Random minibatch projection algorithms for convex problems with functional constraints

Angelia Nedić · Ion Necoara

Received: 28 February 2019 / Accepted: date

Abstract In this paper we consider non-smooth convex optimization problems with (possibly) infinite intersection of constraints. In contrast to the classical approach, where the constraints are usually represented as intersection of simple sets, which are easy to project onto, in this paper we consider that each constraint set is given as the level set of a convex but not necessarily differentiable function. For these settings we propose subgradient iterative algorithms with random minibatch feasibility updates. At each iteration, our algorithms take a step aimed at only minimizing the objective function and then a subsequent step minimizing the feasibility violation of the observed minibatch of constraints. The feasibility updates are performed based on either parallel or sequential random observations of several constraint components. We analyze the convergence behavior of the proposed algorithms for the case when the objective function is restricted strongly convex and with bounded subgradients, while the functional constraints are endowed with a bounded first-order black-box oracle. For a diminishing stepsize, we prove sublinear convergence rates for the expected distances of the weighted averages of the iterates from the constraint set, as well as for the expected suboptimality of the function values along the weighted averages. Our convergence rates are known to be optimal for subgradient methods on this class of problems. Moreover, the rates depend explicitly on the minibatch size and show when minibatching helps a subgradient scheme with random feasibility updates.

A. Nedić
School of Electrical, Computer and Energy Engineering
Arizona State University, Tempe, USA
E-mail: angelia.nedich@asu.edu.

I. Necoara
Department of Automatic Control and Systems Engineering
University Politehnica Bucharest, 060042 Bucharest, Romania
E-mail: ion.necoara@acse.pub.ro (corresponding author).
Keywords Convex minimization · functional constraints · subgradient algorithms · random minibatch projection algorithms · convergence rates.

1 Introduction

The large sum of functions in the objective function and/or the large number of constraints in most of the practical optimization applications led the stochastic optimization field to become an essential tool for many applied mathematics areas, such as machine learning and statistics [11, 25], constrained control [20], sensor networks [1], computer science [9], inverse problems [4]. For example, in machine learning applications the optimization algorithms involve numerical computation of parameters for a system designed to make decisions based on yet unseen data [11, 25]. In particular, in support vector machines one maps the data into a higher dimensional input space and constructs an optimal separating hyperplane in this space by learning, eventually online, the hyperplanes corresponding to each data in the training set [25]. This leads to a convex optimization problem with a large number of functional constraints.

Contributions. To deal with such optimization problems having (possibly) infinite number of functional constraints, we propose subgradient methods with random feasibility updates. At each iteration, the algorithms take a step aimed at only minimizing the objective function, followed by a feasibility step for minimizing the feasibility violation of the observed minibatch of convex constraints. The feasibility updates in the first algorithm are performed using parallel random observations of several constraint components, while in the second algorithm we consider sequential random observations of constraints. Both algorithms are reminiscent of a learning process where we try to learn the constraint set while simultaneously minimizing an objective function. The proposed algorithms are applicable to the situation where the whole constraint set of the problem is not known in advance, but it is rather learned in time through observations. Also, these algorithms are of interest for (non-smooth) constrained optimization problems where the constraints are known but their number is either large or not finite.

We study the convergence properties of the proposed random minibatch projection algorithms for the case when the objective function need not be differentiable but it is restricted strongly convex, while the functional constraints are accessed through a bounded first-order black-box oracle. In doing so, we can avoid the need for projections to the set of constraints, which may be expensive computationally. For a diminishing stepsize, we prove sublinear convergence rates of order $O(1/t)$, where $t$ is the iteration counter, for the expected distances of the weighted averages of the iterates from the constraint set, as well as for the expected suboptimality of the function values along the weighted averages. Our convergence rates are known to be optimal for this class of subgradient schemes for solving non-smooth convex problems with functional constraints. Moreover, our rates depend explicitly on the minibatch size and
show when minibatching works for a subgradient method with random feasibility updates. To the best of our knowledge, this is the first work proving that subgradient methods with random minibatch feasibility steps are better than their non-minibatch variants. More explicitly, the convergence estimate for the parallel algorithm depends on a key parameter $L_N$, which determines whether minibatching helps ($L_N < 1$) or not ($L_N = 1$) and how much (the smaller $L_N$, the better is the complexity), see Theorem 2. For the sequential variant, we show that minibatching always helps and the complexity depends exponentially on the minibatch size (see Theorem 3).

**Related works.** The most prominent work for stochastic optimization problems is stochastic gradient descent (SGD) \cite{ Robbins1951, Polyak1964, B Robbins1951}. Even though SGD is a mature methodology, it only applies to optimization problems with simple constraints, requiring the whole feasible set to be projectable. A line of work that is known as alternating projections, focus on applying random projections for solving problems that are involving the intersection of (infinite) number of sets. The case when the objective function is not present in the formulation, which corresponds to the convex feasibility problem, is studied e.g. in \cite{ Lu2013, Tian2013, Zhang2013}. For this particular setting, \cite{ Lu2013} combines the smoothing technique with (minibatch) SGD, leading to stochastic alternating projection algorithms having linear convergence rates. In \cite{ Zhang2013} stochastic proximal point type steps are combined with alternating projections for solving stochastic optimization problems with infinite intersection of sets. In order to prove sublinear convergence rates $O(1/t)$, \cite{ Zhang2013} requires smooth and strongly convex objective functions, while our results are valid for a more relaxed strong convexity assumption. Lastly, \cite{ Zhang2013} assumes the projectability of individual sets, whereas in our case, the constraints might not be projectable. Stochastic forward-backward algorithms have been also applied to solve optimization problems with many constraints. However, the papers introducing those very general algorithms focused on proving convergence and did not present convergence rates, or they assume the number of constraints is finite, which is more restricted than our settings \cite{ Juditsky2018, Bauschke2017, Bauschke2018}.

In the case where the number of constraints is finite and the objective function is deterministic, Nesterov’s smoothing framework is studied in \cite{ Lin2013, Juditsky2018, Odzak2019} in the setting of accelerated proximal gradient methods. Incremental subgradient or primal-dual approaches were also proposed for solving convex problems with finite intersection of simple sets through an exact penalty reformulation \cite{ Juditsky2018}. The paper most related to our work is \cite{ Juditsky2018}, where iterative subgradient methods with random feasibility steps are proposed for solving convex problems with functional constraints. Our algorithms are minibatch extensions of the algorithm proposed in \cite{ Juditsky2018}. Moreover, in \cite{ Juditsky2018} only sublinear convergence rates of order $O(1/\sqrt{t})$ have been established for convex objective functions, while in this paper we show that $O(1/t)$ rates are valid under a relaxed strong convexity condition. Finally, since we deal with minibatching and a relaxed strong convexity assumption, our convergence analysis requires additional insights that differ from that of \cite{ Juditsky2018}. 
Notation. The inner product of two vectors \(x\) and \(y\) in \(\mathbb{R}^n\) is denoted by \((x, y)\), while \(\|x\|\) denotes the standard Euclidean norm. We write \(\text{dist}(\bar{x}, X)\) for the distance of a vector \(\bar{x}\) from a closed convex set \(X\), i.e., \(\text{dist}(\bar{x}, X) = \min_{x \in X} \|x - \bar{x}\|\), while \(\Pi_X[\bar{x}]\) denotes the projection of \(\bar{x}\) onto \(X\), i.e., \(\Pi_X[\bar{x}] = \text{argmin}_{x \in X} \|x - \bar{x}\|^2\). For a scalar \(a\), we write \(a^+ = \max\{a, 0\}\). For a convex function \(h\), we denote \(s_h(x)\) a subgradient of \(h\) at \(x\) and \(\partial h(x)\) denote the set of all subgradients of \(h\) at \(x\). We write \(\Pr\{\omega\}\) and \(E[\omega]\) to denote respectively the probability distribution and the expectation of a random variable \(\omega\).

