An incremental mirror descent subgradient algorithm with random sweeping and proximal step

Radu Ioan Boţ[*] Axel Böhm[†]

April 26, 2018

Abstract

We investigate the convergence properties of incremental mirror descent type subgradient algorithms for minimizing the sum of convex functions. In each step we only evaluate the subgradient of a single component function and *mirror* it back to the feasible domain, which makes iterations very cheap to compute. The analysis is made for a randomized selection of the component functions, which yields the deterministic algorithm as a special case. Under supplementary differentiability assumptions on the function which induces the mirror map we are also able to deal with the presence of another term in the objective function, which is evaluated via a proximal type step. In both cases we derive convergence rates of $\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ in expectation for the kth best objective function value and illustrate our theoretical findings by numerical experiments in positron emission tomography and machine learning.

Keywords. nonsmooth convex minimization; incremental mirror descent algorithm; global rate of convergence; random sweeping

AMS subject classification. 90C25, 90C90, 90C06

1 Introduction

We consider the problem of minimizing the sum of nonsmooth convex functions

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

where $C \subseteq \mathbb{R}^n$ is a nonempty, convex and closed set and, for every $i = 1, ..., m$, the so-called *component functions* $f_i : \mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ are assumed to be proper and convex and will be evaluated via their respective subgradients. Implicitly we will assume that m is *large* and it is therefore very costly to evaluate all component functions in each iteration. Consequently, we will examine algorithms which only use the subgradient of a single component function in each iteration. These so-called *incremental algorithms*, see [4, 9], have been applied for large-scale problems arising in tomography [3], generalized assignment problems [9] or machine learning [14]. We refer also to [8] for a slightly different approach, where in the spirit of incremental algorithms only the gradient of one of the component functions is evaluated in each step, but gradients at old iterates are used for the other components. Both, subgradient algorithms and

[*] Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria, e-mail: radu.bot@univie.ac.at. Research partially supported by FWF (Austrian Science Fund), project I 2419-N32.

[†] Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria, e-mail: axel.boehm@univie.ac.at. Research supported by the doctoral programme *Vienna Graduate School on Computational Optimization (VGSCO)*, FWF (Austrian Science Fund), project W 1260.

incremental methods usually require decreasing stepsizes in order to converge, which makes them slow near an optimal solution. However, they provide in a very small number of iterations a low accuracy optimal value and possess a rate of convergence which is almost independent of the dimension of the problem. We refer the reader to [13] for a subgradient algorithm designed for the minimization of a nonsmooth nonconvex function under the making use of proximal subgradients.

When solving optimization problems of type (1) one might want to capture in the formulation of the iterative scheme the geometry of the feasible set C. This can be done by a so-called *mirror map*, that mirrors each iterate onto the feasible set. The Bregman distance associated with the function that induces the mirror map plays an essential role in the convergence analysis and in the formulation of convergence rates results (see [1,2]). So-called *mirror descent algorithms* were first discussed in [10] and more recently in [2,11,15] in a very general framework, in [12,14] from a statistical learning point of view, and in [5] for the case of dynamical systems. The mirror map can be viewed as a generalization of the ordinary orthogonal projection on C in Hilbert spaces (see Example 2.4), but allows also for a more careful consideration of the problems structure, as it is the case when the objective function is subdifferentiable only on the relative interior of the feasible set. In such a setting one can design a mirror map which maps not onto the entire feasible set but only on a subset of it where the objective function is subdifferentiable (see Example 2.5).

There exists already a rich literature on incremental algorithms dealing with similar problems. In [4,9] incremental subgradient methods with a random selection of the component functions and even projections onto a feasible set are considered, but no mirror descent. Incremental subgradient algorithms utilizing mirror descent techniques are investigated in [3], however there an additional projection onto the feasible set is required which thus excludes the case where $\text{dom} f \not\supseteq C$ (this is taken care of in our case by the weak assumption that $\text{im}(\nabla H^*) \subseteq \text{dom} f$). Furthermore, the results appearing in Section 4 discussing Bregman proximal steps appear to completely novel for this kind of problems and are only known from a forward-backward setting [1].

The basic concepts in the formulation of mirror descent algorithms are recalled in Section 2. We also provide some illustrating examples, which present some special cases, as the general setting might not be immediately intuitive. In Section 3 we formulate an incremental mirror descent subgradient algorithm with random sweeping of the component functions which we show to have a convergence rate of $\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ in expectation for the kth best objective function value. In Section 4 we ask additionally for differentiability of the function which induces the mirror map and are then able to add another nonsmooth convex function to the objective function which is evaluated in the iterative scheme by a proximal type step. For the resulting algorithm we show a similar convergence result. In the last section we illustrate the theoretical findings by numerical experiments in positron emission tomography and machine learning.

2 Elements of convex analysis and the mirror descent algorithm

Throughout the paper we assume that $\mathbb{R}^n$ is endowed with the Euclidean inner product $\langle \cdot, \cdot \rangle$ and corresponding norm $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. For a nonempty convex set $C \subseteq \mathbb{R}^n$ we denote by $\text{ri} C$ its *relative interior*, which is the interior of C relative to its affine hull. For a convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ we denote by $\text{dom} f := \{x \in \mathbb{R}^n : f(x) < +\infty\}$ its *effective domain* and say that f is *proper*, if $f > -\infty$ and $\text{dom} f \neq \emptyset$. The subdifferential of f at $x \in \mathbb{R}^n$ is defined as $\partial f(x) := \{p \in \mathbb{R}^n : f(y) \geq f(x) + \langle p, y - x \rangle \ \forall y \in \mathbb{R}^n\}$, for $f(x) \in \mathbb{R}$, and as $\partial f(x) := \emptyset$, otherwise. We will write $f'(x)$ for an arbitrary subgradient, i.e. an element of the subdifferential

$\partial f(x)$.

Problem 2.1. *Consider the optimization problem*

$$\min_{x\in C} f(x), \tag{2}$$

where $C \subseteq \mathbb{R}^n$ *is a nonempty, convex and closed set,* $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is a proper and convex function, and* $H : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is a proper, lower semicontinuous and* σ*-strongly convex function such that* $C = \overline{\mathrm{dom} H}$ *and* $\mathrm{im}(\nabla H^*) \subseteq \mathrm{int}(\mathrm{dom} f)$.

We say that $H : \mathbb{R}^n \to \overline{\mathbb{R}}$ is σ-strongly convex for $\sigma > 0$, if for every $x, x' \in \mathbb{R}^n$ and every $\lambda \in [0,1]$ it holds $\frac{\sigma}{2}\lambda(1-\lambda)\|x - x'\|^2 + H(\lambda x + (1-\lambda)x') \leq \lambda H(x) + (1-\lambda)H(x')$. It is well-known that, when H is proper, lower semicontinuous and σ-strongly convex, then its *conjugate function* $H^* : \mathbb{R}^n \to \overline{\mathbb{R}}, H^*(y) = \sup_{x\in\mathbb{R}^n}\{\langle y, x\rangle - H(x)\}$, is Fréchet differentiable (thus it has full domain) and its gradient ∇H^* is $\frac{1}{\sigma}$-Lipschitz continuous or, equivalently, H^* is Fréchet differentiable and its gradient ∇H^* is σ-cocoercive, which means that for every $y, y' \in \mathbb{R}^n$ it holds $\sigma\|\nabla H^*(y) - \nabla H^*(y')\|^2 \leq \langle y - y', \nabla H^*(y) - \nabla H^*(y')\rangle$. Recall that $\mathrm{im}(\nabla H^*) := \{\nabla H^*(y) : y \in \mathbb{R}^n\}$.

