_{k+1}, \psi_{m,k})) - \sum_{i=1}^m \frac{1}{2}\mathbb{E}(D_H(\psi_{i,k}, \psi_{i-1,k})).$$

We add and subtract $t_k\mathbb{E}(\sum_{i=1}^m f_i^{\gamma_k}(x_{k+1}))$ to get

$$t_k\mathbb{E}\left(\left(\sum_{i=1}^m f_i^{\gamma_k} + g\right)(x_{k+1}) - \left(\sum_{i=1}^m f_i^{\gamma_k} + g\right)(y)\right)$$
$$+ t_k\mathbb{E}\left(\sum_{i=1}^m f_i^{\gamma_k}(x_k) - \sum_{i=1}^m f_i^{\gamma_k}(x_{k+1})\right) + \mathbb{E}(D_H(y, x_{k+1}))$$
$$\leq \mathbb{E}(D_H(y, x_k)) + \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2\left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)$$
$$- \mathbb{E}(D_H(x_{k+1}, \psi_{m,k})) - \sum_{i=1}^m \frac{1}{2}\mathbb{E}(D_H(\psi_{i,k}, \psi_{i-1,k})).$$

Because of the differentiability and convexity of $f_i^{\delta/\sigma t_k}$, $(i = 1, \ldots, m)$ for all $k \geq 0$ and Lemma 3.4 we have

$$-t_k\mathbb{E}\left(\sum_{i=1}^m f_i^{\gamma_k}(x_{k+1}) - \sum_{i=1}^m f_i^{\gamma_k}(x_k)\right) \geq -t_k\mathbb{E}\left(\left\|\sum_{i=1}^m \nabla f_i^{\gamma_k}(x_{k+1})\right\| \|x_k - x_{k+1}\|\right)$$
$$\geq -t_k\mathbb{E}\left(\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\|x_k - x_{k+1}\|\right)$$
$$\geq -t_k\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\mathbb{E}\left(\|x_k - x_{k+1}\|\right) \quad (12)$$

and from (4) it follows that

$$t_k\left(\mathbb{E}\left(\left(\sum_{i=1}^m f_i + g\right)(x_{k+1}) - \left(\sum_{i=1}^m f_i + g\right)(y)\right) - \gamma_k\sum_{i=1}^m D_{U_i}\right)$$
$$- t_k\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\mathbb{E}\left(\|x_k - x_{k+1}\|\right) + \mathbb{E}(D_H(y, x_{k+1}))$$
$$\leq \mathbb{E}(D_H(y, x_k)) + \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2\left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)$$
$$- \mathbb{E}(D_H(x_{k+1}, \psi_{m,k})) - \sum_{i=1}^m \frac{1}{2}\mathbb{E}(D_H(\psi_{i,k}, \psi_{i-1,k})). \quad (13)$$

By the triangle inequality we get for every $k \geq 0$

$$t_k\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\mathbb{E}\left(\|x_k - x_{k+1}\|\right) \leq t_k\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\mathbb{E}\left(\|x_k - \psi_{m,k}\|\right)$$
$$+ t_k\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\mathbb{E}\left(\|\psi_{m,k} - x_{k+1}\|\right). \quad (14)$$

Using Young's inequality and the strong convexity of H we have

$$t_k \sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \mathbb{E}\left(\|\psi_{m,k} - x_{k+1}\|\right)$$

$$\leq \frac{1}{2\sigma} t_k^2 \left(\sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \right)^2 + \frac{\sigma}{2} \mathbb{E}\left(\|\psi_{m,k} - x_{k+1}\|^2\right)$$

$$\leq \frac{1}{2\sigma} t_k^2 \left(\sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \right)^2 + \mathbb{E}(H(x_{k+1}) - H(\psi_{m,k}) - \langle \nabla H(x_{k+1}), x_{k+1} - \psi_{m,k} \rangle)$$

$$= \frac{1}{2\sigma} t_k^2 \left(\sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \right)^2 + \mathbb{E}(D_H(x_{k+1}, \psi_{m,k})),$$

and since

$$\|x_k - \psi_{m,k}\| = \left\| \sum_{i=1}^{m} (\psi_{i-1,k} - \psi_{i,k}) \right\| \leq \sum_{i=1}^{m} \|\psi_{i-1,k} - \psi_{i,k}\|,$$

the inequality (14) becomes

$$t_k \sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \mathbb{E}\left(\|x_k - x_{k+1}\|\right) \leq \frac{1}{2\sigma} t_k^2 \left(\sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \right)^2$$

$$+\mathbb{E}(D_H(x_{k+1}, \psi_{m,k})) \quad + t_k \sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \mathbb{E}\left(\sum_{i=1}^{m} \|\psi_{i-1,k} - \psi_{i,k}\| \right).$$

Using Young's inequality and the strong convexity of H we get for every $i = 1, \ldots, m$, and every $k \geq 0$

$$t_k \sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \mathbb{E}(\|\psi_{i-1,k} - \psi_{i,k}\|) \leq \frac{1}{\sigma} t_k^2 \left(\sum_{j=1}^{m} \|A_j\| \sqrt{4D_{U_j} + 2\|u_j^c\|^2} \right)^2$$

$$+\frac{\sigma}{4} \mathbb{E}(\|\psi_{i-1,k} - \psi_{i,k}\|^2) \leq \frac{1}{\sigma} t_k^2 \left(\sum_{j=1}^{m} \|A_j\| \sqrt{4D_{U_j} + 2\|u_j^c\|^2} \right)^2 + \frac{1}{2} \mathbb{E}(D_H(\psi_{i,k}, \psi_{i-1,k})),$$

so we have

$$t_k \sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \mathbb{E}\left(\|x_k - x_{k+1}\|\right) \leq \frac{1}{2\sigma} t_k^2 \left(\sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \right)^2$$

$$+\mathbb{E}(D_H(x_{k+1}, \psi_{m,k})) + \frac{1}{\sigma} m t_k^2 \left(\sum_{j=1}^{m} \|A_j\| \sqrt{4D_{U_j} + 2\|u_j^c\|^2} \right)^2 + \sum_{i=1}^{m} \frac{1}{2} \mathbb{E}(D_H(\psi_{i,k}, \psi_{i-1,k})). \tag{15}$$

Combining (15) and (13) we obtain

$$t_k \mathbb{E}\left(\left(\sum_{i=1}^{m} f_i + g \right)(x_{k+1}) - \left(\sum_{i=1}^{m} f_i + g \right)(y) \right) + \mathbb{E}(D_H(y, x_{k+1})) \leq \mathbb{E}(D_H(y, x_k))$$

$$+ \frac{1}{\sigma} t_k^2 \left(\sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \right)^2 \left(\left(\sum_{i=1}^{m} \frac{1}{p_i^2} \right)^{\frac{1}{2}} + \frac{3}{2} + m \right) + t_k \gamma_k \sum_{i=1}^{m} D_{U_i}.$$