1.1 Problem formulation

In this paper we are interested in solving the following convex constrained minimization problem:

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in X, \quad X \triangleq Y \cap (\cap_{\omega \in \mathcal{A}} X_{\omega}), \\
\text{with} & \quad X_{\omega} = \{x \in \mathbb{R}^n \mid g_{\omega}(x) \leq 0\} \quad \text{for every } \omega \in \mathcal{A}, \tag{1}
\end{align*}
\]

where \(\mathcal{A}\) is an arbitrary collection of indices. The objective function \(f\) and all constraint functions \(g_{\omega}\) are assumed convex. We also assume that the optimization problem (1) has finite optimum and we let \(f^*\) and \(X^*\) denote the optimal value and the optimal set, respectively,

\[
f^* = \inf_{x \in X} f(x), \quad X^* = \{x \in X \mid f(x) = f^*\}.
\]

We work under the premise that the collection \(\mathcal{A}\) is large, possibly infinite (even uncountable). Let us now formally state the assumptions on the functions \(f\) and \(g_{\omega}\), \(\omega \in \mathcal{A}\), of problem (1).

Assumption 1 Let the following hold:

(a) The set \(Y\) is closed and convex, and the constraint set \(X\) is nonempty.

(b) The objective function \(f : \mathbb{R}^n \to \mathbb{R}\) is restricted strongly convex on the set \(Y\) with a constant \(\mu > 0\), i.e., there exists \(x^* \in X^*\) such that

\[
f(x) \geq f(x^*) + \frac{\mu}{2} \|x - x^*\|^2 \quad \forall x \in Y.
\]

The subgradients of the function \(f\) are uniformly bounded on the set \(Y\), i.e., there is \(M_f > 0\) such that

\[
\|s_f(x)\| \leq M_f \quad \forall s_f(x) \in \partial f(x) \text{ and } x \in Y.
\]

(c) The functional constraints \(g_{\omega} : \mathbb{R}^n \to \mathbb{R}\) are convex, not necessarily differentiable, and have bounded subgradients on the set \(Y\), i.e., there is \(M_g > 0\) such that

\[
\|d\| \leq M_g \quad \forall d \in \partial g_{\omega}(x), \ x \in Y \text{ and } \omega \in \mathcal{A}.
\]
It is known that restricted strong convexity is a weaker condition than strong convexity since it is implied by strong convexity along the direction of the solution, see [12] for more details. Note that the conditions of Assumption 1(b) may look contradictory since the following relations need to hold:

\[ \frac{\mu}{2} \| x - x^* \|^2 \leq f(x) - f(x^*) \leq \langle s_f(x), x - x^* \rangle \leq M_f \| x - x^* \| \quad \forall x \in Y, x^* \in X^*, \]

where the second inequality follows from the convexity of \( f \) and the third one from the Cauchy-Schwartz inequality. This implies that \( \| x - x^* \| \leq 2M_f/\mu \) for any \( x \in Y \). Note that this inequality is always valid provided that the set \( Y \) is compact and our optimization model [1] allows us to impose such an assumption on the set \( Y \). In fact, the reader should note that we can replace the boundedness on the subgradients of \( f \), i.e. assumption \( \| s_f(x) \| \leq M_f \), with a more general assumption, that is there exist two constants \( M_{f,1}, M_{f,2} \geq 0 \) and \( x^* \in X^* \) such that:

\[ \| s_f(x) \| \leq M_{f,1} + M_{f,2} \| x - x^* \| \quad \forall s_f(x) \in \partial f(x) \text{ and } x \in Y. \]

Clearly, this condition covers the class of functions with bounded gradients, e.g. take \( M_{f,2} = 0 \), and also the class of functions with Lipschitz continuous gradients [15]. Indeed, if there is \( L_f > 0 \) such that \( \| \nabla f(x) - \nabla f(y) \| \leq L_f \| x - y \| \) for all \( x, y \in Y \), then \( \| \nabla f(x) \| \leq \| \nabla f(x^*) \| + \| \nabla f(x) - \nabla f(x^*) \| \leq \| \nabla f(x^*) \| + L_f \| x - x^* \| \), which proves our inequality for \( M_{f,1} = \max_{x \in X^*} \| \nabla f(x^*) \| \) and \( M_{f,2} = L_f \). All our derivations from the present work will hold under this more general assumption, however, the recurrence relations will be more cumbersome. Therefore, for the sake of simplicity, our convergence analysis is derived under Assumption 1(b). Moreover, when the sets \( X_\omega \) are simple for projection operation, then one may choose an alternative equivalent description of the constraint sets by letting \( g_\omega(x) = \text{dist}(x, X_\omega) \) for all \( x \in \mathbb{R}^n \). Note that in this case \( d(x) = \frac{x - R_{X_\omega}[x]}{\text{dist}(x, X_\omega)} \in \partial g_\omega(x) \) for all \( x \not \in X_\omega \). Moreover, \( \| d(x) \| = 1 \), thus the subgradients are bounded with \( M_g = 1 \) in this case. Therefore, our approach is more general than those from most of the existing works (see Related works paragraph from Section 1), which usually assume projectability of each \( X_\omega \).

2 Parallel random minibatch projections algorithm

To solve the convex problem with functional constraints [1], we first propose a subgradient method with parallel random minibatch feasibility updates. More precisely, our first algorithm is a parallel minibatch extension of the algorithm proposed in [13]. Let \( x_{k-1} \) be available at iteration \( k \), and define the update:

\[ v_k = \Pi_Y [x_{k-1} - \alpha_{k-1} s_f(x_{k-1})], \quad (2a) \]

\[ z^i_k = v_k - \beta \frac{\partial g_\omega^i(v_k)}{\| d_k \|^2} d_k^i \quad \text{for } i = 1, \ldots, N, \quad (2b) \]

\[ x_k = \Pi_Y [\bar{z}_k], \quad \text{with } \bar{z}_k = \frac{1}{N} \sum_{i=1}^N z^i_k. \quad (2c) \]
Here, \( \alpha_k > 0 \) and \( \beta > 0 \) are deterministic stepsizes and recall that \( s_f(x) \) denotes a subgradient of \( f \) at \( x \) and \( g^+_{\omega_k}(x) = \max\{g_\omega(x), 0\} \). The method takes one subgradient step for the objective function, followed by \( N \) feasibility updates in parallel, which are then averaged and projected onto the set \( Y \).