The following mirror descent algorithm has been introduced in [11] under the name *dual averaging.*

Algorithm 2.2. *Consider for some initial values* $x_0 \in \mathrm{int}(\mathrm{dom} f), y_0 \in \mathbb{R}^n$ *and sequence of positive stepsizes* $(t_k)_{k\geq 0}$ *the following iterative scheme:*

$$(\forall k \geq 0) \quad \left[\begin{array}{l} y_{k+1} = y_k - t_k f'(x_k) \\ x_{k+1} = \nabla H^*(y_{k+1}). \end{array}\right.$$

We notice that, since the sequence $(x_k)_{k\geq 0}$ is contained in the interior of the effective domain of f, the algorithm is well-defined. The assumptions concerning the function H, which induces the mirror map ∇H^*, are not consistent in the literature. Sometimes H is assumed to be a *Legendre function* as in [1], or strongly convex and differentiable as in [2,15]. In the following section we will only assume that H is proper, lower semicontinuous and strongly convex.

Example 2.3. *For* $H = \frac{1}{2}\|\cdot\|^2$ *we have that* $H^* = \frac{1}{2}\|\cdot\|^2$ *and thus* ∇H^* *is the identity operator on* $\mathbb{R}^n$. *Consequently, Algorithm 2.2 reduces to the classical subgradient method:*

$$(\forall k \geq 0) \quad x_{k+1} = x_k - t_k f'(x_k).$$

Example 2.4. *For* $C \subseteq \mathbb{R}^n$ *a nonempty, convex and closed set, take* $H(x) = \frac{1}{2}\|x\|^2$*, for* $x \in C$*, and* $H(x) = +\infty$*, otherwise. Then* $\nabla H^* = P_C$*, where* P_C *denotes the orthogonal projection onto* C*. In this setting, Algorithm 2.2 becomes:*

$$(\forall k \geq 0) \quad \left[\begin{array}{l} y_{k+1} = y_k - t_k f'(x_k) \\ x_{k+1} = P_C(y_{k+1}). \end{array}\right.$$

This iterative scheme is similar to, but different from the well-known subgradient projection algorithm, which reads:

$$(\forall k \geq 0) \quad \left[\begin{array}{l} y_{k+1} = x_k - t_k f'(x_k) \\ x_{k+1} = P_C(y_{k+1}). \end{array}\right.$$

Example 2.5. *When considering numerical experiments in positron emission tomography, one often minimizes over the unit simplex* $\Delta := \{x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n : \sum_{j=1}^n x_j = 1, x \geq 0\}$*. An* *appropriate choice for the function H is $H(x)=\sum_{j=1}^n x_j\log(x_j)$ for $x\in\Delta$, where $0\log(0)=0$, and $H(x)=+\infty$, if $x\notin\Delta$. In this case ∇H^* is given for every $y\in\mathbb{R}^n$ by*

$$\nabla H^*(y)=\frac{1}{\sum_{i=1}^n\exp(y_i)}\left(\exp(y_1),\exp(y_2),\dots,\exp(y_n)\right)^T,$$

and maps into the relative interior of Δ.

The following result will play an important role in the convergence analysis that we will carry out in the next sections.

Lemma 2.6. *Let $H:\mathbb{R}^n\to\overline{\mathbb{R}}$ be a proper, lower semicontinuous and σ-strongly convex function, for $\sigma>0$, $x\in\mathbb{R}^n$ and $y\in\partial H(x)$. Then it holds*

$$H(x)+\langle y,x'-x\rangle+\frac{\sigma}{2}\|x'-x\|^2\le H(x')\quad\forall x'\in\mathbb{R}^n.$$

Proof. The function $\widetilde{H}(\cdot):=H(\cdot)-\frac{\sigma}{2}\|\cdot\|^2$ is convex and $y-\sigma x\in\partial\widetilde{H}(x)$. Thus

$$\widetilde{H}(x)+\langle y-\sigma x,\tilde{x}-x\rangle\le\widetilde{H}(\tilde{x})\quad\forall\tilde{x}\in\mathbb{R}^n$$

or, equivalently,

$$H(x)-\frac{\sigma}{2}\|x\|^2+\langle y-\sigma x,\tilde{x}-x\rangle\le H(\tilde{x})-\frac{\sigma}{2}\|\tilde{x}\|^2\quad\forall\tilde{x}\in\mathbb{R}^n.$$

Rearranging the terms, leads to the desired conclusion. □

3 A stochastic incremental mirror descent algorithm

In this section we will address the following optimization problem.

Problem 3.1. *Consider the optimization problem*

$$\min_{x\in C}\sum_{i=1}^m f_i(x),\tag{3}$$

where $C\subseteq\mathbb{R}^n$ is a nonempty, convex and closed set, for every $i=1,...,m$, the functions $f_i:\mathbb{R}^n\mapsto\overline{\mathbb{R}}$ are proper and convex, and $H:\mathbb{R}^n\to\overline{\mathbb{R}}$ is a proper, lower semicontinuous and σ-strongly convex function such that $C=\overline{\mathrm{dom}H}$ and $\mathrm{im}(\nabla H^*)\subseteq\mathrm{int}\left(\cap_{i=1}^m\mathrm{dom}f_i\right)$.

In this section we apply the dual averaging approach described in Algorithm 2.2 to the optimization problem (3) by only using the subgradient of a component function at a time. This *incremental* approach (see, also, [4, 9]) is similar to but slightly different from the extension suggested in [2]. Furthermore, we introduce a stochastic sweeping of the component functions. For a similar strategy, but in the random selection of coordinates we refer to [6].

Algorithm 3.2. *Consider for some initial values $x_0\in\mathrm{int}\left(\cap_{i=1}^m\mathrm{dom}f_i\right),y_{m,-1}\in\mathbb{R}^n$ and sequence of positive stepsizes $(t_k)_{k\ge0}$ the following iterative scheme:*

```
(∀k ≥ 0)  ⎡ $\psi_{0,k}=x_k$
          ⎢ $y_{0,k}=y_{m,k-1}$
          ⎢ for $i=1,\dots,m$
          ⎢         $y_{i,k}=y_{i-1,k}-\epsilon_{i,k}\frac{t_k}{p_i}f_i'(\psi_{i-1,k})$
          ⎢         $\psi_{i,k}=\nabla H^*(y_{i,k})$
          ⎢ end
          ⎣ $x_{k+1}=\psi_{m,k}$,
```

where $\epsilon_{i,k}$ is a $\{0,1\}$ valued random variable for every $i=1,...,m$ and $k\ge0$, such that $\epsilon_{i,k}$ is independent of $\psi_{i-1,k}$ and $\mathbb{P}(\epsilon_{i,k}=1)=p_i$ for every $i=1,...,m$ and $k\ge0$.

One can notice that in the above iterative scheme $y_{i,k} \in \partial H(\psi_{i,k})$ for every $i = 1, ..., m$ and $k \geq 0$.

In the convergence analysis of Algorithm 3.2 we will make use of the following *Bregman-distance-like function* associated to the proper and convex function $H : \mathbb{R}^n \to \overline{\mathbb{R}}$ and defined as

$$d_H : \mathbb{R}^n \times \mathrm{dom} H \times \mathbb{R}^n \to \overline{\mathbb{R}},\ d_H(x, y, z) := H(x) - H(y) - \langle z, x - y\rangle. \tag{4}$$

We notice that $d_H(x, y, z) \geq 0$ for every $(x, y) \in \mathbb{R}^n \times \mathrm{dom} H$ and every $z \in \partial H(y)$, due to subgradient inequality.