Summing up this inequality from $k = 0$ to $N-1$, for $N \geq 1$, we get

$$\sum_{k=0}^{N-1} t_k \mathbb{E}\left(\left(\sum_{i=1}^{m} f_i + g\right)(x_{k+1}) - \left(\sum_{i=1}^{m} f_i + g\right)(y)\right) + \mathbb{E}(D_H(y, x_N)) \leq \mathbb{E}(D_H(y, x_0))$$

$$+\frac{1}{\sigma}\left(\sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=1}^{m} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + \frac{3}{2} + m\right) \sum_{k=0}^{N-1} t_k^2 + \sum_{i=1}^{m} D_{U_i} \sum_{k=0}^{N-1} \gamma_k t_k.$$

Since $\mathbb{E}(D_H(y, x_N)) \geq 0$ and $\gamma_k = t_k \delta/\sigma$, $k \geq 0$, we obtain

$$\mathbb{E}\left(\min_{0 \leq k \leq N-1} \left(\sum_{i=1}^{m} f_i + g\right)(x_{k+1}) - \left(\sum_{i=1}^{m} f_i + g\right)(y)\right) \leq$$

$$\frac{D_H(y, x_0) + \frac{1}{\sigma}\left(\delta \sum_{i=1}^{m} D_{U_i} + 2\left(\sum_{i=1}^{m} \|A_i\| \sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=1}^{m} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + \frac{3}{2} + m\right)\right) \sum_{k=0}^{N-1} t_k^2}{\sum_{k=0}^{N-1} t_k}.$$

□

In the following corollary we give the optimal stepsize choice for Algorithm 4.3, which follows from [7, Proposition 4.1].

Corollary 4.6. *Let $x^* \in \operatorname{dom} H$ be an optimal solution to (9) and for a constant $\delta > 0$ let $\gamma_k := t_k \delta/\sigma$, $k \geq 0$. Then the optimal stepsize for Algorithm 4.3 is given by*

$$t_k := \sqrt{\frac{\sigma D_H(x^*, x_0)}{\delta \sum_{i=1}^{m} D_{U_i} + 2\left(\sum_{i=1}^{m} \|A_i\| \sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=1}^{m} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + \frac{3}{2} + m\right)}} \frac{1}{\sqrt{k}} \quad \forall k \geq 0$$

which yields for every $N \geq 1$

$$\mathbb{E}\left(\min_{0 \leq k \leq N-1} \left(\sum_{i=1}^{m} f_i + g\right)(x_k) - \left(\sum_{i=1}^{m} f_i + g\right)(x^*)\right)$$

$$\leq 2\sqrt{\frac{D_H(,x_0)\left(\delta \sum_{i=1}^{m} D_{U_i} + 2\left(\sum_{i=1}^{m} \|A_i\| \sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=1}^{m} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + \frac{3}{2} + m\right)\right)}{\sigma}} \frac{1}{\sqrt{N}}.$$

As in the previous section we modify the considered problem by composing the smoothed functions with the linear operators used in their construction.

Problem 4.7. *We consider the convex optimization problem*

$$\min_{x \in C} \left\{\sum_{i=1}^{m} f_i(A_i x) + g(x)\right\},$$

where $C \subseteq \mathbb{R}^n$ is a nonempty, convex and closed set, $f_i : \mathbb{R}^p \to \overline{\mathbb{R}}$, $i = 1, \ldots, m$ are proper, convex and semicontinuous functions and $A_i : \mathbb{R}^n \to \mathbb{R}^p$ linear operators, such that $(2')$ holds for them and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper, convex and lower semicontinuous function such that $C \cap (\cap_{i=1}^{m} \operatorname{dom}(f_i \circ A_i)) \cap \operatorname{dom} g \neq \emptyset$. Furthermore, let the function H defined as in Problem 4.1.

By smoothing the functions f_i $(i = 1, \dots, m)$ via the Moreau-envelope, we obtain

$$(f_i \circ A_i)^\gamma(\cdot) = (f_i \square (1/2\gamma)\|\cdot\|^2)(A_i \cdot)$$

with the gradients

$$\nabla (f_i \circ A_i)^\gamma(\cdot) = (1/\gamma)(\cdot - A_i^* \operatorname{Prox}_{\gamma f_i}(A_i \cdot))$$

as in the previous section.

Remark 4.8. The linear SVM classification problem, which we will study for numerical experiments in section 6.3, is a special case of the optimization problem considered above.

We obtain from Algorithm 4.3 the following mirror descent proximal point algorithm for solving Problem 4.7.

Algorithm 4.9

Choose $x_0 \in \operatorname{Im} \nabla H^* \cap C$, the smoothing parameters $\gamma_i > 0$ and the stepsizes $t_k > 0$, $k \geq 0$:
for all $k \geq 0$ **do**
  $\psi_{0,k} := x_k$
  for all $i := 1, \dots, m$ **do**
    $\psi_{i,k} := \nabla H^*(\nabla H(\psi_{i-1,k}) - \epsilon_{i,k} \frac{t_k}{\gamma_k p_i} (\psi_{i-1,k} - A_i^* \operatorname{Prox}_{\gamma_i f_i}(A_i \psi_{i-1,k}))$
  end for
  $x_{k+1} := \operatorname{Prox}^H_{t_k g}(\psi_{m,k})$,
end for
where $\epsilon_{i,k} \in \{0, 1\}$ is a random variable independent of $\psi_{i-1,k}$ and $\mathbb{P}(\epsilon_{i,k} = 1) = p_i$ for all $1 \leq i \leq m$ and $k \geq 0$.

Because this algorithm is derived from Algorithm 4.3 the convergence result follows directly from Theorem 4.5, where $D_{U_i} = D_{\operatorname{dom} f_i^*} = \sup_{u \in \operatorname{dom} f_i^*} \frac{1}{2}\|u\|^2$ and $\|u_i^c\| = 0$.

Theorem 4.10. *Let the sequence $\{x_k\}_k$ generated by Algorithm 4.9 and for a constant $\delta > 0$ let $\gamma_k := t_k \delta / \sigma$, $k \geq 0$. Then for all $N \geq 1$ and all $y \in \mathbb{R}^n$ one has*

$$\mathbb{E}\left(\min_{0 \leq k \leq N-1} \left(\sum_{i=1}^m f_i \circ A_i + g \right)(x_{k+1}) - \left(\sum_{i=1}^m f_i \circ A_i + g \right)(y) \right) \leq$$

$$\frac{D_H(y, x_0) + \frac{1}{\sigma} \left(\delta \sum_{i=1}^m D_{dom f_i^*} + 4 \left(\sum_{i=1}^m \|A_i\| \sqrt{D_{dom f_i^*}} \right)^2 \left(\left(\sum_{i=1}^m \frac{1}{p_i^2} \right)^{\frac{1}{2}} + \frac{3}{2} + m \right) \right) \sum_{k=0}^{N-1} t_k^2}{\sum_{k=0}^{N-1} t_k}.$$

Analogously, the optimal stepsize choice for Algorithm 4.9 is given by the following corollary.