At each of the feasibility update step a random constraint is selected from the collection of the constraint sets, i.e., the index variable \( \omega^i_k \) is random with values in the set \( \mathcal{A} \). The vector \( d^i_k \) is chosen as \( d^i_k \in \partial g^+_{\omega_k^i}(v_k) \) if \( g^+_{\omega_k^i}(v_k) > 0 \) and \( d^i_k = d \) for some \( d \neq 0 \) if \( g^+_{\omega_k^i}(v_k) = 0 \). When \( g^+_{\omega_k^i}(v_k) = 0 \), we have \( z_k^i = v_k \) for any choice of \( d \neq 0 \). The initial point \( x_0 \in Y \) is selected randomly with an arbitrary distribution. The projection on the set \( Y \) in the updates (2a) and (2c) is used to ensure that each \( v_k \) and \( x_k \) remain in the set \( Y \), over which the functions \( f \) and \( g_\omega \) are assumed to have bounded subgradients. Our next assumption deals with the random variables \( \omega^i_k \). For this, we introduce the sigma-field \( \mathcal{F}_k \) induced by the history of the method, i.e., by the realizations of the initial point \( x_0 \) and the variables \( \omega^i_k \) up to main iteration \( k \):

\[
\mathcal{F}_k = \{x_0\} \cup \{\omega^j_t \mid 1 \leq t \leq k, 1 \leq j \leq N\},
\]

which contains the same information as the set \( \{x_0\} \cup \{\{v_t, x_t\} \mid 1 \leq t \leq k\} \).

For notational convenience, we will allow \( k = 0 \) by letting \( \mathcal{F}_0 = \{x_0\} \). We impose the following assumption.

**Assumption 2** There exists a constant \( c \in (0, \infty) \) such that

\[
\text{dist}^2(y, X) \leq c \cdot \mathbb{E} \left[ (g^+_{\omega^i_k}(y))^2 \mid \mathcal{F}_{k-1} \right], \quad \forall y \in Y, \ k \geq 1 \text{ and } i = 1, \ldots, N.
\]

Assumption 2 does not require that \( \omega^1_k, \ldots, \omega^N_k \) are conditionally independent, given \( \mathcal{F}_{k-1} \). For example, when the collection \( \mathcal{A} \) is finite, the indices \( i \in \mathcal{A} \) can be selected randomly without replacement, i.e., given the realizations of \( \omega^1_k = j_1, \ldots, \omega^{i-1}_k = j_{i-1} \), the index \( \omega^i_k \) can be random with realizations in \( \mathcal{A} \setminus \{j_1, \ldots, j_{i-1}\} \). As another example, the index set \( \mathcal{A} \) can be partitioned in \( N \) disjoint sets \( \cup_{i=1}^N \mathcal{A}_i = \mathcal{A} \), and each \( \omega^i_k \) can be uniformly distributed over the index set \( \mathcal{A}_i \). Such a sampling allows for a parallel computation of all \( z_k^i \) in the algorithm (2). One can also combine the preceding two possibilities, by using a smaller partition of the set \( \mathcal{A} \), and in each of the partitions choose the corresponding \( \omega^i_k \) sequentially, without replacement. Assumption 2 is crucial in our convergence analysis of method (2). It summarizes all the information we need regarding the distributions of the random variables \( \omega^i_k \) and the initial point \( x_0 \). A discussion on the connection between the constant \( c \) of Assumption 2 and the linear regularity constant for the sets \( \mathcal{A}_\omega \) can be found in [12,15,13]. When each set \( \mathcal{A}_\omega \) is given by either linear inequality or a linear equality, one can verify that the sets are linearly regular, see [4,13]. Hence, Assumption 2 is also satisfied in this case. However, Assumption 2 holds for more general sets, e.g., when a strengthened Slater condition holds for the collection of functional constraints \( (X_\omega)_{\omega \in \mathcal{A}} \) such as the generalized Robinson condition, as detailed in Corollary 2 of [10].
2.1 Preliminary results

In this section, we derive some preliminary results for later use in the convergence analysis of method (2). We start by recalling a basic property of the projection operation on a closed convex set $Y \subseteq \mathbb{R}^n$:

$$\|\Pi_Y[v] - y\|^2 \leq \|v - y\|^2 - \|\Pi_Y[v] - v\|^2 \quad \text{for any } v \in \mathbb{R}^n \text{ and } y \in Y.$$  \hfill (3)

We now show that the parameter $c$ in Assumption 2 satisfies the following inequality:

**Lemma 1** Let Assumption 1(c) and Assumption 2 hold. Then, we have:

$$cM_g^2 \geq 1.$$  

**Proof** Let $y \in Y$ be such that $y \notin X$. Then, there exists $\bar{\omega} \in \mathcal{A}$ such that the convex function $g_{\bar{\omega}}$ satisfies $g_{\bar{\omega}}(y) > 0$. Consequently, for any $s_g(y) \in \partial g_{\bar{\omega}}(y)$ we also have $s_g(y) \in \partial g_{\bar{\omega}}^+(y)$, and using convexity of $g_{\bar{\omega}}^+$, we obtain:

$$0 = g_{\bar{\omega}}^+(\Pi_X[y]) \geq g_{\bar{\omega}}^+(y) + (s_g(y), \Pi_X[y] - y) \geq g_{\bar{\omega}}^+(y) - M_g \|\Pi_X[y] - y\|,$$

or equivalently

$$g_{\bar{\omega}}^+(y) \leq M_g \|\Pi_X[y] - y\|.$$  

On the other hand for those $\omega \in \mathcal{A}$ for which $g_{\omega}(y) = 0$ we automatically have

$$0 = g_{\omega}^+(y) \leq M_g \|\Pi_X[y] - y\|.$$  

In conclusion, for any $\omega \in \mathcal{A}$ there holds:

$$g_{\omega}^+(y) \leq M_g \|\Pi_X[y] - y\|.$$  

Combining the preceding inequality and Assumption 2, we obtain:

$$\text{dist}^2(y, X) = \|\Pi_X[y] - y\|^2 \leq cE[(g_{\omega_k}^+(y))^2 \mid \mathcal{F}_{k-1}]$$

$$\leq cE[M_g^2 \|\Pi_X[y] - y\|^2 \mid \mathcal{F}_{k-1}] = cM_g^2 \text{dist}^2(y, X),$$

which proves our relation $cM_g^2 \geq 1$. \hfill $\square$

We now derive a relation between the iterates $v_{k+1}$ and $x_k$.

**Lemma 2** Let Assumptions 1(a) and 1(b) hold. Let $v_{k+1}$ be obtained via equation (2a) for a given $x_k \in Y$. Then, for any optimal solution $x^*$ of the problem (1) and any $\rho \in (0, 1)$, we have:

$$\|v_{k+1} - x^*\|^2 + 2\alpha_k(1 - \rho)(f(\Pi_X[x_k]) - f^*)$$

$$\leq (1 - \alpha_k \rho \mu)\|x_k - x^*\|^2 + 2\alpha_k(1 - \rho)M_f \|\Pi_X[x_k] - x_k\| + \alpha_k^2 M_f^2.$$
The best choice for the parameter $\rho$ will just do fine.