The function d_H is an extension of the *Bregman distance* (see [1, 14, 15]), which is associated to a proper and convex function $H : \mathbb{R}^n \to \overline{\mathbb{R}}$ fulfilling $\mathrm{dom}\nabla H := \{x \in \mathbb{R}^n : H \text{ is differentiable at } x\} \neq \emptyset$ and defined as

$$D_H : \mathbb{R}^n \times \mathrm{dom}\nabla H \to \overline{\mathbb{R}},\ D_H(x, y) = H(x) - H(y) - \langle \nabla H(y), x - y\rangle. \tag{5}$$

Theorem 3.3. *In the setting of Problem 3.1, assume that the functions f_i are L_{f_i}-Lipschitz continuous on $\mathrm{im}(\nabla H^*)$ for $i = 1, ..., m$. Let $(x_k)_{k\geq 0}$ be a sequence generated by Algorithm 3.2. Then for every $N \geq 1$ and every $y \in \mathbb{R}^n$ it holds*

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

Proof. Let $y \in \cap_{i=1}^m \mathrm{dom} f_i \cap \mathrm{dom} H$ be fixed. For y outside this set the conclusion follows automatically.

For every $i = 1, ..., m$ and every $k \geq 0$ it holds

$$\begin{aligned} d_H(y, \psi_{i,k}, y_{i,k}) &= H(y) - H(\psi_{i,k}) - \langle y_{i,k}, y - \psi_{i,k}\rangle \\ &= H(y) - H(\psi_{i,k}) - \left\langle y_{i-1,k} - \frac{t_k}{p_i}\epsilon_{i,k} f_i'(\psi_{i-1,k}), y - \psi_{i,k}\right\rangle. \end{aligned}$$

Rearranging the terms, this yields for every $i = 1, ..., m$ and every $k \geq 0$ to

$$\begin{aligned} d_H(y, \psi_{i,k}, y_{i,k}) &= d_H(y, \psi_{i-1,k}, y_{i-1,k}) - d_H(\psi_{i,k}, \psi_{i-1,k}, y_{i-1,k}) + \frac{t_k}{p_i}\epsilon_{i,k}\langle f_i'(\psi_{i-1,k}), y - \psi_{i,k}\rangle \\ &= d_H(y, \psi_{i-1,k}, y_{i-1,k}) - d_H(\psi_{i,k}, \psi_{i-1,k}, y_{i-1,k}) + \frac{t_k}{p_i}\epsilon_{i,k}\langle f_i'(\psi_{i-1,k}), y - \psi_{i-1,k}\rangle \\ &\quad - \frac{t_k}{p_i}\epsilon_{i,k}\langle f_i'(\psi_{i-1,k}), \psi_{i,k} - \psi_{i-1,k}\rangle \\ &\leq d_H(y, \psi_{i-1,k}, y_{i-1,k}) - d_H(\psi_{i,k}, \psi_{i-1,k}, y_{i-1,k}) + \frac{t_k}{p_i}\epsilon_{i,k}(f_i(y) - f_i(\psi_{i-1,k})) \\ &\quad + \frac{t_k}{p_i}\epsilon_{i,k}\|f_i'(\psi_{i-1,k})\|\|\psi_{i-1,k} - \psi_{i,k}\|. \end{aligned}$$

From here we get for every $i = 1, ..., m$ and every $k \geq 0$

$$\begin{aligned} d_H(y, \psi_{i,k}, y_{i,k}) \leq{}& d_H(y, \psi_{i-1,k}, y_{i-1,k}) - d_H(\psi_{i,k}, \psi_{i-1,k}, y_{i-1,k}) + \frac{t_k}{p_i}\epsilon_{i,k}(f_i(y) - f_i(\psi_{i-1,k})) \\ &+ \frac{1}{\sigma} t_k^2 \frac{1}{p_i^2}\epsilon_{i,k}^2 \|f_i'(\psi_{i-1,k})\|^2 + \frac{\sigma}{4}\|\psi_{i-1,k} - \psi_{i,k}\|^2 \end{aligned}$$

which, by using that H is σ-strongly convex and Lemma 2.6, yields

$$\begin{aligned} d_H(y,\psi_{i,k},y_{i,k}) &\le d_H(y,\psi_{i-1,k},y_{i-1,k}) - d_H(\psi_{i,k},\psi_{i-1,k},y_{i-1,k}) + \frac{t_k}{p_i}\epsilon_{i,k}(f_i(y)-f_i(\psi_{i-1,k})) \\ &\quad + \frac{1}{\sigma}t_k^2\frac{1}{p_i^2}\epsilon_{i,k}\|f_i'(\psi_{i-1,k})\|^2 + \frac{1}{2}d_H(\psi_{i,k},\psi_{i-1,k},y_{i-1,k}) \\ &= d_H(y,\psi_{i-1,k},y_{i-1,k}) + \frac{t_k}{p_i}\epsilon_{i,k}(f_i(y)-f_i(\psi_{i-1,k})) + \frac{1}{\sigma}t_k^2\frac{1}{p_i^2}\epsilon_{i,k}\|f_i'(\psi_{i-1,k})\|^2 \\ &\quad - \frac{1}{2}d_H(\psi_{i,k},\psi_{i-1,k},y_{i-1,k}). \end{aligned}$$

Using the fact that f_i is L_{f_i}-Lipschitz continuous, it follows that $\|f_i'(\psi_{i-1,k})\| \le L_{f_i}$, for every $i=1,...,m$ and every $k \ge 0$, thus

$$\begin{aligned} d_H(y,\psi_{i,k},y_{i,k}) &\le d_H(y,\psi_{i-1,k},y_{i-1,k}) + \frac{t_k}{p_i}\epsilon_{i,k}(f_i(y)-f_i(\psi_{i-1,k})) + \frac{1}{\sigma}t_k^2\frac{1}{p_i^2}\epsilon_{i,k}L_{f_i}^2 \\ &\quad - \frac{1}{2}d_H(\psi_{i,k},\psi_{i-1,k},y_{i-1,k}). \end{aligned}$$

Since all the involved functions are measurable, we can take the expected value on both sides of the above inequality and get due to the assumed independence of $\epsilon_{i,k}$ and $\psi_{i-1,k}$ for every $i=1,...,m$ and every $k \ge 0$

$$\begin{aligned} \mathbb{E}\left(d_H(y,\psi_{i,k},y_{i,k})\right) &\le \mathbb{E}\left(d_H(y,\psi_{i-1,k},y_{i-1,k})\right) + \mathbb{E}\left(\frac{t_k}{p_i}(f_i(y)-f_i(\psi_{i-1,k}))\right)\mathbb{E}(\epsilon_{i,k}) \\ &\quad + \frac{1}{\sigma}t_k^2\frac{1}{p_i^2}L_{f_i}^2\mathbb{E}(\epsilon_{i,k}) - \mathbb{E}\left(\frac{1}{2}d_H(\psi_{i,k},\psi_{i-1,k},y_{i-1,k})\right). \end{aligned}$$

Since $\mathbb{E}(\epsilon_{i,k}) = p_i$, we get for every $i=1,...,m$ and every $k \ge 0$

$$\begin{aligned} \mathbb{E}\left(d_H(y,\psi_{i,k},y_{i,k})\right) &\le \mathbb{E}\left(d_H(y,\psi_{i-1,k},y_{i-1,k})\right) + \mathbb{E}\left(t_k(f_i(y)-f_i(\psi_{i-1,k}))\right) \\ &\quad + \frac{1}{\sigma}t_k^2\frac{1}{p_i}L_{f_i}^2 - \mathbb{E}\left(\frac{1}{2}d_H(\psi_{i,k},\psi_{i-1,k},y_{i-1,k})\right). \end{aligned}$$