Corollary 4.11. *Let $x^* \in \operatorname{dom} H$ be an optimal solution to Problem 4.7 and for a constant $\delta > 0$ let $\gamma_k := t_k \delta / \sigma$, $k \geq 0$. Then the optimal stepsize for Algorithm 4.9 is given by*

$$t_k := \sqrt{\frac{\sigma D_H(y, x_0)}{\delta \sum_{i=1}^m D_{dom f_i^*} + 4 \left(\sum_{i=1}^m \|A_i\| \sqrt{D_{dom f_i^*}} \right)^2 \left(\left(\sum_{i=1}^m \frac{1}{p_i^2} \right)^{\frac{1}{2}} + \frac{3}{2} + m \right)}} \frac{1}{\sqrt{k}}, \quad \forall k \geq 0$$

which yields for every $N \geq 1$

$$\mathbb{E}\left(\min_{0\leq k\leq N-1}\left(\sum_{i=1}^{m} f_i \circ A_i + g\right)(x_k) - \left(\sum_{i=1}^{m} f_i \circ A_i + g\right)(x^*)\right)$$

$$\leq 2\sqrt{\frac{D_H(y,x_0)\left(\delta\sum_{i=1}^{m} D_{domf_i^*} + 4\left(\sum_{i=1}^{m}\|A_i\|\sqrt{D_{domf_i^*}}\right)^2\left(\left(\sum_{i=1}^{m}\frac{1}{p_i^2}\right)^{\frac{1}{2}} + \frac{3}{2} + m\right)\right)}{\sigma}}\frac{1}{\sqrt{N}}.$$

Remark 4.12. One can provide counterparts to [10, Remark 3.7 and Remark 4.7] in the framework considered in this work, too. We leave them to the interested reader.

5 Stochastic incremental mirror descent algorithms with subgradient and Nesterov smoothing

In the following we combine the mirror descent algorithms proposed above, which use the Nesterov smoothing approach, and the mirror descent algorithms proposed in [10], which use the subgradients of the objective functions to minimize.

Problem 5.1. *We consider the convex optimization problem*

$$\min_{x\in C}\left\{\sum_{i=1}^{m_1} f_i(x) + \sum_{i=m_1+1}^{m_2} f_i(x)\right\}, \tag{16}$$

where $C \subseteq \mathbb{R}^n$ *is a nonempty, convex and closed set such that* $C \cap (\cap_{i=1}^{m_2} \operatorname{dom} f_i) \neq \emptyset$*, for all* $i = 1,\ldots,m_1$ *(*$m_1 \in \mathbb{N}$*), the functions* $f_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ *are proper, convex and* L_{f_i}*-Lipschitz continuous on* $\operatorname{Im}\nabla H^*$*, where* H *is defined as in Problem 4.1, and for all* $i = m_1+1,\ldots,m_2$ *(*$m_1 \leq m_2 \in \mathbb{N}$*), the functions* $f_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ *fulfill* $f_i(x) = \max_{u\in U_i}\{\langle A_i x, u\rangle - \phi_i(u)\}$ *for* $x \in \operatorname{dom} f_i$*, where* $U_i \subseteq \mathbb{R}^p$ *is a compact and convex set,* $A_i : \mathbb{R}^n \to \mathbb{R}^p$ *are linear operators and* $\phi_i : \mathbb{R}^p \to \overline{\mathbb{R}}$ *are proper, lower semicontinuous and convex functions.*

For the following algorithm we use subgradients of the first m_1 functions f_i and the gradients of smooth functions $f_i^{\gamma_k}$ for $i = m_1+1,\ldots,m_2$.

Algorithm 5.2

Choose $x_0 \in \bigcap_{i=1}^{m_2} \operatorname{dom} f_i \cap C$, $y_{m_2,-1} \in \mathbb{R}^n$, the smoothing parameters $\gamma_k > 0$ and the stepsizes $t_k > 0$, $k \geq 0$:
for all $k \geq 0$ **do**
  $\psi_{0,k} := x_k$
  $y_{0,k} := y_{m_2,k-1}$
  for all $i := 1, \ldots, m_1$ **do**
    $y_{i,k} := y_{i-1,k} - \epsilon_{i,k} \frac{t_k}{p_i} v_i(\psi_{i-1,k})$
    $\psi_{i,k} := \nabla H^*(y_{i,k})$
  end for
  for all $i := m_1 + 1, \ldots, m_2$ **do**
    $y_{i,k} := y_{i-1,k} - \epsilon_{i,k} \frac{t_k}{p_i} \nabla f_i^{\gamma_k}(\psi_{i-1,k})$
    $\psi_{i,k} := \nabla H^*(y_{i,k})$
  end for
  $x_{k+1} := \psi_{m_2,k}$
end for,
where $\epsilon_{i,k} \in \{0, 1\}$ is a random variable independent of $\psi_{i-1,k}$ and $\mathbb{P}(\epsilon_{i,k} = 1) = p_i$ for all $1 \leq i \leq m_2$ and $k \geq 0$, and $v_i(\psi_{i-1,k})$ is a subgradient of f_i at $\psi_{i-1,k}$.

In the following statement we give the convergence result for this algorithm. The proof is basically a combination of the ones of Theorem 3.12 and [10, Theorem 3.3], hence it is skipped.

Theorem 5.3. *For Problem 5.1 let the sequence $\{x_k\}_k$ generated by the algorithm above and for a $\delta > 0$ $\gamma_k := t_k \delta / \sigma$. Then for all $N \geq 1$ and $y \in \mathbb{R}^n$ it holds*

$$\mathbb{E}\left(\min_{0 \leq k \leq N-1} \sum_{i=1}^{m_2} f_i(x_k) - \sum_{i=1}^{m_2} f_i(y) \right) \leq \frac{d_H(y, x_0, y_{0,0}) + \frac{1}{\sigma} C \sum_{k=0}^{N-1} t_k^2}{\sum_{k=0}^{N-1} t_k},$$

where

$$\begin{aligned} C \quad = \quad & \delta \sum_{i=m_1+1}^{m_2} D_{U_i} + \left(\sum_{i=1}^{m_1} L_{f_i} \right)^2 \left(\left(\sum_{i=1}^{m_1} \frac{1}{p_i^2} \right)^{\frac{1}{2}} + 1 \right) \\ & + 2 \left(\sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{2 D_{U_i} + \|u_i^c\|^2} \right)^2 \left(\left(\sum_{i=m_1+1}^{m_2} \frac{1}{p_i^2} \right)^{\frac{1}{2}} + 1 \right). \end{aligned}$$

The optimal stepsize choice for Algorithm 5.2 can be deduced from [7, Proposition 4.1].

Corollary 5.4. *Let $x^* \in \operatorname{dom} H$ be an optimal solution to (17), for a $\delta > 0$ take $\gamma_k := t_k \delta / \sigma$, $k \geq 0$, and*

$$\begin{aligned} P := \; & \delta \sum_{i=m_1+1}^{m_2} D_{U_i} + \left(\sum_{i=1}^{m_1} L_{f_i} \right)^2 \left(\left(\sum_{i=1}^{m_1} \frac{1}{p_i^2} \right)^{\frac{1}{2}} + 1 \right) \\ & + 2 \left(\sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{2 D_{U_i} + \|u_i^c\|^2} \right)^2 \left(\left(\sum_{i=m_1+1}^{m_2} \frac{1}{p_i^2} \right)^{\frac{1}{2}} + 1 \right). \end{aligned}$$

Then the optimal stepsize for the algorithm above is given by

$$t_k := \sqrt{\frac{\sigma d_H(x^*, x_0, y_{0,0})}{P}} \frac{1}{\sqrt{k}},$$

for all $k \geq 0$, which yields for every $N \geq 1$

$$\mathbb{E}\left(\min_{0\leq k\leq N-1} \sum_{i=1}^{m_2} f_i(x_k) - \sum_{i=1}^{m_2} f_i(x^*)\right) \leq 2\sqrt{\frac{d_H(x^*, x_0, y_{0,0})P}{\sigma}} \frac{1}{\sqrt{N}}.$$

Adding another nonsmooth function to the objective function of Problem 5.1 brings into attention the following problem, which can be solved by the algorithm below it.