Using the standard analysis of the projected subgradient method and the fact that the subgradients of $f$ are uniformly bounded on $Y$, we have for any $x^* \in X^*$:

$$
\|v_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - 2\alpha_k (f(x_k) - f(x^*)) + \alpha_k^2 M_f^2. \tag{4}
$$

We provide a lower bound on $f(x_k) - f(x^*)$. We consider two choices, namely, one is based on the restricted strong convexity of $f$ and the other is based on considering another intermittent point. By the restricted strong convexity of $f$, we have

$$
f(x_k) - f(x^*) \geq \frac{\rho}{2} \|x_k - x^*\|^2. \tag{5}
$$

The other choice consists of adding and subtracting $f(\Pi_Y x_k)$, which yields

$$
f(x_k) - f(x^*) = f(x_k) - f(\Pi_Y x_k) + f(\Pi_Y x_k) - f(x^*)
\geq -\|s_f(\Pi_Y x_k)\| \|\Pi_Y x_k - x_k\| + f(\Pi_Y x_k) - f(x^*),
$$

where the last inequality follows by the convexity of $f$ and the Cauchy-Schwarz inequality. By Assumption 1(b), the subgradients of $f$ are uniformly bounded on $Y$ and hence, also on $Z$, implying that

$$
f(x_k) - f(x^*) \geq f(\Pi_Y x_k) - f(x^*) - M_f \|\Pi_Y x_k - x_k\|. \tag{6}
$$

We now let $\rho \in (0, 1)$ be arbitrary. By multiplying relation (5) with $\rho$ and relation (6) with $(1 - \rho)$, and by adding the resulting relations, we obtain

$$
f(x_k) - f(x^*) \geq \frac{\rho \mu}{2} \|x_k - x^*\|^2
+(1 - \rho) (f(\Pi_Y x_k) - f(x^*)) - (1 - \rho) M_f \|\Pi_Y x_k - x_k\|. \tag{7}
$$

By using the estimate (2) in relation (4), we obtain

$$
\|v_{k+1} - x^*\|^2 \leq (1 - \alpha_k \rho \mu) \|x_k - x^*\|^2 - 2\alpha_k (1 - \rho) (f(\Pi_Y x_k) - f(x^*))
+ 2\alpha_k (1 - \rho) M_f \|\Pi_Y x_k - x_k\| + \alpha_k^2 M_f^2, \tag{8}
$$

and after re-arranging some of the terms we get the relation of the lemma. □

Remark 1 The best choice for the parameter $\rho$ is not apparent at this point. It is important to have it in order to have the function value involved in the expression, but it can be that $\rho = \frac{1}{2}$ will just do fine.

We next state a result that will be used to provide a basic relation between the iterates $v_k$ and $x_{k-1}$. The relation is stated in a generic form, and its proof can be found in [21].

**Lemma 3** [21] Let $g$ be a convex function over a closed convex set $Z$, and let $y$ be given by

$$
y = \Pi_Z \left[ v - \beta g^+(v) \frac{d}{\|d\|^2} d \right] \quad \text{for } v \in Z, \ d \in \partial g^+(v), \ \beta > 0,
$$

where $d \neq 0$. Then, for any $\bar{z} \in Z$ such that $g^+(\bar{z}) = 0$, we have

$$
\|y - \bar{z}\|^2 \leq \|v - \bar{z}\|^2 - \beta(2 - \beta) \frac{(g^+(v))^2}{\|d\|^2}.
$$
In the analysis, we will also make use of the relation for averages, stating that for given vectors \( u_1, \ldots, u_N \in \mathbb{R}^n \) and their average \( \bar{u} = \frac{1}{N} \sum_{i=1}^{N} u_i \), the following relation is valid for any vector \( w \in \mathbb{R}^n \):

\[
\| \bar{u} - w \|^2 = \frac{1}{N} \sum_{i=1}^{N} \| u_i - w \|^2 - \frac{1}{N} \sum_{j=1}^{N} \| u_j - \bar{u} \|^2.
\] (9)

Now we provide a basic relation for the iterate \( x_k \) upon completion of the \( N \) randomly sampled feasibility updates.

**Lemma 4** Let Assumption 1(a) hold. Let \( x_k \) be obtained via updates (23) and (24) for a given \( v_k \in Y \) and \( \beta > 0 \). Then, the following relation holds:

\[
\text{dist}^2(x_k, X) \leq \text{dist}^2(v_k, X) - \frac{\beta(2 - \beta)}{N} \sum_{i=1}^{N} \frac{(g_{\omega_k}^i(v_k))^2}{\| d_k^i \|^2} - \beta^2 V_N(v_k),
\]

where \( V_N(v_k) \) is the total variation of the minibatch subgradients, i.e.,

\[
V_N(v_k) = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{g_{\omega_k}^i(v_k)}{\| d_k^i \|^2} d_k^i - \frac{1}{N} \sum_{j=1}^{N} \frac{g_{\omega_k}^j(v_k)}{\| d_k^j \|^2} d_k^j \right)^2.
\]

**Proof** By the projection property (3) and the definition of \( x_k \), we have for any \( y \in X \) that:

\[
\| x_k - y \|^2 \leq \| \bar{z}_k - y \|^2 - \| x_k - \bar{z}_k \|^2.
\] (10)

By the definition we have \( \bar{z}_k = \frac{1}{N} \sum_{i=1}^{N} z_k^i \). Thus, by using relation (9) for the collection \( z_k^1, \ldots, z_k^N \), we have for any \( w \in \mathbb{R}^n \),

\[
\| \bar{z}_k - w \|^2 = \frac{1}{N} \sum_{i=1}^{N} \| z_k^i - w \|^2 - \frac{1}{N} \sum_{i=1}^{N} \| z_k^i - \bar{z}_k \|^2.
\] (11)

Letting \( w = y \) in the preceding relation and combining the resulting relation with (10), we obtain

\[
\| x_k - y \|^2 \leq \frac{1}{N} \sum_{i=1}^{N} \| z_k^i - y \|^2 - \frac{1}{N} \sum_{i=1}^{N} \| z_k^i - \bar{z}_k \|^2 - \| x_k - \bar{z}_k \|^2.
\]

Now, we use the definition of the iterates \( z_k^i \) in algorithm (2) and Lemma 3 with \( Z = \mathbb{R}^n \). Thus, we obtain for any \( y \in X \) (for which we would have \( g_{\omega_k}^i(y) = 0 \) for any realization of \( \omega_k^i \)) and for any \( i = 1, \ldots, N \),

\[
\| z_k^i - y \|^2 \leq \| v_k - y \|^2 - \beta(2 - \beta) \frac{(g_{\omega_k}^i(v_k))^2}{\| d_k^i \|^2}.
\]