Summing the above inequality for $i=1,...,m$ and using that

$$\sum_{i=1}^m L_{f_i}^2\frac{1}{p_i} \le \left(\sum_{i=1}^m L_{f_i}^4\right)^{\frac{1}{2}}\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} \le \left(\sum_{i=1}^m L_{f_i}\right)^2\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}},$$

it yields for every $k \ge 0$

$$\begin{aligned} \mathbb{E}(d_H(y,\psi_{m,k},y_{m,k})) &\le \mathbb{E}(d_H(y,x_k,y_{0,k})) + \mathbb{E}\left(t_k\sum_{i=1}^m(f_i(y)-f_i(\psi_{i-1,k}))\right) \\ &\quad + \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m L_{f_i}\right)^2\left(\sum_{i=1}^m\frac{1}{p_i^2}\right)^{\frac{1}{2}} - \mathbb{E}\left(\sum_{i=1}^m\frac{1}{2}d_H(\psi_{i,k},\psi_{i-1,k},y_{i-1,k})\right) \end{aligned}$$

or, equivalently,

$$\begin{aligned} \mathbb{E}(d_H(y,\psi_{m,k},y_{m,k})) &\le \mathbb{E}(d_H(y,x_k,y_{0,k})) + \mathbb{E}\left(t_k\sum_{i=1}^m(f_i(y)-f_i(x_k)+f_i(x_k)-f_i(\psi_{i-1,k}))\right) \\ &\quad + \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m L_{f_i}\right)^2\left(\sum_{i=1}^m\frac{1}{p_i^2}\right)^{\frac{1}{2}} - \mathbb{E}\left(\sum_{i=1}^m\frac{1}{2}d_H(\psi_{i,k},\psi_{i-1,k},y_{i-1,k})\right). \end{aligned}$$

Thus, for every $k \geq 0$,

$$\begin{aligned}\mathbb{E}(d_H(y,\psi_{m,k},y_{m,k})) \leq{}& \mathbb{E}(d_H(y,x_k,y_{0,k})) + t_k\mathbb{E}\left(\sum_{i=1}^m f_i(y) - \sum_{i=1}^m f_i(x_k)\right)\\ &+ \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m L_{f_i}\right)^2\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} - \mathbb{E}\left(\sum_{i=1}^m \frac{1}{2}d_H(\psi_{i,k},\psi_{i-1,k},y_{i-1,k})\right)\\ &+ \mathbb{E}\left(t_k\sum_{i=1}^m(f_i(x_k) - f_i(\psi_{i-1,k}))\right). \qquad (6)\end{aligned}$$

On the other hand, by using the Lipschitz continuity of ∇H^* it yields for every $k \geq 0$

$$\begin{aligned}\sum_{i=1}^m (f_i(x_k) - f_i(\psi_{i-1,k})) &= \sum_{i=2}^m\sum_{j=1}^{i-1}(f_i(\psi_{j-1,k}) - f_i(\psi_{j,k}))\\ &\leq \sum_{i=2}^m\sum_{j=1}^{i-1} L_{f_i}\|\psi_{j-1,k} - \psi_{j,k}\| \leq \left(\sum_{l=1}^m L_{f_l}\right)\sum_{i=2}^m\|\psi_{i-1,k} - \psi_{i,k}\|,\\ &\leq \left(\sum_{l=1}^m L_{f_l}\right)\sum_{i=2}^m\|\nabla H^*(y_{i-1,k}) - \nabla H^*(y_{i,k})\|\\ &\leq \frac{1}{\sigma}\left(\sum_{l=1}^m L_{f_l}\right)\sum_{i=2}^m\|y_{i-1,k} - y_{i,k}\|\\ &= \frac{1}{\sigma}\left(\sum_{l=1}^m L_{f_l}\right)\sum_{i=2}^m\left\|\epsilon_{i,k}\frac{t_k}{p_i}f_i'(\psi_{i-1,k})\right\|\\ &\leq \frac{1}{\sigma}t_k\left(\sum_{l=1}^m L_{f_l}\right)\left(\sum_{i=1}^m \frac{\epsilon_{i,k}}{p_i}L_{f_i}\right).\end{aligned}$$

Therefore, for every $k \geq 0$

$$\begin{aligned}\mathbb{E}\left(t_k\sum_{i=1}^m(f_i(x_k) - f_i(\psi_{i-1,k}))\right) &\leq \frac{1}{\sigma}t_k^2\left(\sum_{l=1}^m L_{f_l}\right)\mathbb{E}\left(\sum_{i=1}^m \frac{\epsilon_{i,k}}{p_i}L_{f_i}\right)\\ &\leq \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m L_{f_i}\right)^2. \qquad (7)\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}(d_H(y,\psi_{m,k},y_{m,k})) \leq{}& \mathbb{E}(d_H(y,x_k,y_{0,k})) + t_k\mathbb{E}\left(\sum_{i=1}^m f_i(y) - \sum_{i=1}^m f_i(x_k)\right)\\ &+ \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m L_{f_i}\right)^2\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} - \mathbb{E}\left(\sum_{i=1}^m \frac{1}{2}d_H(\psi_{i,k},\psi_{i-1,k},y_{i-1,k})\right)\\ &+ \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m L_{f_i}\right)^2. \qquad (8)\end{aligned}$$

Since $\psi_{m,k} = x_{k+1}$ and $y_{m,k} = y_{0,k+1}$ we get for every $k \geq 0$ that

$$\mathbb{E}(d_H(y, x_{k+1}, y_{0,k+1})) \leq \mathbb{E}(d_H(y, x_k, y_{0,k})) + t_k \mathbb{E}\left(\sum_{i=1}^m f_i(y) - \sum_{i=1}^m f_i(x_k)\right)$$
$$+ \frac{1}{\sigma} t_k^2 \left(\sum_{i=1}^m L_{f_i}\right)^2 \left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^2 + 1\right).$$

By summing up this inequality from $k = 0$ to $N-1$, where $N \geq 1$, we get

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

Since $d_H(y, x_N, y_{0,N}) \geq 0$, as $y_{0,N} \in \partial H(x_N)$, we get

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

and this finishes the proof. □

Remark 3.4. The set from which the variable y is chosen in the previous theorem might seems to be restrictive, however, we would like to recall that in many applications $\mathrm{dom} H$ is the set of feasible solutions of the optimization problem (3). Since $\mathrm{im}(\nabla H^*) = \mathrm{dom}\partial H := \{x \in \mathbb{R}^n : \partial H(x) \neq \emptyset\} \subseteq \mathrm{dom} H$, the inequality in Theorem 3.3 is fulfilled for every $y \in \mathrm{im}(\nabla H^*)$.

Remark 3.5. Note furthermore that so far we have not made any assumptions about the stepsizes in Theorem 3.3. It is however clear from the statement that in the case where $y = x^*$ for an optimal solution x^* and the stepsizes $(t_k)_{k \in \mathbb{N}}$ fulfill the classical condition that $\sum_{k=1}^\infty t_k = +\infty$ and $\sum_{k=1}^\infty t_k^2 < +\infty$ it follows that $\lim_{N \in \mathbb{N}} \mathbb{E}\left(\min_{0 \leq k \leq N-1} \sum_{i=1}^m f_i(x_k) - \sum_{i=1}^m f_i(x^*)\right) = 0$.

The optimal stepsize choice, which we provide in the following corollary, is a consequence of [2, Proposition 4.1], which states that the function

$$z \mapsto \frac{c + (2\sigma)^{-1} z^T D z}{b^T z},$$

where $c > 0, b \in \mathbb{R}^d_{++} := \{(z_1, ..., z_d)^T \in \mathbb{R}^d : z_i > 0, i = 1, ..., d\}$ and $D \in \mathbb{R}^{d \times d}$ is a symmetric positive definite matrix, attains its minimum on $\mathbb{R}^d_{++}$ at $z^* = \sqrt{\frac{2c\sigma}{b^T D^{-1} b}} D^{-1} b$ and this provides $\sqrt{\frac{2c}{\sigma b^T D^{-1} b}}$ as optimal objective value.