Problem 5.5. *We consider the convex optimization problem*

$$\min_{x\in C} \left\{ \sum_{i=1}^{m_1} f_i(x) + \sum_{i=m_1+1}^{m_2} f_i(x) + g(x) \right\}, \tag{17}$$

where $(m_1 + 1 < m_2 \in \mathbb{N})$, $C \subseteq \mathbb{R}^n$ is a nonempty, convex and closed set, the functions $f_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ (for $i = 1, \dots m_1$) and $f_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ (for $i = m_1 + 1, \dots m_2$) are defined like in Problem 5.1 and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper, convex and lower semicontinuous function such that $C \cap (\cap_{i=1}^{m_2} \operatorname{dom} f_i \cap \operatorname{dom} g) \neq \emptyset$. Let $H : \mathbb{R}^n \to \overline{\mathbb{R}}$ be defined like in Problem 4.1.

Algorithm 5.6

Choose $x_0 \in \operatorname{Im} \nabla H^* \cap C$, $y_{m_2,-1} \in \mathbb{R}^n$, the smoothing parameters $\gamma_k > 0$ and the stepsizes $t_k > 0$, $k \geq 0$:
for all $k \geq 0$ **do**
  $\psi_{0,k} := x_k$
  $y_{0,k} := y_{m_2,k-1}$
  for all $i := 1, \dots, m_1$ **do**
    $\psi_{i,k} := \nabla H^* \left(\nabla H(\psi_{i-1,k}) - \epsilon_{i,k} \frac{t_k}{p_i} v_i(\psi_{i-1,k}) \right)$
  end for
  for all $i := m_1 + 1, \dots, m_2$ **do**
    $\psi_{i,k} := \nabla H^* \left(\nabla H(\psi_{i-1,k}) - \epsilon_{i,k} \frac{t_k}{p_i} \nabla f_i^{\gamma_k}(\psi_{i-1,k}) \right)$
  end for
  $x_{k+1} := \operatorname{Prox}_{t_k g}^H(\psi_{m_2,k})$.
end for,
where $\epsilon_{i,k} \in \{0, 1\}$ is random variable independent of $\psi_{i-1,k}$ and let $\mathbb{P}(\epsilon_{i,k} = 1) = p_i$ for all $1 \leq i \leq m_2$ and $k \geq 0$, and $v_i(\psi_{i-1,k})$ is a subgradient of f_i at $\psi_{i-1,k}$.

The convergence result and the optimal stepsize t_k, $k \geq 0$, for this algorithm are derivable via Theorem 4.5 and [10, Theorem 4.5], and [7, Proposition 4.1], respectively.

Theorem 5.7. *Let the sequence $\{x_k\}_k$ generated by Algorithm 5.6 and for a $\delta > 0$ $\gamma_k := t_k \delta / \sigma$. Then for all $N \geq 1$*

$$\mathbb{E}\left(\min_{0\leq k\leq N-1} \left(\sum_{i=1}^{m_2} f_i + g\right)(x_{k+1}) - \left(\sum_{i=1}^{m_2} f_i + g\right)(y)\right) \leq \frac{D_H(y, x_0) + \frac{1}{\sigma} C \sum_{k=0}^{N-1} t_k^2}{\sum_{k=0}^{N-1} t_k},$$

where

$$P = \delta \sum_{i=m_1+1}^{m_2} D_{U_i} + 2\left(\sum_{i=m_1+1}^{m_2} \|A_i\|\sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=m_1+1}^{m_2} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)$$

$$+ \left(\sum_{i=1}^{m_1} L_{f_i}\right)^2 \left(\left(\sum_{i=1}^{m_1} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right) + \frac{3}{2}\left(\left(\sum_{i=1}^{m_1} L_{f_i}\right) + 2\sum_{i=m_1+1}^{m_2} \|A_i\|\sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2.$$

Corollary 5.8. *Let $x^* \in \operatorname{dom} H$ be an optimal solution to* (17)*, for a $\delta > 0$ take $\gamma_k := t_k\delta/\sigma$, $k \geq 0$, and*

$$P = \delta \sum_{i=m_1+1}^{m_2} D_{U_i} + 2\left(\sum_{i=m_1+1}^{m_2} \|A_i\|\sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=m_1+1}^{m_2} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)$$

$$+ \left(\sum_{i=1}^{m_1} L_{f_i}\right)^2 \left(\left(\sum_{i=1}^{m_1} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right) + \frac{3}{2}\left(\left(\sum_{i=1}^{m_1} L_{f_i}\right) + 2\sum_{i=m_1+1}^{m_2} \|A_i\|\sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2.$$

Then the optimal stepsize for Algorithm 5.6 is given by

$$t_k := \sqrt{\frac{\sigma D_H(y, x_0)}{P}}\frac{1}{\sqrt{k}},$$

for all $k \geq 0$, which yields for every $N \geq 1$

$$\mathbb{E}\left(\min_{0\leq k\leq N-1}\left(\sum_{i=1}^{m_2} f_i + g\right)(x_k) - \left(\sum_{i=1}^{m_2} f_i + g\right)(x^*)\right) \leq 2\sqrt{\frac{D_H(y, x_0)P}{\sigma}}\frac{1}{\sqrt{N}}.$$

6 Applications

We consider three applications that can be modeled as optimization problems of the format considered in this work. The first of them stems from Logistics and was modeled in [41] as a continuous location optimization problem. We compare the performance of our algorithm with those of three versions of the method proposed in [10]. The other two applications, one in Medical Imaging (more precisely in Tomography) and one in Machine Learning (Support Vector Machines) were discussed in [10], too, and we compare the performance of our algorithm to the stochastic version of the method introduced there. We use the proximal points of the smoothed objective functions instead of their subgradients, motivated also by the fact (noted, for instance, in [18]) that proximal point algorithms tend to solve certain optimization problems faster and cheaper than subgradient methods. To this end we smooth the involved functions in the second and third application with the Moreau-envelope, in the first application with Nesterov's smoothing approach. The experiments were carried out for one run of the algorithms and then averaged over 10 runs (and 100 runs for the first application) of the algorithms as the stochastic methods perform slightly differently on each run due to the stochastic component.