Hence, it follows that for any \( y \in X \),

\[
\| x_k - y \|^2 \leq \| v_k - y \|^2 - \beta(2 - \beta) \frac{1}{N} \sum_{i=1}^{N} \frac{(g_{\omega_k}^i(v_k))^2}{\| d_k^i \|^2} - \frac{1}{N} \sum_{i=1}^{N} \| z_k^i - \bar{z}_k \|^2 - \| x_k - \bar{z}_k \|^2.
\]
From the definition of the iterates $z_k^i$ in algorithm (2), we see that
\[
\|z_k^i - z_k\|^2 = \beta^2 \left\| \frac{g_{\omega_k^i}(v_k)}{\|d_k^i\|^2} d_k^i - \frac{1}{N} \sum_{j=1}^N \frac{g_{\omega_k^j}(v_k)}{\|d_k^j\|^2} d_k^j \right\|^2.
\]
By defining
\[
V_N(v_k) = \frac{1}{N} \sum_{i=1}^N \left\| \frac{g_{\omega_k^i}(v_k)}{\|d_k^i\|^2} d_k^i - \frac{1}{N} \sum_{j=1}^N \frac{g_{\omega_k^j}(v_k)}{\|d_k^j\|^2} d_k^j \right\|^2,
\]
we have
\[
\frac{1}{N} \sum_{i=1}^N \|z_k^i - z_k\|^2 = \beta^2 V_N(v_k).
\]
Therefore, we obtain for any $y \in X$,
\[
\|x_k - y\|^2 - \|v_k - y\|^2 - \beta(2 - \beta) \sum_{i=1}^N \frac{(g_{\omega_k^i}(v_k))^2}{\|d_k^i\|^2} - \beta^2 V_N(v_k).
\] (12)
The statement of the lemma follows by letting $y = \Pi_X[v_k]$ in the preceding relation and using the fact that $\|x_k - \Pi_X[x_k]\| \leq \|x_k - \Pi_X[v_k]\|$. \hfill \Box

Let us define the following parameters:
\[
L_N^k = \left\| \frac{1}{N} \sum_{i=1}^N \frac{g_{\omega_k^i}(v_k)}{\|d_k^i\|^2} d_k^i \right\|^2 / \left\| \frac{1}{N} \sum_{i=1}^N \frac{(g_{\omega_k^i}(v_k))^2}{\|d_k^i\|^2} \right\| \quad \text{and} \quad L_N = \max_{k \geq 0} L_N^k. \quad (13)
\]
From Jensen’s inequality it follows that $L_N^k \leq 1$. However, there are also convex functions $g_{\omega_k}$ such that $L_N^k < 1$. We postpone the derivation of such examples of functional constraints satisfying condition $L_N^k < 1$ until Section 2.3. The parameter $L_N \leq 1$ will play a key role in our derivations below. In particular, we obtain the following simplification for Lemma 4.

**Lemma 5** Let Assumptions 1(a) and 1(c) hold. Let $L_N \leq 1$ as defined in (13) and $x_k$ be obtained via updates (2b) and (2a) for a given $v_k \in Y$ and $\beta \in (0, 2/L_N)$. Then, the following relation holds:
\[
\text{dist}(x_k, X) \leq \text{dist}(v_k, X) - \frac{\beta(2 - \beta L_N)}{NM_2} \sum_{i=1}^N (g_{\omega_k^i}(v_k))^2.
\]

**Proof** Note that the total variation of the minibatch subgradients $V_N(v_k)$ can be written equivalently as:
\[
V_N(v_k) = \frac{1}{N} \sum_{i=1}^N \left( \frac{g_{\omega_k^i}(v_k)}{\|d_k^i\|^2} \right)^2 - \left\| \frac{1}{N} \sum_{i=1}^N \frac{g_{\omega_k^i}(v_k)}{\|d_k^i\|^2} d_k^i \right\|^2.
\]
Using the previous expression of $V_N$ and the definitions of $L_N^k$ and $L_N$ from (3) in Lemma 3 we get:

$$\text{dist}^2(x_k, X) \leq \text{dist}^2(v_k, X) - \frac{\beta(2 - \beta)}{N} \sum_{i=1}^{N} \frac{(g_{\omega_k}^+(v_k))^2}{\|d_k^i\|^2} - \beta^2 V_N(v_k)$$

$$= \text{dist}^2(v_k, X) - \frac{\beta(2 - \beta)}{N} \sum_{i=1}^{N} \frac{(g_{\omega_k}^+(v_k))^2}{\|d_k^i\|^2} - \beta^2 \frac{(1 - L_N^k)}{N} \sum_{i=1}^{N} \frac{(g_{\omega_k}^+(v_k))^2}{\|d_k^i\|^2}$$

$$\leq \text{dist}^2(v_k, X) - \frac{\beta(2 - \beta L_N)}{N} \sum_{i=1}^{N} \frac{(g_{\omega_k}^+(v_k))^2}{\|d_k^i\|^2}$$

By Assumption (1c) each function $g_i$ has bounded subgradients uniformly on $Y$. Hence, we have $\|d_k^i\| \leq M_g$, which used in the previous inequality implies the statement of the lemma.

Note that the largest decrease in Lemma 5 is obtained by maximizing $\beta(2 - \beta L_N)$, that is, the optimal stepsize is $\beta = 1/L_N$. We now combine Lemma 2 and Lemma 5 to provide a basic relation for the subsequent analysis.

**Lemma 6** Consider the method in (2), and let Assumption (1) hold. Let the stepsize $\alpha_k$ be such that $1 - \frac{\alpha_k \mu}{2} > 0$ for all $k \geq 0$ and stepsize $\beta \in (0, 2/L_N)$, with $L_N \leq 1$ defined in (3). Then, the iterates of the method (2) satisfy the following recurrence for any optimal solution $x^* \in X$ and for all $k \geq 0$:

$$\|v_{k+1} - x^*\|^2 + \alpha_k (f(\Pi_X[x_k]) - f^*) \leq \left(1 - \frac{\alpha_k \mu}{2}\right) \|v_k - x^*\|^2$$

$$- \left(1 - \frac{\alpha_k \mu}{2}\right) \frac{\beta(2 - \beta L_N)}{N M_g^2} \sum_{i=1}^{N} (g_{\omega_k}^+(v_k))^2 + \frac{\eta}{2} \|\Pi_X[x_k] - x_k\|^2$$

$$+ \alpha_k \left(1 + \frac{1}{2\eta}\right) M_f^2,$$

where $\eta > 0$ is arbitrary.

**Proof** Let $x^* \in X$ be an optimal solution of problem (1). Then, we use Lemma 2 for $\rho = \frac{1}{2}$ so that for all $k \geq 0$, we have

$$\|v_{k+1} - x^*\|^2 + \alpha_k (f(\Pi_X[x_k]) - f^*) \leq \left(1 - \frac{\alpha_k \mu}{2}\right) \|x_k - x^*\|^2$$

$$+ \alpha_k M_f \|\Pi_X[x_k] - x_k\| + \alpha_k^2 M_f^2.$$

Using the same reasoning as in the proof of Lemma 5 for the inequality (12) with $y = x^*$ gives:

$$\|x_k - x^*\|^2 \leq \|v_k - x^*\|^2 - \frac{\beta(2 - \beta)}{N} \sum_{i=1}^{N} \frac{(g_{\omega_k}^+(v_k))^2}{\|d_k^i\|^2} - \beta^2 V_N(v_k)$$

$$\leq \|v_k - x^*\|^2 - \frac{\beta(2 - \beta L_N)}{N M_g^2} \sum_{i=1}^{N} (g_{\omega_k}^+(v_k))^2.$$
Combining the preceding two relations yields
\[
\|v_{k+1} - x^*\|^2 + \alpha_k (f(\Pi_X[x_k]) - f^*) \leq \left(1 - \frac{\alpha_k \mu}{2}\right) \|v_k - x^*\|^2 + \left(1 - \frac{\alpha_k \mu}{2}\right) \beta (2 - \beta L_N) \sum_{i=1}^{N} (g^+_\omega(v_k))^2 + \alpha_k M_f \|\Pi_X[x_k] - x_k\| + \alpha_k^2 M_f^2.
\] (14)