Corollary 3.6. *In the setting of Problem 3.1, assume that the functions f_i are L_{f_i}-Lipschitz continuous on $\mathrm{im}(\nabla H^*)$ for $i = 1, ..., m$. Let $x^* \in \mathrm{dom} H$ be an optimal solution of (3) and $(x_k)_{k \geq 0}$ be a sequence generated by Algorithm 3.2 with optimal stepsize*

$$t_k := \frac{1}{\sum_{i=1}^m L_{f_i}} \sqrt{\frac{d_H(x^*, x_0, y_{0,0})}{\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^2 + 1}} \frac{1}{\sqrt{k}} \ \forall k \geq 0.$$

Then for every $N \geq 1$ it holds

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

Remark 3.7. In the last step of the proof of Theorem 3.3 one could have chosen to use the following inequality

$$\left(\sum_{k=0}^{N-1}t_k\right)\mathbb{E}\left(\sum_{i=1}^m f_i\left(\frac{\sum_{k=0}^{N-1}t_kx_k}{\sum_{k=0}^{N-1}t_k}\right)-\sum_{i=1}^m f_i(y)\right)\leq\sum_{k=0}^{N-1}t_k\mathbb{E}\left(\sum_{i=1}^m f_i(x_k)-\sum_{i=1}^m f_i(y)\right)$$

given by the convexity of $\sum_{i=1}^m f_i(\cdot)$ in order to prove convergence of the function values for the ergodic sequence $\bar{x}_k := \frac{1}{\sum_{i=0}^k t_i}\sum_{i=0}^k t_ix_i$ for all $k \geq 0$. This would lead for every $N \geq 1$ and every $y \in \mathbb{R}^n$ to

$$\mathbb{E}\left(\sum_{i=1}^m f_i(\bar{x}_{N-1})-\sum_{i=1}^m f_i(y)\right)\leq\frac{d_H(y,x_0,y_{0,0})+\frac{1}{\sigma}\left(\sum_{i=1}^m L_{f_i}\right)^2\left(\left(\sum_{i=1}^m\frac{1}{p_i^2}\right)^2+1\right)\sum_{k=0}^{N-1}t_k^2}{\sum_{k=0}^{N-1}t_k}$$

and for the optimal stepsize choice from Corollary 3.6 to

$$\mathbb{E}\left(\sum_{i=1}^m f_i(\bar{x}_{N-1})-\sum_{i=1}^m f_i(y)\right)\leq 2\left(\sum_{i=1}^m L_{f_i}\right)\sqrt{\frac{d_H(x^*,x_0,y_{0,0})\left(\left(\sum_{i=1}^m\frac{1}{p_i^2}\right)^2+1\right)}{\sigma}}\frac{1}{\sqrt{N}},$$

and might be beneficial, as it does not require the computation of objective function values, which are by our implicit assumption of m being large expensive to compute.

4 A stochastic incremental mirror descent algorithm with Bregman proximal step

In this section we add another nonsmooth convex function to the objective function of the optimization problem (3) and provide an extension of Algorithm 3.2, which evaluates in particular the new summand by a proximal type step. However, this asks for supplementary differentiability assumption on the function inducing the mirror map.

Problem 4.1. *Consider the optimization problem*

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

where $C \subseteq \mathbb{R}^n$ is a nonempty, convex and closed set, for every $i = 1,...,m$, the functions $f_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ are proper and convex and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper, convex and lower semicontinuous function, and $H : \mathbb{R}^m \to \overline{\mathbb{R}}$ is a proper, σ-strongly convex and lower semicontinuous function such that $C = \overline{\mathrm{dom}H}$, H is continuously differentiable on $\mathrm{int}(\mathrm{dom}H)$*,* $\mathrm{im}(\nabla H^*) \subseteq \mathrm{int}\left(\cap_{i=1}^m \mathrm{dom}f_i\right) \cap \mathrm{int}(\mathrm{dom}H)$ *and* $\mathrm{int}(\mathrm{dom}H) \cap \mathrm{dom}g \neq \emptyset$*.*

For a proper, convex, lower semicontinuous function $h:\mathbb{R}^n \to \overline{\mathbb{R}}$ we define its *Bregman-proximal operator* with respect to the proper, σ-strongly convex and lower semicontinuous function $H:\mathbb{R}^n \to \overline{\mathbb{R}}$ as being

$$\operatorname{prox}_h^H : \operatorname{dom}\nabla H \to \mathbb{R}^n,\ \operatorname{prox}_h^H(x) := \operatorname*{arg\,min}_{u\in\mathbb{R}^n} \left\{h(u) + D_H(u,x)\right\}.$$

Due to the strong convexity of H, the Bregman-proximal operator is well-defined. For $H = \frac{1}{2}\|\cdot\|^2$ it coincides with the classical proximal operator.

We are now in the position to formulate the iterative scheme we would like to propose for solving (9). In case $g = 0$, this algorithm gives exactly the incremental version of the iterative method in [2], actually suggested by the two authors in this paper.

Algorithm 4.2. *Consider for some initial value $x_0 \in \operatorname{im}(\nabla H^*)$ and sequence of positive stepsizes $(t_k)_{k\geq 0}$ the following iterative scheme:*

$$(\forall k \geq 0) \quad \left[\begin{array}{l} \psi_{0,k} = x_k \\ \textit{for } i = 1,\ldots,m \\ \qquad\qquad \psi_{i,k} = \nabla H^*(\nabla H(\psi_{i-1,k}) - \epsilon_{i,k}\frac{t_k}{p_i} f_i'(\psi_{i-1,k})) \\ \textit{end} \\ x_{k+1} = \operatorname{prox}_{t_k g}^H(\psi_{m,k}), \end{array}\right.$$

where $\epsilon_{i,k}$ is a $\{0,1\}$ valued random variable for every $i = 1,...,m$ and $k \geq 0$, such that $\epsilon_{i,k}$ is independent from $\psi_{i-1,k}$ and $\mathbb{P}(\epsilon_{i,k} = 1) = p_i$ for every $i = 1,...,m$ and $k \geq 0$.

Lemma 4.3. *In the setting of Problem 4.1, Algorithm 4.2 is well-defined.*

Proof. As $\operatorname{im}(\nabla H^*) \subseteq \operatorname{int}\left(\cap_{i=1}^m \operatorname{dom} f_i\right)$, it follows for every $i = 2,...,m$ and every $k \geq 0$ immediately that $\psi_{i-1,k} \in \operatorname{int}\operatorname{dom} f_i$, thus a subgradient of f_i at $\psi_{i-1,k}$ exists.