6.1 Continuous location problem

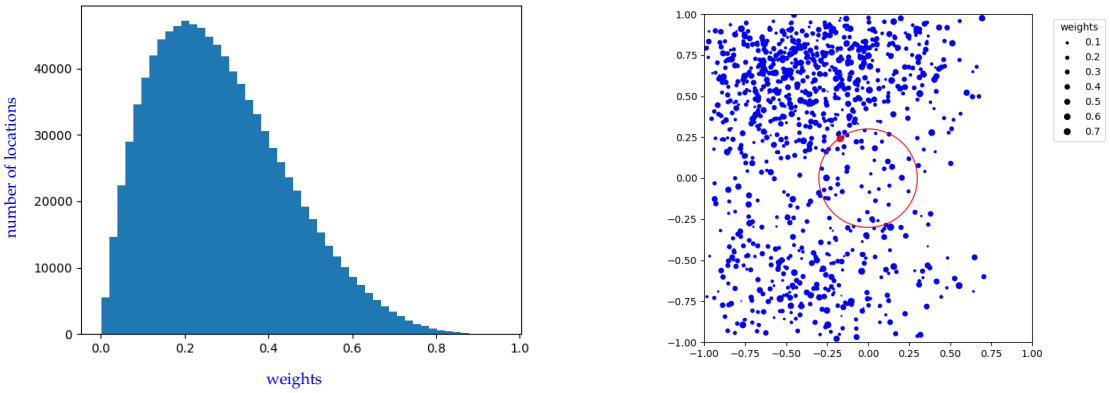

Figure 1: The left picture shows the histogram for the number of locations for the different weights. The right picture shows the position of 1000 locations randomly selected from the set of the m locations as blue dots, the service center, which is determined with Algorithm 3.16, as a red dot and the feasible set as the red circle.

We consider the following location problem, which is a special case of Problem 3.1: given m locations placed at points $c_i \in \mathbb{R}^2$, each of them weighted with a parameter $w_j > 0$, $i = 1, \ldots, m$, find a position $x \in \mathbb{R}^2$ for a service center so that the sum of the distances from it to the m locations is minimized under the restriction that the distance from the service center to the origin is less than or equal to a given radius $r > 0$. We can write this problem as

$$\min_{x \in S} \left\{ \sum_{i=1}^{m} f_i(x) \right\},$$

where $S = \{x \in \mathbb{R}^2 : \|x\| \leq r\}$ and $f_i(x) = w_i \|x - c_i\|$. We can write (see [43, Example 2.22])

$$f_i(x) = w_i \|x - c_i\| = \max_{y \in \mathbb{B}} \{\langle w_i x, y\rangle - \langle w_i c_i, y\rangle\},$$

where $\mathbb{B}$ is the closed unit ball of $\mathbb{R}$. By chosing $b_{\mathbb{B}}(y) = \frac{1}{2}\|y\|^2$ we approximate the functions f_i, using Nesterov's smoothing approach, for all $k \geq 0$ by

$$f_i^{\gamma_k}(x) = \max_{y \in \mathbb{B}} \left\{ \langle w_i x, y\rangle - \langle w_i c_i, y\rangle - \frac{\gamma_k}{2}\|y\|^2 \right\}$$

and according to [43, Corollary 2.20] (note that for our setting we have $A = w_i$, $b = w_i c_i$, $y_0 = 0$, $\mu = \gamma_k$ and $Q = \mathbb{B}$) with $b_{\mathbb{B}}(y) = \frac{1}{2}\|y\|^2$ we have

$$f_i^{\gamma_k}(x) = w_i^2 \frac{\|x - c_i\|^2}{2\gamma_k} - \frac{\gamma_k}{2} \left[d\left(\frac{w_i(x - c_i)}{\gamma_k}, \mathbb{B} \right) \right]^2,$$

where $d(x, \mathbb{B})$ is the Euclidean distance from x to $\mathbb{B}$. Then the gradients $\nabla f_i^{\gamma_k}$ can be written in terms of the projection operator $\mathcal{P}_{\mathbb{B}}$ on $\mathbb{B}$:

$$\nabla f_i^{\gamma_k} = w_i \mathcal{P}_{\mathbb{B}} \left(\gamma_k^{-1} w_i (x - c_i) \right).$$

In the following we apply Algorithm 3.10 for solving this problem. We choose $H(x) = \frac{1}{2}\|x\|^2$ for $x \in S$, and $H(x) = +\infty$ otherwise, so that we obtain for the mirror map the orthogonal projection onto the set S.

In our numerical experiments we choose $m = 1000000$, $r = 0.3$ and the m locations such that $c_i \in [-1,1] \times [-1,1]$ and the weights $w_i \in (0,1)$, $i = 1,\dots,m$, are beta randomly distributed. A histogram for the the number of locations for the different weights is presented in the left picture of Figure 1. In the right picture of Figure 1, the positions of the m locations are shown as blue dots. The greater the weight of the respective location, the larger the point. The red circle with radius r represents the feasible set for the position of the service center. The calculated position of the service center is shown as the red dot.

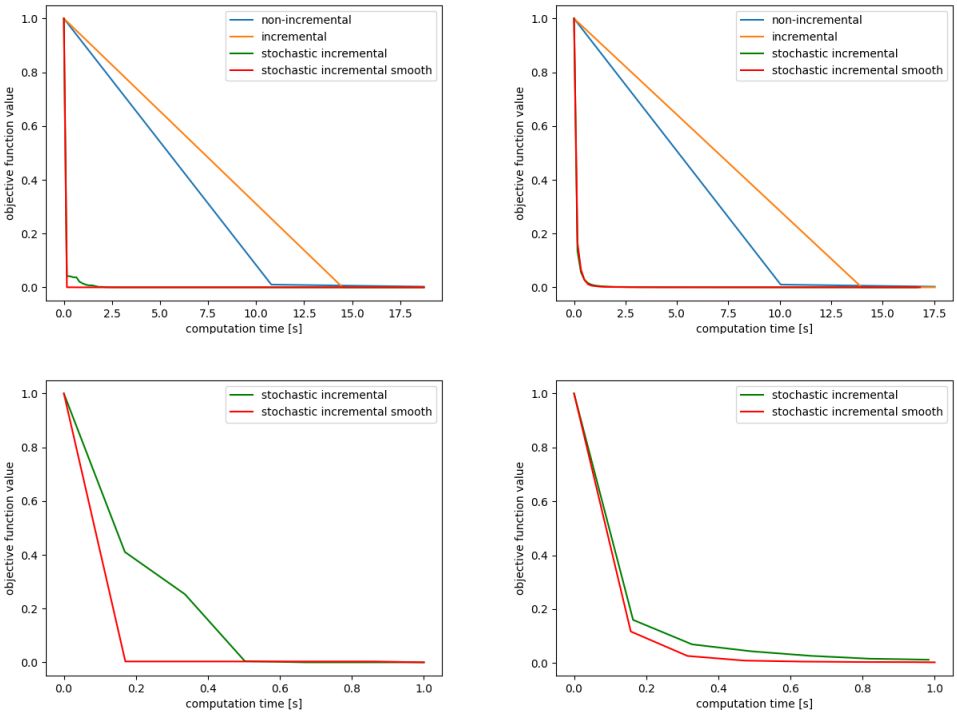

Figure 2: The plots show $(f_N - f(x_{\text{best}})/(f(x_0) - f(x_{\text{best}}))$, where $f_N := \min_{0\le k\le N} f(x_k)$, as a function of time, so x_k is the last iterate before a given point in time. In the first row we see the results after 17.5 seconds for one run left and 100 runs right. In the second row we see the results after 1 second for one run left and 100 runs right.