We next approximate the term that is linear in \(\alpha_k\), i.e. \(\alpha_k M_f \|\Pi_X[x_k] - x_k\|\), with a sum of two quadratic terms, one of which is in the order of \(\alpha_k^2\), as:
\[
\alpha_k M_f \|\Pi_X[x_k] - x_k\| = (\alpha_k \sqrt{\eta}^{-1} M_f) (\sqrt{\eta} \|\Pi_X[x_k] - x_k\|)
\leq \frac{1}{2} \left( \alpha_k^2 \eta^{-1} M_f^2 + \eta \|\Pi_X[x_k] - x_k\|^2 \right),
\]
for any arbitrary \(\eta > 0\). Substituting the preceding estimate in (14), we obtain the stated relation.

2.2 Convergence rates

In this section we derive the convergence rates of algorithm (2). For this, we first provide a recurrence relation for the iterates in expectation, which is the key relation for our convergence rate results. Note that \(c M_f^2 \geq 1\) according to Lemma 4 and \(L_N \in (0, 1]\). Thus, by increasing \(c\), \(M_f\) and/or \(L_N\), we can always ensure that \(c M_f^2 L_N > 1\). In the sequel, we assume that \(c M_f^2 L_N > 1\).

**Theorem 1** Consider the iterative process (2), and let Assumption 1 and Assumption 2 hold. Let the stepsizes \(\alpha_k\) be such that \(1 - \frac{\alpha_k \mu}{2} > 0\) for all \(k \geq 0\) and \(\beta \in (0, 2/L_N]\), with \(L_N \leq 1\) defined in (13). Then, for the algorithm (2), by defining \(q_N = \frac{\alpha_k^2 \eta^{-1} M_f^2}{c M_f^2 L_N} < 1\), we have almost surely for all \(k \geq 0\),
\[
\mathbb{E}[\|v_{k+1} - x^*\|^2 | \mathcal{F}_{k-1}] + \alpha_k \mathbb{E}[(f(\Pi_X[x_k]) - f^*) | \mathcal{F}_{k-1}]
\leq \left(1 - \frac{\alpha_k \mu}{2}\right) \|v_k - x^*\|^2 - \frac{1}{2} \left(1 - \frac{\alpha_k \mu}{2}\right) \frac{q_N}{1 - q_N} \mathbb{E}[\delta^2(x_k, X) | \mathcal{F}_{k-1}]
+ \alpha_k^2 \left(1 + \frac{1}{2 q_N (2 - \alpha_k \mu)}\right) M_f^2.
\]

**Proof** From Lemma 4 by taking the conditional expectation on the past \(\mathcal{F}_{k-1}\), we have almost surely for all \(k \geq 0\),
\[
\mathbb{E}[\|v_{k+1} - x^*\|^2 | \mathcal{F}_{k-1}] + \alpha_k \mathbb{E}[(f(\Pi_X[x_k]) - f^*) | \mathcal{F}_{k-1}]
\leq \left(1 - \frac{\alpha_k \mu}{2}\right) \|v_k - x^*\|^2 - \left(1 - \frac{\alpha_k \mu}{2}\right) \frac{\beta (2 - \beta L_N)}{N M_f^2} \sum_{i=1}^{N} \mathbb{E}\left[(g^+_\omega(v_k))^2 | \mathcal{F}_{k-1}\right]
+ \frac{\eta}{2} \mathbb{E}[(\Pi_X[x_k] - x_k)^2 | \mathcal{F}_{k-1}] + \alpha_k^2 \left(1 + \frac{1}{2 q_N}\right) M_f^2.
\] (15)
where $\eta > 0$ is arbitrary. By Assumption 2, it follows that
\[
\mathbb{E}\left[(g_{\omega_i}^+(v_k))^2 \mid \mathcal{F}_{k-1}\right] \geq \frac{1}{c} \text{dist}^2(v_k, X) \quad \text{for all } i = 1, \ldots, N.
\]
Hence
\[
\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[(g_{\omega_i}^+(v_k))^2 \mid \mathcal{F}_{k-1}\right] \geq \frac{1}{c} \text{dist}^2(v_k, X). \quad (16)
\]
Taking the conditional expectation on the past $\mathcal{F}_{k-1}$ in the relation of Lemma 4, and using relation (16), we obtain almost surely
\[
\mathbb{E}[\text{dist}^2(x_k, X) \mid \mathcal{F}_{k-1}] \leq (1 - q_N) \text{dist}^2(v_k, X),
\]
where we denote
\[
q_N = \frac{\beta(2 - \beta L_N)}{cM^2_f}.
\] (17)
Recall that we assume $cM^2_f L_N > 1$, then $q_N < 1$ (since $\max_{\delta} \beta(2 - \beta L_N) = 1/L_N$). Hence, $1 - q_N > 0$. By dividing with $1 - q_N$, we further obtain
\[
\text{dist}^2(v_k, X) \geq \frac{1}{1 - q_N} \mathbb{E}[\text{dist}^2(x_k, X) \mid \mathcal{F}_{k-1}].
\]
Substituting the preceding estimate in relation (16), yields
\[
\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[(g_{\omega_i}^+(v_k))^2 \mid \mathcal{F}_{k-1}\right] \geq \frac{1}{c(1 - q_N)} \mathbb{E}[\text{dist}^2(x_k, X) \mid \mathcal{F}_{k-1}]. \quad (18)
\]
We now use estimate (18) in relation (15), and thus obtain
\[
\mathbb{E}[(\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_{k-1} \big) + \alpha_k \mathbb{E}[(f(\Pi_X[x_k]) - f^\star) \mid \mathcal{F}_{k-1}] \\
\leq \left(1 - \frac{\alpha_k \mu}{2}\right) \|x_k - x^*\|^2 - \left(1 - \frac{\alpha_k \mu}{2}\right) \frac{\beta(2 - \beta L_N)}{(1 - q_N)cM^2_f} \mathbb{E}[\text{dist}^2(x_k, X) \mid \mathcal{F}_{k-1}] \\
+ \frac{\eta}{2} \mathbb{E}[\|\Pi_X[x_k] - x_k\|^2 \mid \mathcal{F}_{k-1}] + \alpha_k^2 \left(1 + \frac{1}{2\eta}\right) M^2_f.
\]
By the definition of $q$ (see (17)), we have
\[
\frac{\beta(2 - \beta L_N)}{(1 - q_N)cM^2_f} = \frac{q_N}{1 - q_N}.
\]
Hence,
\[
\mathbb{E}[(\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_{k-1} \big) + \alpha_k \mathbb{E}[(f(\Pi_X[x_k]) - f^\star) \mid \mathcal{F}_{k-1}] \\
\leq \left(1 - \frac{\alpha_k \mu}{2}\right) \|x_k - x^*\|^2 - \left(1 - \frac{\alpha_k \mu}{2}\right) \frac{q_N}{1 - q_N} - \frac{\eta}{2} \mathbb{E}[\text{dist}^2(x_k, X) \mid \mathcal{F}_{k-1}] \\
+ \alpha_k^2 \left(1 + \frac{1}{2\eta}\right) M^2_f,
\]
and by letting $\eta = \left(1 - \frac{\alpha_k \mu}{2}\right) \frac{q_N}{1 - q_N} > 0$, the desired relation follows. \qed
We now turn our attention to the stepsize $\alpha_k$. We consider $\alpha_k$ of the form:

$$\alpha_k = \frac{2}{\mu} \gamma_k$$

for all $k \geq 0$, for some diminishing sequence $\gamma_k$ as detailed below. Indeed, for this choice, the recurrence from Theorem 1 becomes:

$$E[\|v_{k+1} - x^*\|^2 | \mathcal{F}_{k-1}] + \frac{2}{\mu} \gamma_k E[(f(\Pi_X[x_k]) - f^*) | \mathcal{F}_{k-1}]$$

$$\leq (1 - \gamma_k) \|v_k - x^*\|^2 - \frac{1}{2} (1 - \gamma_k) \frac{qN}{1 - qN} E[dist^2(x_k, X) | \mathcal{F}_{k-1}]$$

$$+ \frac{4}{\mu^2} \gamma_k^2 \left(1 + \frac{1 - qN}{qN (2 - 2\gamma_k)}\right) M_f^2,$$

where recall that $q_N = \frac{b(2 - \beta L)}{cM_g}$. Let $\gamma_k$ be given by

$$\gamma_k = \frac{2}{k+1},$$

hence the stepsize $\alpha_k = \frac{4}{\mu (k+1)}$, $\forall k \geq 0$.

Since the sequence $\gamma_k$ is decreasing, we have

$$\gamma_k \leq \frac{2}{3}$$

for all $k \geq 1$,

implying that

$$1 - \gamma_k \geq \frac{1}{3}$$

for all $k \geq 1$.

Using this estimate in (19), we obtain

$$E[\|v_{k+1} - x^*\|^2 | \mathcal{F}_{k-1}] + \frac{2}{\mu} \gamma_k E[(f(\Pi_X[x_k]) - f^*) | \mathcal{F}_{k-1}]$$

$$\leq (1 - \gamma_k) \|v_k - x^*\|^2 - \frac{1}{6} \frac{qN}{1 - qN} E[dist^2(x_k, X) | \mathcal{F}_{k-1}]$$

$$+ \frac{4}{\mu^2} \gamma_k^2 \left(1 + \frac{2(1 - qN)}{qN}\right) M_f^2.$$

Next, we note that

$$\frac{1 - \gamma_k}{\gamma_k^2} \leq \frac{1}{\gamma_k^2 - 1}$$

for all $k \geq 1$.

Dividing (20) by $\gamma_k^2$ and using the preceding inequality we have for all $k \geq 1$, after taking total expectations and rearranging terms:

$$\gamma_k^{-2} E[\|v_{k+1} - x^*\|^2] + \frac{2}{\mu} \gamma_k^{-1} E[(f(\Pi_X[x_k]) - f^*)] + \frac{4}{\mu^2} \frac{qN}{1 - qN} E[dist^2(x_k, X)]$$

$$\leq \gamma_k^{-2} E[\|v_k - x^*\|^2] + \frac{4}{\mu^2} \left(1 + \frac{2(1 - qN)}{qN}\right) M_f^2.$$
Summing these over \( k = 1, \ldots, t \), for some \( t > 0 \), we obtain

\[
\gamma_t^{-2} E[||v_{t+1} - x^*||^2] + \frac{2}{\mu} \sum_{k=1}^{t} \gamma_k^{-1} E[(f(\Pi_X [x_k]) - f^*)]
\]

(21)

\[ + \frac{1}{6} \frac{qN}{1 - qN} \sum_{k=1}^{t} \gamma_k^{-2} E[\text{dist}^2(x_k, X)] \leq \gamma_0^{-2} E[||v_1 - x^*||^2] + t \frac{4}{\mu^2} \left(1 + \frac{2(1 - qN)}{qN}\right) M_f^2.
\]

Using the definition of \( \gamma_k \), (21) implies

\[
\frac{(t + 1)^2}{4} E[||v_{t+1} - x^*||^2] + \frac{1}{(t + 1)\mu} E \left[ \sum_{k=1}^{t} (k + 1)^2 (f(\Pi_X [x_k]) - f^*) \right] 
\]

\[ + \frac{qN}{24(1 - qN)} \sum_{k=1}^{t} (k + 1)^2 \|x_k - \Pi_X [x_k]\|^2 \]

(22)

\[ \leq \frac{1}{4} E[||v_1 - x^*||^2] + \frac{4t}{\mu^2} \left(1 + \frac{2(1 - qN)}{qN}\right) M_f^2.
\]

We finally obtain by the linearity of the expectation operation:

\[
\frac{(t + 1)^2}{4} E[||v_{t+1} - x^*||^2] + \frac{1}{(t + 1)\mu} E \left[ \sum_{k=1}^{t} (k + 1)^2 (f(\Pi_X [x_k]) - f^*) \right] 
\]

\[ + \frac{qN}{24(1 - qN)} \sum_{k=1}^{t} (k + 1)^2 \|x_k - \Pi_X [x_k]\|^2 \]

(22)

\[ \leq \frac{1}{4} E[||v_1 - x^*||^2] + \frac{4t}{\mu^2} \left(1 + \frac{2(1 - qN)}{qN}\right) M_f^2.
\]

Define for \( t \geq 1 \) the sum

\[ S_t = \sum_{k=1}^{t} (k + 1)^2 \sim O(t^3).
\]

Define also the following weighted averages (convex combinations)

\[ \hat{x}_t = \sum_{k=1}^{t} a_k x_k, \quad \hat{w}_t = \sum_{k=1}^{t} a_k \Pi_X [x_k], \]

(23)

with \( a_k = \frac{(k+1)^2}{S_t} \), hence satisfying \( \sum_{k=1}^{t} a_k = 1 \). Using convexity of the function \( f \) and of the norm-squared, we have

\[
\frac{(t + 1)^2}{4} E[||v_{t+1} - x^*||^2] + \frac{S_t}{(t + 1)\mu} E[(f(\hat{w}_t) - f^*)] + \frac{qNS_t}{24(1 - qN)} E[||\hat{w}_t - \hat{x}_t||^2]
\]

\[ \leq \frac{1}{4} E[||v_1 - x^*||^2] + \frac{4t}{\mu^2} \left(1 + \frac{2(1 - qN)}{qN}\right) M_f^2.
\]

(24)