In what follows we prove that this is the case also for $\psi_{0,k}$, for every $k \geq 0$. To this aim it is enough to show that $x_k \in \operatorname{im}(\nabla H^*)$ for every $k \geq 0$. For $k = 0$ this statement is true by the choice of the initial value. For every $k \geq 0$ we have that

$$0 \in \partial\left(t_k g + H - \langle \nabla H(\psi_{m,k}), \cdot\rangle\right)(x_{k+1}),$$

which, according to $\operatorname{int}(\operatorname{dom} H) \cap \operatorname{dom} g \neq \emptyset$, is equivalent to

$$0 \in t_k \partial g(x_{k+1}) + \partial H(x_{k+1}) - \nabla H(\psi_{m,k}).$$

Thus $x_{k+1} \in \operatorname{dom}\partial H = \operatorname{im}(\nabla H^*)$ for every $k \geq 0$ and this concludes the proof. □

Example 4.4. *Consider the case when $m = 1$, $\epsilon_{1,k} = 1$ for every $k \geq 0$ and $H(x) = \frac{1}{2}\|x\|^2$ for $x \in C$, while $H(x) = +\infty$ for $x \notin C$, where $C \subseteq \mathbb{R}^n$ is a nonempty, convex and closed set. In this setting, ∇H^* is equal to the orthogonal projection P_C onto the set C. Algorithm 4.2 yields the following iterative scheme, which basically minimizes the sum $f_1 + g$ over the set C:*

$$(\forall k \geq 0) \quad x_{k+1} = \operatorname{prox}_{t_k g}^H(P_C(x_k - t_k f_1'(x_k))). \tag{10}$$

The difficulty in Example 4.4, assuming that it is reasonably possible to project onto the set C, lies in evaluating $\operatorname{prox}_{t_k g}^H$, for every $k \geq 0$, as this itself is a constraint optimization problem

$$\operatorname{prox}_{t_k g}^H(x) = \operatorname*{arg\,min}_{u\in C} \left\{ t_k g(u) + \frac{1}{2}\|x-u\|^2 \right\}.$$

When $C = \mathbb{R}^n$, the iterative scheme (10) becomes the proximal subgradient algorithm investigated in [7].

Theorem 4.5. *In the setting of Problem 4.1, assume that the functions f_i are L_{f_i}-Lipschitz continuous on* $\operatorname{im}(\nabla H^*)$ *for $i = 1, ..., m$. Let $(x_k)_{k\geq 0}$ be a sequence generated by Algorithm 4.2. Then for every $N \geq 1$ and every $y \in \mathbb{R}^n$ it holds*

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

$$\frac{2\sigma D_H(y,x_0) + \left(2\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 3 + 2m\right)\left(\sum_{i=1}^m L_{f_i}\right)^2 \sum_{k=0}^{N-1} t_k^2}{2\sigma \sum_{k=0}^{N-1} t_k}.$$

Proof. Let $y \in \cap_{i=1}^m \operatorname{dom} f_i \cap \operatorname{dom} g \cap \operatorname{dom} H$ be fixed. For y outside this set the conclusion follows automatically.

As in the first part of the proof of Theorem 3.3, we obtain instead of (8) the following inequality which holds for every $i = 1, ..., m$ and every $k \geq 0$

$$\mathbb{E}\left(D_H(y,\psi_{m,k})\right) \leq \mathbb{E}(D_H(y,x_k)) + t_k\mathbb{E}\left(\sum_{i=1}^m f_i(y) - \sum_{i=1}^m f_i(x_k)\right)$$
$$+ \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m L_{f_i}\right)^2\left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right) - \mathbb{E}\left(\sum_{i=1}^m \frac{1}{2}D_H(\psi_{i,k},\psi_{i-1,k})\right). \quad (11)$$

As pointed out in the proof of Lemma 4.3, for every $k \geq 0$ we have

$$0 \in t_k\partial g(x_{k+1}) + \nabla H(x_{k+1}) - \nabla H(\psi_{m,k}),$$

thus

$$t_k(g(y) - g(x_{k+1})) \geq \langle \nabla H(\psi_{m,k}) - \nabla H(x_{k+1}), y - x_{k+1}\rangle .$$

The three point identity leads to

$$t_k(g(y) - g(x_{k+1})) \geq -(D_H(y,\psi_{m,k}) - D_H(y,x_{k+1}) - D_H(x_{k+1},\psi_{m,k}))$$

or, equivalently,

$$t_k(g(x_{k+1}) - g(y)) + D_H(y,x_{k+1}) \leq D_H(y,\psi_{m,k}) - D_H(x_{k+1},\psi_{m,k})$$

for every $k \geq 0$. Since the involved functions are measurable, we can take the expected value on both sides and obtain for every $k \geq 0$

$$t_k\mathbb{E}((g(x_{k+1}) - g(y))) + \mathbb{E}(D_H(y,x_{k+1})) \leq \mathbb{E}(D_H(y,\psi_{m,k})) - \mathbb{E}(D_H(x_{k+1},\psi_{m,k})). \quad (12)$$

Combining (11) and (12) gives for every $k \geq 0$

$$t_k\mathbb{E}((g(x_{k+1}) - g(y))) + t_k\mathbb{E}\left(\sum_{i=1}^m f_i(x_k) - \sum_{i=1}^m f_i(y)\right) + \mathbb{E}(D_H(y,x_{k+1})) \leq$$
$$\mathbb{E}(D_H(y,x_k)) + \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m L_{f_i}\right)^2\left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)$$
$$- \mathbb{E}(D_H(x_{k+1},\psi_{m,k})) - \sum_{i=1}^m \frac{1}{2}\mathbb{E}\left(D_H(\psi_{i,k},\psi_{i-1,k})\right).$$

By adding and subtracting $\mathbb{E}\left(\sum_{i=1}^m f_i(x_{k+1})\right)$ and by using afterwards the Lipschitz continuity of $\sum_{i=1}^m f_i$, we get for every $k \geq 0$

$$\begin{aligned}
&t_k\mathbb{E}\left(\left(\sum_{i=1}^m f_i + g\right)(x_{k+1}) - \left(\sum_{i=1}^m f_i + g\right)(y)\right)\\
&-t_k\left(\sum_{i=1}^m L_{f_i}\right)\mathbb{E}(\|x_k - x_{k+1}\|) + \mathbb{E}(D_H(y, x_{k+1})) \leq\\
&\mathbb{E}\left(D_H(y, x_k)\right) + \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m L_{f_i}\right)^2\left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)\\
&-\mathbb{E}\left(D_H(x_{k+1}, \psi_{m,k})\right) - \sum_{i=1}^m \frac{1}{2}\mathbb{E}\left(D_H(\psi_{i,k}, \psi_{i-1,k})\right).
\end{aligned}$$

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

$$\begin{aligned}
&t_k\mathbb{E}\left(\left(\sum_{i=1}^m f_i + g\right)(x_{k+1}) - \left(\sum_{i=1}^m f_i + g\right)(y)\right) + \mathbb{E}(D_H(y, x_{k+1})) \leq\\
&\mathbb{E}(D_H(y, x_k)) + \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m L_{f_i}\right)^2\left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right) - \mathbb{E}(D_H(x_{k+1}, \psi_{m,k}))\\
&+ t_k\left(\sum_{i=1}^m L_{f_i}\right)\mathbb{E}(\|x_k - \psi_{m,k}\|) + t_k\left(\sum_{i=1}^m L_{f_i}\right)\mathbb{E}(\|\psi_{m,k} - x_{k+1}\|) - \sum_{i=1}^m \frac{1}{2}\mathbb{E}\left(D_H(\psi_{i,k}, \psi_{i-1,k})\right),
\end{aligned}$$

which, due to Young's inequality and the strong convexity of H, leads to

$$\begin{aligned}
&t_k\mathbb{E}\left(\left(\sum_{i=1}^m f_i + g\right)(x_{k+1}) - \left(\sum_{i=1}^m f_i + g\right)(y)\right) + \mathbb{E}(D_H(y, x_{k+1})) \leq\\
&\mathbb{E}(D_H(y, x_k)) + \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m L_{f_i}\right)^2\left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right) - \mathbb{E}(D_H(x_{k+1}, \psi_{m,k}))\\
&+ t_k\left(\sum_{i=1}^m L_{f_i}\right)\mathbb{E}(\|x_k - \psi_{m,k}\|) + \frac{1}{2\sigma}t_k^2\left(\sum_{i=1}^m L_{f_i}\right)^2\\
&+ \mathbb{E}(D_H(x_{k+1}, \psi_{m,k})) - \sum_{i=1}^m \frac{1}{2}\mathbb{E}\left(D_H(\psi_{i,k}, \psi_{i-1,k})\right).
\end{aligned}$$