We compared our algorithm 3.10 (*stochastic incremental smooth*) to three versions of the algorithms described in [10]. The *stochastic incremental version* is the basic version of algorithm proposed in [10]. The *non-incremental version* takes a full subgradient step of the objective function $f(x)$ in each iteration instead of the single components $f_i(x)$, so basically it is a special case of the stochastic incremental version for $m = 1$ and $\epsilon_{1,k} = 1$ for every $k \ge 0$. The *incremental version* is the same as the stochastic incremental version, if we choose $\epsilon_{i,k} = 1$ for every $i = 1,\dots,m$, and every $k \ge 0$, so that we use the subgradient of all single components instead of a random choice. We choose $p_i = 0.000001$ for every $i = 1,\dots,m$, for the stochastic algorithms. In Figure 2 in the first row one can see the comparison of all four algorithms after one run in the left and 100 runs in the right respectively. There one can note that the stochastic algorithms clearly outperform their non-stochastic versions. In the second row we compared only the stochastic algorithms to have a better look after 1 second CPU time for one run and 100 runs. Here we can see that our algorithm performs slightly better than the stochastic incremental one.

6.2 Tomography

We consider the following optimization problem, which was proposed in [9] for reconstructing images in PET and and is in the setting of Problem 3.14,

$$\min_{x\in\Delta}\left\{-\sum_{i=1}^{m} y_i \log\left(\sum_{j=1}^{n} r_{ij}x_j\right)\right\},$$

where $\Delta := \{x \in \mathbb{R}^n : \sum_{j=1}^n x_j = 1, x \geq 0\}$ and $r_{ij} > 0$ is for $i = 1,\dots,m$, and $j = 1,\dots,n$, the entry of the i-th row and the j-th column of the matrix $R \in \mathbb{R}^{m\times n}$. Furthermore, y_i is for $i = 1,\dots,m$, the positive number of photons measured in the i-th bin. As mirror map we choose $H(x) = \sum_{i=1}^n x_i \log(x_i)$ for $x \in \Delta$ and $H(x) = +\infty$, otherwise.

The function $f_i(x) = -y_i \log\left(\sum_{j=1}^n r_{ij}x_j\right)$ is Lipschitz continuous for all $i = 1,\dots,m$, and so it follows that dom f_i^* is bounded, so we can apply Algorithm 3.16. The proximal point mapping of the function f_i can be deduced from [6, Lemma 6.5 and Theorem 6.15] and is given by

$$\mathrm{Prox}_{\gamma f_i}(v) = v + \frac{1}{\alpha}R_i\frac{\sqrt{\langle R_i, v\rangle^2 + 4\gamma\alpha y_i} - \langle R_i, v\rangle}{2},$$

where

$$R_i = (r_{i1},\dots,r_{in})^\top, \quad \alpha = \sum_{j=1}^{n} r_{ij}^2.$$

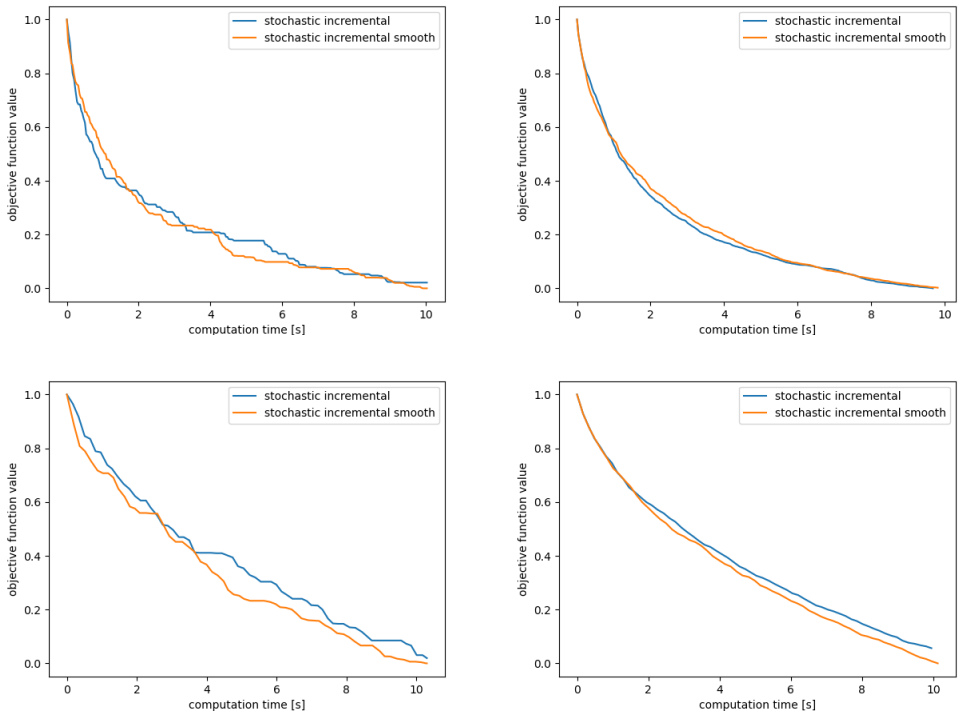

Figure 3: The plots show $(f_N - f(x_{\text{best}})/(f(x_0) - f(x_{\text{best}}))$, where $f_N := \min_{0\leq k\leq N} f(x_k)$, as a function of time, so x_k is the last iterate before a given point in time. In the first row we see the results for $n = 1000$ and $m = 6000$ for one run in the left plot and the average values of 10 runs in the right plot (with $p_i = 0.01667$, $i = 1,\dots,m$). In the second row we see the results for $n = 5000$ and $m = 15000$ for one run and the average values of 10 runs , respectively (with $p_i = 0,00667$, $i = 1,\dots,m$).

We can see in Figure 3 that both algorithms (our stochastic incremental smooth algorithm and the stochastic incremental version from [10]) have similar numerical performances, with the one proposed in this work reaching slightly lower objective function values.

6.3 SVM

In this subsection we consider an optimization problem of classifying images via support vector machines, which is a special case of Problem 4.7. The given data set for classification consists of 11339 training images and 1850 test images of size 28×28 and was taken from [47]. In the following optimization problem we search for a decision function x based on a pool of handwritten digits showing either the number 5 or the number 6, labeled by $+1$ and -1, respectively,

$$\min_{x \in \mathbb{R}^d} \left\{ \sum_{i=1}^{m} \max\{1 - Y_i \langle x, X_i \rangle, 0\} + \lambda \|x\|_1 \right\}, \tag{18}$$

$\{(X_1, Y_1), \ldots, (X_m, Y_m)\} \subseteq \mathbb{R}^d \times \{+1, -1\}$ is the given training data set with the training images X_i and the labels Y_i (here $d = 28 \cdot 28 = 784$). The 1-norm is a regularization term with the regularization parameter $\lambda > 0$. We set as in [10] $H = (1/2)\|\cdot\|^2$, so we obtain the identity as mirror map as this problem is unconstrained.

We can write the optimization problem as

$$\min_{x \in \mathbb{R}^d} \left\{ \sum_{i=1}^{m} f_i(x) + g(x) \right\},$$

where $f_i(x) = \max\{1 - Y_i \langle x, X_i \rangle, 0\}$ and $g(x) = \lambda \|x\|_1$. The function f_i is Lipschitz continuous for all $i = 1, \ldots, m$, and so it follows that $\operatorname{dom} f_i^*$ is bounded, so we can employ our algorithm. The proximal point mapping of the function f_i can be found in [13] and is given by

$$\operatorname{Prox}_{\gamma f_i}(v) = v + \begin{cases} \gamma Y_i X_i, & s_i \geq \gamma \|X_i\|^2 \\ 0, & s_i \leq 0 \\ \frac{Y_i s_i X_i}{\|X_i\|^2} & \text{otherwise,} \end{cases}$$

where $s_i = 1 - Y_i \langle v, X_i \rangle$, $i = 1, \ldots, m$.