Since

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

we get for every $k \geq 0$ that

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

$$\mathbb{E}(D_H(y, x_k)) + \frac{1}{2\sigma} t_k^2 \left(\sum_{i=1}^m L_{f_i}\right)^2 \left(2\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 3\right)$$

$$+ t_k \left(\sum_{i=1}^m L_{f_i}\right) \mathbb{E}\left(\sum_{i=1}^m \|\psi_{i-1,k} - \psi_{i,k}\|\right) - \sum_{i=1}^m \frac{1}{2} \mathbb{E}\left(D_H(\psi_{i,k}, \psi_{i-1,k})\right).$$

Young's inequality and the strong convexity of H imply that for every $i = 1, ..., m$ and every $k \geq 0$

$$\begin{aligned} t_k \left(\sum_{i=1}^m L_{f_i}\right) \|\psi_{i-1,k} - \psi_{i,k}\| &\leq \frac{1}{\sigma} t_k^2 \left(\sum_{i=1}^m L_{f_i}\right)^2 + \frac{\sigma}{4} \|\psi_{i-1,k} - \psi_{i,k}\|^2 \\ &\leq \frac{1}{\sigma} t_k^2 \left(\sum_{i=1}^m L_{f_i}\right)^2 + \frac{1}{2} D_H(\psi_{i,k}, \psi_{i-1,k}) \end{aligned}$$

and thus

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

$$\mathbb{E}(D_H(y, x_k)) + \frac{1}{2\sigma} t_k^2 \left(\sum_{i=1}^m L_{f_i}\right)^2 \left(2\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 3 + 2m\right).$$

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

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

$$\mathbb{E}(D_H(y, x_0)) + \frac{1}{2\sigma} \left(\sum_{i=1}^m L_{f_i}\right)^2 \left(2\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 3 + 2m\right) \sum_{k=0}^{N-1} t_k^2.$$

This shows that

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

$$\frac{2\sigma D_H(y, x_0) + \left(2\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 3 + 2m\right) \left(\sum_{i=1}^m L_{f_i}\right)^2 \sum_{k=0}^{N-1} t_k^2}{2\sigma \sum_{k=0}^{N-1} t_k}$$

and therefore finishes the proof. □

The following result is again a consequence of [2, Proposition 4.1].

Corollary 4.6. *In the setting Problem 4.1, assume that the functions f_i are L_{f_i}-Lipschitz continuous on* $\operatorname{im}(\nabla H^*)$ *for* $i = 1, ..., m$*. Let* $x^* \in \operatorname{dom} H$ *be an optimal solution of* (9) *and* $(x_k)_{k \geq 0}$ *be a sequence generated by Algorithm 4.2 with optimal stepsize*

$$t_k := \frac{1}{\sum_{i=1}^m L_{f_i}} \sqrt{\frac{2D_H(x^*, x_0)}{2\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^2 + 3 + 2m}} \frac{1}{\sqrt{k}} \ \forall k \geq 0.$$

Then for every $N \geq 1$ *it holds*

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

$$\left(\sum_{i=1}^m L_{f_i}\right) \sqrt{\frac{2D_H(x^*, x_0)\left(2\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^2 + 3 + 2m\right)}{\sigma}} \frac{1}{\sqrt{N}}.$$

Remark 4.7. The same considerations as in Remark 3.7 about ergodic convergence are applicable also for the rates provided in Theorem 4.5 and Corollary 4.6.

Remark 4.8. Note the straightforward dependence of the optimal stepsizes as well as the right hand side of the convergence statement on the data, i.e. the distance of the initial point to optimality, the Lipschitz constants L_{f_i} and the probabilities p_i. This backs up the intuition that the decreased gradient evaluation, i.e. smaller p_i, does not come for free but at the cost of a worse constant in the convergence rate.

5 Applications

In the numerical experiments carried out in this section we will compare three versions of the provided algorithms. First of all, the *non-incremental* version, which takes *full* subgradient steps with respect to the sum of all component functions instead of every single one individually. This can be viewed as a special case of the algorithms given, when $m = 1$ and $\epsilon_{1,k} = 1$ for all $k \geq 0$. Secondly, we discuss the *non-stochastic incremental* version, which uses the subgradient of every single component function in every iteration and thus corresponds to the case when $\epsilon_{i,k} = 1$ for every $i = 1, ..., m$ and every $k \geq 0$. Lastly, we apply the algorithms as intended by evaluating the subgradients of the respective component functions incrementally with a probability different from 1.

5.1 Tomography

This application can be found in [3] and arises in the reconstruction of images in positron emission tomography (PET). We consider the following problem

$$\min_{x \in \Delta} \quad -\sum_{i=1}^m y_i \log\left(\sum_{j=1}^n r_{ij} x_j\right), \tag{13}$$

where $\Delta := \left\{x \in \mathbb{R}^n : \sum_{j=1}^n x_j = 1, \, x \geq 0\right\}$ and r_{ij} denotes for $i = 1, ..., m$ and $j = 1, ..., n$ the entry of the matrix $R \in \mathbb{R}^{m \times n}$ in the i-th row and j-th column and all of these are assumed to be strictly positive. Furthermore, y_i denotes for $i = 1, ..., m$ the positive number of photons

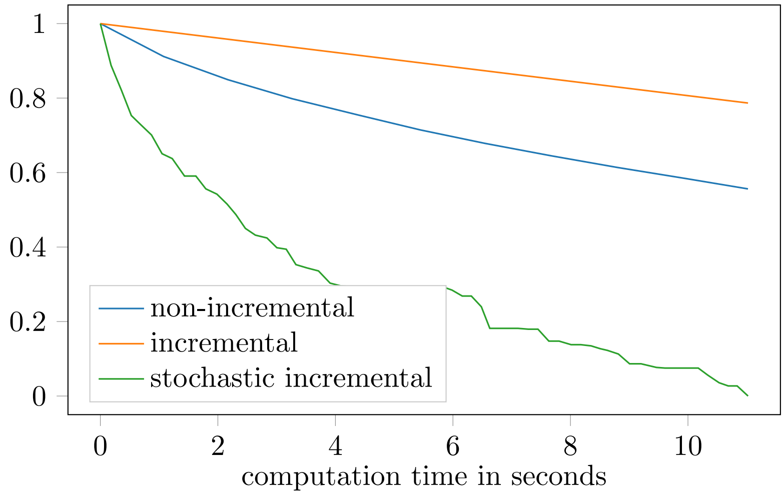

Figure 1: Results for the optimization problem (13). A plot of $\frac{f_N - f(x_{best})}{f(x_0) - f(x_{best})}$, where $f_N := \min_{0 \leq k \leq N} f(x_k)$, as a function of time, i.e. x_N is the last iterate computed before a given point in time.

		NI	DI	SI
$n = 10^4$, $m = 3n$ $p_i = 0.003, \forall i$	decrease obj.fun.val. [%]	0.066	0.032	0.15
	# outer loops	10	1	63
	# subgrad. evaluations	300000	6939	6216
$n = 10^3$, $m = 6n$ $p_i = 0.0016, \forall i$	decrease in obj.fun.val. [%]	0.196	0.515	0.671
	# outer loops	71	8	1769
	# subgrad. evaluations	426000	47435	17734

Table 1: Results for the optimization problem (13), where NI denotes the non-incremental, DI the deterministic incremental and SI the stochastic incremental version of Algorithm 3.2.

measured in the i-th bin. As discussed in Example 2.5 this can be incorporated into our framework with the mirror map $H(x) = \sum_{i=1}^{n} x_i \log(x_i)$ for $x \in \Delta$ and $H(x) = +\infty$, otherwise. As initial value we use the all ones vector divided by the dimension n.

We also want to point out that a similar example given in [2] in which the minimization of a convex function over the unit simplex Δ somehow does not match the assumption made throughout the paper as the interior of Δ is empty and the function H can therefore not be continuously differentiable in a classical sense. However, with the setting of Section 3 we are able to tackle this problem.

The bad performance, see Figure 1, of the deterministic incremental version of Algorithm 3.2 can be explained by the fact that many more evaluations of the mirror map are needed, which increases the overall computation time dramatically. The stochastic version, however, performs rather well, after only evaluating merely roughly a fifth of the total number of component functions, see Table 1.

5.2 Support Vector Machines

We deal with the classic machine learning problem of binary classification based on the well-known MNIST dataset, which contains 28 by 28 images of handwritten numbers on a grey-scale pixel map. For each of the digits the dataset comprises around 6000 training images and roughly

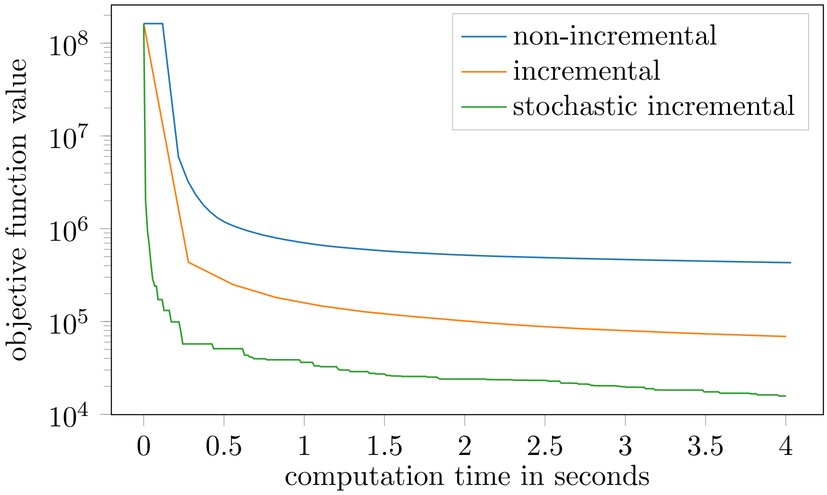

Figure 2: Numerical results for the optimization problem (14) with $\lambda = 0.01$. The plot shows $\min_{0\leq k\leq N} f(x_k)$ as a function of time, i.e. x_N is the last iterate computed before a given point in time.

1000 test images. In line with [14], we train a classifier to distinguish the numbers 6 and 7, by solving the following optimization problem

$$\min_{w\in\mathbb{R}^{784}} \quad \sum_{i=1}^{m} \max\{0, 1 - y_i\langle w, x_i\rangle\} + \lambda\|w\|_1, \tag{14}$$

where, for $i = 1, ..., m$, $x_i \in \{0, 1, \dots, 255\}^{784}$ denotes the i-th training image and $y_i \in \{-1, 1\}$ denotes the label of the i-th training image. The 1-norm serves as a regularization term and $\lambda > 0$ balances the two objectives of minimizing the classification error and reducing the 1-norm of the classifier w. To incorporate this problem into our framework, we set $H = \frac{1}{2}\|\cdot\|^2$ which leaves us with the identity as mirror map as this problem is unconstrained. The results comparing the three versions of Algorithm 4.2 discussed in the beginning of this section are illustrated in Figure 2. As initial value we simply use the all ones vector. All three versions show classical first-order behaviour, giving a fast decrease in objective function value first but then slowing down dramatically. More information about the performance can be seen in Table 2. All three algorithms results in a significant decrease in objective function after being run for only 4 seconds each. However, from a machine learning point of view, only the misclassification rate is of actual importance. In both regards, the stochastic incremental version clearly trumps the other two implementations. It is also interesting to note that it needs only a small fraction of the number of subgradient evaluations in comparison to the full non-incremental algorithm.

6 Conclusion

In this paper we present two algorithms to solve nonsmooth convex optimization problems where the objective function is a sum of *many* functions which are evaluated by their respective subgradients under the implicit presence of a constraint set which is dealt with by a so-called *mirror map*. By allowing for a random selection of each component function to evaluate in each iteration, the proposed methods become suitable even for very large-scale problems. We prove a convergence order of $\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ in expectation for the kth best objective function value, which is standard for subgradient methods. However, even for the case where all the objective functions

		NI	DI	SI
$\lambda = 0.01$	decrease in obj.fun.val. [%]	99.735	99.958	99.99
	#outer loops	83	16	370
	#subgrad. evaluations	999006	179531	36962
	misclassified[%]	1.057	0.856	0.604
$\lambda = 0.001$	decrease obj.fun.val. [%]	99.728	99.958	99.985
	#iter	75	15	336
	#subgrad. eval.	913725	179320	33777
	misclassified[%]	1.007	0.856	0.403

Table 2: Numerical results for the optimization problem (14), where NI denotes the non-incremental, DI the deterministic incremental and SI the stochastic incremental version of Algorithm 4.2. The computation for different regularization parameters λ show similar performances of the algorithms, but a lower misclassification rate for the lower value.

are differentiable it is not clear if better theoretical estimates can be achieved, due to the need of using diminishing stepsizes in order to obtain convergence in incremental algorithms. Future work could comprise the investigation of different stepsizes, such as constant or dynamic stepsizes as in [9]. Another possible extension of this would be to use different selection procedures such as random subsets of fixed size. Our framework, however, does not provide the right setting for such a *batch* approach as it would leave $\epsilon_{i,k}$ and $\epsilon_{j,k}$ dependent.

References

[1] Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

[2] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[3] Aharon Ben-Tal, Tamar Margalit, and Arkadi Nemirovski. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM Journal on Optimization*, 12(1):79–108, 2001.

[4] Dimitri P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. In: Suvrit Sra, Sebastian Nowozin, Stephen J. Wright (Eds.). *Optimization for Machine Learning*, Neural Information Processing Series, MIT Press, Massachusetts, 85–120, 2012.

[5] Jérôme Bolte and Marc Teboulle. Barrier operators and associated gradient-like dynamical systems for constrained minimization problems. *SIAM journal on control and optimization*, 42(4):1266–1292, 2003.

[6] Patrick L. Combettes and Jean-Christophe Pesquet. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, 25(2):1221–1248, 2015.

[7] José Yunier Bello Cruz. On proximal subgradient splitting method for minimizing the sum of two nonsmooth convex functions. *Set-Valued and Variational Analysis*, 25(2):245–263, 2017.

[8] Mert Gurbuzbalaban, Asuman Ozdaglar, and Pablo A. Parrilo. On the convergence rate of incremental aggregated gradient algorithms. *SIAM Journal on Optimization*, 27(2):1035–1048, 2017.

[9] Angelia Nedic and Dimitri P. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001.

[10] Arkadii S. Nemirovskii and David B. Yudin. Problem Complexity and Method Efficiency in Optimization. John Wiley & Sons Ltd, New Jersey, 1983.

[11] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

[12] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[13] Zhou Wei and Qing Hai He Nonsmooth steepest descent method by proximal subdifferentials in Hilbert spaces. *Journal of Optimization Theory and Applications*, 161(2):465–477, 2014.

[14] Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

[15] Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen Boyd, and Peter Glynn. Mirror descent in non-convex stochastic programming. *arXiv:1706.05681*, 2017.