Table 1: Numerical results for the SVM problem for stochastic incremental algorithm [10, Algorithm 4.2] (SI) and stochastic incremental smoothing algorithm (Algorithm 4.9) (SIS). The results are for one run and in the brackets over 10 runs for $p_i = 0,0082$, $i = 1, \ldots, m$.

regularization parameter	algorithm	decrease obj.function value	misclassified in %
$\lambda = 0.01$	SI	99.928 (99.923)	2.595 (2.595)
	SIS	99.929 (99.924)	2.324 (2.654)
$\lambda = 0.001$	SI	99.923 (99.927)	3.027 (2.605)
	SIS	99.922 (99.923)	2.432 (2.568)

The plots presented in Figure 4 show also for this application similar numerical performance of the employed algorithms (our stochastic incremental smooth algorithm and the stochastic incremental version from [10]), with a slightly better classification delivered by the method proposed in this work.

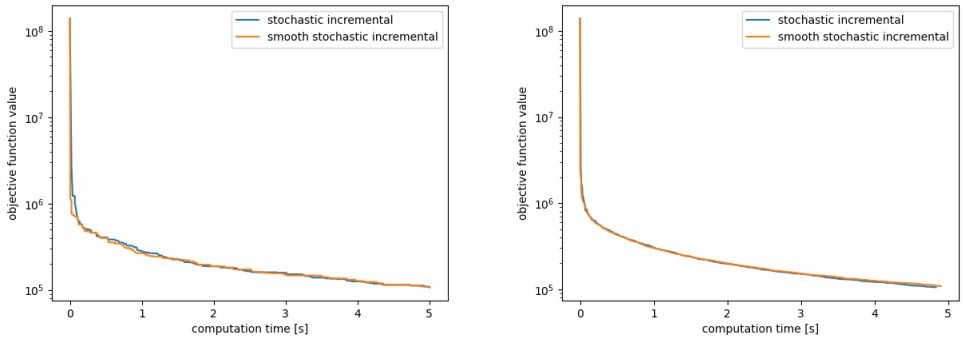

Figure 4: The plots show $f := \min_{0 \leq k \leq N} f(x_k)$ as a function of time, so x_k is the last iterate before a given point in time. We see the results for $\gamma = 0.001$ and $p_i = 0,0082$, $i = 1, \ldots, m$, for one run in the left plot and the average values of 10 experiments in the right plot.

7 Conclusions

In this paper we present two incremental stochastic mirror descent algorithms meant to minimize sums of finitely many nonsmooth convex functions over convex sets. Different to the similar approach from [10], we use the gradients of the smoothed summands of the objective function of the problem instead of their subgradients. For this we use the Nesterov smoothing technique, but since the Moreau-envelope is a special case of this smoothing technique, these algorithms can also be formulated with proximal steps, too. Moreover these algorithms can be modified for minimizing sums of finitely many compositions of convex functions with linear operators in similar contexts. We managed to obtain the same convergence order $\mathcal{O}(1/\sqrt{k})$ in expectation for the kth best objective function value and could show similar numerical performance as in [10], with slight improvements. Due to the fact that we do not need subgradients of the summands of the objective function, we have more variations of the proposed algorithms, so the most suitable smoothing method can be chosen depending on the structure of the considered optimization problem. If we use the Moreau-envelope we have uniquely defined proximal points, which have closed formulae for a variety of commonly used functions, instead of subgradients which one would have to pick from the subdifferentials of the involved functions at the given points and can sometimes be hard to determine. Moreover, the involved functions are not required to be (Lipschitz) continuous or differentiable, as they are usually taken in the literature on mirror descent methods. As subsequent developments we are interested in accelerating the proposed algorithms using for instance Nesterov's accelerated gradient or Polyak's heavy ball method. In [26] the autors combined Nesterov's accelerated gradient method and Nemirovski's mirror descent method both in continuous and discrete time, but for smooth convex functions. We are aim to provide a similar approach for nonsmooth functions using Nesterov smoothing. Furthermore it might be interesting to modify our algorithms in order to solve optimization problems like (9) where the prox-friendly proper, convex and lower semicontinuous function is composed with a linear operator or consists in a sum of such functions.

Acknowledgements. The work of the first named author was supported by the German Research Foundation (DFG), project WA922/9-1. The work of the second named author was partially supported by the Austrian Science Fund (FWF), project M-2045, by the Hi! PARIS Center, and by a public grant as part of the Investissement d'avenir project,

reference ANR-11-LABX-0056-LMH, LabEx LMH. For useful discussions regarding this paper the authors are thankful to Radu Ioan Boţ and Axel Böhm, to whom we are also grateful for providing us the program codes for their numerical experiments from [10], and to Gert Wanka. Valuable comments and suggestions from two anonymous reviewers are gratefully acknowledged, too.

8 Statements and declarations

There are no competing interests to declare.

Data availability statement The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

[1] M. Ahookhosh: *Optimal subgradient methods: computational properties for large-scale linear inverse problems*, Optim Eng 19:815–844 (2018)

[2] Z. Allen-Zhu, L. Orecchia: *Linear coupling: An ultimate unification of gradient and mirror descent*, in: C.H. Papadimitrou (ed.), Innovations in Theoretical Computer Science (ITCS 2017), Leibniz Int Pr Infor 8, Art. No. 3, 3:1–3:22 (2017)

[3] M. Amini, F. Yousefian: *An iterative regularized mirror descent method for ill-posed nondifferetiable stochastic optimization*, arXiv:1901.09506 (2019)

[4] N. Azizan, B. Hassibi: *A characterization of stochastic mirror descent algorithms and their convergence properties*, in: Int Conf Acoust Spee (ICASSP-2019), 5167–5171 (2019)

[5] F. Bach: *Duality between subgradient and conditional gradient methods*, SIAM J Optim 25:115–129 (2015)

[6] A. Beck: *First Order Methods in Optimization*, SIAM, Philadelphia (2017)

[7] A. Beck, M. Teboulle: *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Oper Res Lett 31:167–175 (2003)

[8] A. Beck, M. Teboulle: *Smoothing and first order methods: a unified framework*, SIAM J Optim 22:557–580 (2012)

[9] A. Ben-Tal, T. Margalit, A. Nemirovski:*The ordered subsets mirror descent optimization method with applications to tomography*, SIAM J Optim 12:79–108 (2001)

[10] R.I. Boţ, A. Böhm: *An incremental mirror descent subgradient algorithm with random sweeping and proximal step*, Optimization 68:1–18 (2018)

[11] R.I. Boţ, A. Böhm: *Variable smoothing for convex optimization problems using stochastic gradients*, J Sci Comput 85:33 (2020)

[12] J.-P. Calliess: *Lipschitz optimisation for Lipschitz interpolation*, in: 2017 American Control Conference (ACC2017), 17000349 (2017)

[13] A. Defazio: *A simple practical accelerated method for finite sums*, in: D.D. Lee, U. von Luxburg, R. Garnett, M. Sugiyama and I.M.Guyon (eds.), Adv Neur In 29 (NIPS 2016)

[14] T.T. Doan, S. Bose, D.H. Nguyen, C.L. Beck: *Convergence of the iterates in mirror descent methods*, IEEE Contr Syst Lett 3:114–119 (2019)

[15] J.C. Duchi, A. Agarwal, M. Johansson, M.I. Jordan: *Ergodic mirror descent*, SIAM J Optim 22:1549–1578 (2012)

[16] R.M. Freund, P. Grigas, R. Mazumder: *AdaBoost and forward stagewise regression are first-order convex optimization methods*, arXiv:1307.1192 (2013)

[17] G. Goh: *Optimization with Costly Subgradients*, ProQuest Dissertations Publishing, 2017.10685037 (2017)

[18] S.-M. Grad, O. Wilfer: *A proximal method for solving nonlinear minmax location problems with perturbed minimal time functions via conjugate duality*, J Glob Optim 74:121–160 (2019)

[19] V. Guigues: *Inexact stochastic mirror descent for two-stage nonlinear stochastic programs*, Math Program 187:533–577 (2021)

[20] F. Hanzely, P. Richtárik: *Fastest rates for stochastic mirror descent methods*, Comput Optim Appl 79:717–766 (2021)

[21] L.T.K. Hien, N. Gillis, P. Patrinos: *Inertial block mirror descent method for non-convex non-smooth optimization*, arXiv:1903.01818 (2019)

[22] L.T.K. Hien, C.V. Nguyen, H. Xu, C. Lu, J. Feng: *Accelerated randomized mirror descent algorithms for composite non-strongly convex optimization*, J Optim Theory Appl 181:541–566 (2019)

[23] V. Hovhannisyan, P. Parpas, S. Zafeiriou: *MAGMA - multilevel accelerated gradient mirror descent algorithm for large-scale convex composite minimization*, SIAM J Imaging Sci 9:1829–1857 (2016)

[24] A. Ivanova, F. Stonyakin, D. Pasechnyuk, E. Vorontsova, A. Gasnikov: *Adaptive mirror descent for the network utility maximization problem*, IFAC-PapersOnLine 53:7851–7856 (2020)

[25] A. Juditsky, J. Kwon, É. Moulines: *Unifying mirror descent and dual averaging*, Math Program, DOI:10.1007/s10107-022-01850-3

[26] W. Krichene, A. M. Bayen, P. L. Bartlett: *Accelerated mirror descent in continuous and discrete time*, Adv Neur In 2 (NIPS 2015), 2845–2853 (2015)

[27] G. Kunapuli, J. Shavlik: *Mirror descent for metric learning: a unified approach*, in: P.A. Flach, T. De Bie and N. Cristianini (eds.), Machine Learning and Knowledge Discovery in Databases - ECML PKDD 2012, Lect Notes Artif Int 7523, 859-874 (2012)

[28] Y.-H. Li, C.A. Riofrío, V. Cevher: *A general convergence result for mirror descent with Armijo line search*, arXiv:1805.12232 (2018)

[29] H. Lu: *Relative continuity for non-Lipschitz nonsmooth convex optimization using stochastic (or deterministic) mirror descent*, INFORMS J Optim 4:288–303 (2019)

[30] D.V.N. Luong, P. Parpas, D. Rueckert, B. Rustem: *Solving MRF minimization by mirror descent*, in: G. Bebis et al. (eds.), Advances in Visual Computing (ISVC 2012), Lect Notes Comput Sci 7431, Springer, 587–598 (2012)

[31] S. Mahadevan, B. Liu, P. Thomas, W. Dabney, S. Giguere, N. Jacek, I. Gemp, J. Liu: *Proximal reinforcement learning: a new theory of sequential decision making in primal-dual spaces*, arXiv:1405.6757 (2014)

[32] H.B. McMahan: *A unified view of regularized dual averaging and mirror descent with implicit updates*, arXiv:1009.3240v2 (2011)

[33] P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, G. Piliouras: *Optimistic mirror descent in saddle-point problems - going the extra (gradient) mile*, International Conference on Learning Representations (ICLR 2019), 1–23 (2019)

[34] P. Mertikopoulos, M. Staudigl: *Stochastic mirror descent dynamics and their convergence in monotone variational inequalities*, J Optim Theory Appl 179:838–867 (2018)

[35] K. Mishchenko: *Sinkhorn algorithm as a special case of stochastic mirror descent*, arXiv:1909.06918 (2019)

[36] A.V. Nazin, S. Anulova, A. Tremba: *Application of the mirror descent method to minimize average losses coming by a Poisson flow*, in: Proceedings of the European Control Conference (ECC14), 2194–2197 (2014)

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

[38] A. Nemirovski: *Efficient methods for large-scale convex optimization problems*, Ékon Mat Metody 2:135–152 (1979) (in Russian)

[39] A. Nemirovski, D.B. Yudin: *Problem Complexity and Method Efficiency in Optimization*, J. Wiley & Sons, New York (1983)

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

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

[42] Y. Nesterov: *Lectures on Convex Optimization*, Springer (2018)

[43] M.N. Nguyen, T.H.A. Le, D. Giles, T.A. Nguyen: *Smoothing techniques and difference of convex functions algorithms for image reconstruction*, Optimization 69(7–8):1601–1633 (2019)

[44] R. Paulavičius, J Žilinskas: *Analysis of different norms and corresponding Lipschitz constants for global optimization*, Inf Technol Control 12:301-306 (2006)

[45] T.-D. Quoc: *Adaptive smoothing algorithms for nonsmooth composite convex minimization*, Comput Optim Appl 66:425–451 (2017)

[46] R.T. Rockafellar: *Convex Analysis*, Princeton University Press (1970)

[47] S. Roweis: *Data for* MATLAB *hackers*, http://www.cs.nyu.edu/~roweis/data.html

[48] F. Schäfer, A. Anandkumar, H. Owhadi: *Competitive mirror descent*, arXiv:2006.10179 (2020)

[49] V.V. Semenov: *A version of the mirror descent method to solve variational inequalities*, Cybern Syst Anal 53:234–243 (2017)

[50] A. Titov, F. Stonyakin, M. Alkousa, S. Ablaev, A. Gasnikov: *Analogues of switching subgradient schemes for relatively Lipschitz-continuous convex programming problems*, in: MOTOR 2020, 133–149 (2020)

[51] S. Zhang, N. He: *On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization*, arXiv:1806.04781 (2018)

[52] X. Zhou, C. Du, X. Cai: *An efficient smoothing proximal gradient algorithm for convex clustering*, arXiv:2006.12592 (2020)

[53] Y. Zhou, Y. Liang, L. Shen: *A unified approach to proximal algorithms using Bregman distance*, Technical Report, Syracuse University (2016)

[54] Z. Zhou, P. Mertikopoulos, N. Bambos, S.P. Boyd, P.W. Glynn: *On the convergence of mirror descent beyond stochastic convex programming*, SIAM J Optim 30:687–716 (2020)

[55] J. Zimmert, T. Lattimore: *Connections between mirror descent, Thompson sampling and the information ratio*, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett (eds.), Adv Neur In 32 (NIPS 2019), 11973–11982 (2019)