If we define \( b_N^p = qN(1 - qN)^{-1} = (1 - qN)^{-1} - 1 \), then (24) becomes:

\[
\frac{(t + 1)^2}{4} E[||v_{t+1} - x^*||^2] + \frac{S_t}{(t + 1)\mu} E[(f(\hat{w}_t) - f^*)] + \frac{b_m^p S_t}{24} E[||\hat{w}_t - \hat{x}_t||^2]
\]
Next theorem summarizes the convergence rates followed from the previous discussion. For simplicity of the exposition, we omit the constants and express the rates only in terms of the dominant powers of $t$:

**Theorem 2** Let Assumption 1 and Assumption 2 hold and the stepsizes $\alpha_k = \frac{4}{\mu(k+1)}$ and $\beta \in (0, 2/L_N)$, with $L_N \leq 1$ defined in (13). Let also $q_N = \frac{\beta}{\beta(2-\beta L_N)} < 1$ and $b_p^N = (1-q_N)^{-1}-1$. Then, the following sublinear rates for suboptimality and feasibility violation hold for the average sequence $\hat{x}_t$ generated by the parallel algorithm (2):

\[
E[|f(\hat{x}_t) - f^*|] \leq O\left(\frac{1}{t} + \frac{1}{\sqrt{b_p^N t}}\right), \quad E[\text{dist}_X(\hat{x}_t)] \leq O\left(\frac{1}{\sqrt{b_p^N t}}\right).
\]

**Proof** From the recurrence (25), omitting the constants but keeping the terms depending on $b_p^N = (1-q_N)^{-1}-1$, we get the following convergence rates in terms of these weighted averages $\hat{w}_t$ and $\hat{x}_t$:

\[
E[f(\hat{w}_t) - f^*] \leq O\left(\frac{1}{t} + \frac{1}{b_p^N t}\right) \quad \text{and} \quad E[||\hat{w}_t - \hat{x}_t||^2] \leq O\left(\frac{1}{b_p^N t^2}\right).
\]

Since $\hat{w}_t \in X$ and using the Jensen’s inequality we get the following convergence rate for the feasibility violation of the constraints:

\[
E[\text{dist}_X(\hat{x}_t)] \leq E[||\hat{w}_t - \hat{x}_t||] \leq \sqrt{E[||\hat{w}_t - \hat{x}_t||^2]} \leq O\left(\frac{1}{\sqrt{b_p^N t}}\right).
\]

Since $\hat{x}_t \in Y$ and $\hat{w}_t \in X \subset Y$, by the subgradient boundedness of $f$ on $Y$, it follows that

\[
E[|f(\hat{x}_t) - f(\hat{w}_t)|] \leq M_f E[||\hat{x}_t - \hat{w}_t||] \leq O\left(\frac{1}{\sqrt{b_p^N t}}\right),
\]

which combined with $E[f(\hat{w}_t) - f^*] \leq O\left(\frac{1}{t} + \frac{1}{\sqrt{b_p^N t}}\right)$, yields also the following convergence rate for suboptimality

\[
E[|f(\hat{x}_t) - f^*|] \leq O\left(\frac{1}{t} + \frac{1}{\sqrt{b_p^N t}}\right),
\]

which proves our theorem. \(\square\)

We observe that the convergence estimate for the feasibility violation depends explicitly on the minibatch size $N$ via the key parameter $L_N$. For the optimal stepsize $\beta = 1/L_N$ we get $q_N = 1/cM_f^2 L_N$ and $b_p^N = 1/(cM_f^2 L_N - 1)$. Hence, $b_p^N$ is large provided that $L_N \ll 1$ (small). Note that if $L_N = 1$, then $b_p^N$ does not depend on $N$ and hence complexity does not improve with minibatch size.
N. However, as long as $L_N < 1$ (and it can be also the case that $L_N \sim 0$), then $b^N_p$ becomes large, which shows that minibatching improves complexity. To the best of our knowledge, this is the first time that a subgradient method with random minibatch feasibility updates is shown to be better than its non-minibatch variant. We have identified $L_N$ as the key quantity determining whether minibatching helps ($L_N < 1$) or not ($L_N = 1$), and how much (the smaller $L_N$, the more it helps). Note also that the suboptimality estimate contains a term which does not depend on the minibatch size $N$ as it happens for feasibility violation estimate. This is natural, since the minibatch feasibility steps have no effect on the minimization step of the objective function.

Remark 2 Note that the convergence rates $O\left(\frac{1}{t}\right)$ for feasibility and suboptimality are known to be optimal for the stochastic subgradient method for solving the optimization problem \[\text{(1)}\] under Assumption 1, see \[\text{[17,18]}\]. Moreover, the iterative process \[\text{(2)}\] does not require knowledge of the subgradient norm bounds $M_f$ and $M_g$ from Assumption 1 nor the constant $c$ from Assumption 2. These values are only affecting the constants in the convergence rates, they are not needed for the stepsize selection. The stepsize $\alpha_k$ requires only knowledge of some estimate of the restricted strong convexity constant $\mu$. Moreover, since $L_N \leq 1$, we can use e.g., stepsizes $\beta \in (0, 2) \subseteq (0, 2/L_N)$. Of course, a larger stepsizes $\beta$ leads to a faster convergence. Hence, if $L_N < 1$ and it can be computed, then we should choose an extrapolated steplength $\beta = (2 - \delta)/L_N$ for some $\delta \in (0, 1)$ small. When $L_N$ cannot be computed explicitly, we propose to approximate it online with $L^i_N$, and use at each iteration an adaptive extrapolated stepsize $\beta_k$ of the form:

$$\beta_k = \frac{2 - \delta}{L^k_N} = \frac{2 - \delta}{L_N} \frac{1}{N} \sum_{i=1}^{N} \left(\frac{g^+_i(v_k))}{d^i_k}\right)^2 \left\|\frac{1}{N} \sum_{j=1}^{N} \frac{g^+_i(v_k)}{d^j_k} d^j_k\right\|^2,$$

(26)

for some $\delta \in (0, 1)$ sufficiently small. The convergence rate of algorithm \[\text{(2)}\] for this adaptive choice of the stepsize $\beta_k$ will be analyzed in our future work.

2.3 Example of functional constraints having $L_N < 1$

Let us recall the definition of the parameters $L^k_N$ and $L_N$ from \[\text{[13]}\]:

$$L^k_N = \left\|\frac{1}{N} \sum_{i=1}^{N} g^+_i(v_k) \right\|^2 / \left\|\frac{1}{N} \sum_{i=1}^{N} g^+_i(v_k) \frac{d^i_k}{\|d^i_k\|^2}\right\|^2$$

and $L_N = \max_{k \geq 0} L^k_N$.

From Jensen’s inequality we have $L^k_N \leq 1$ and consequently $L_N \leq 1$. On the other hand, Theorem 2 shows that $L_N \ll 1$ is beneficial for a subgradient scheme with minibatch feasibility updates. In this section we provide an example of functional constraints $g_{\omega}$ for which $L_N < 1$. Let us consider $m$ linear inequality constraints for the convex problem \[\text{(1)}\]:

$$g_{\omega}(x) = a^T_{\omega} x + b^T_{\omega} \leq 0 \quad \forall \omega \in \mathcal{A} = \{1, 2, \cdots, m\}.$$
Without loss of generality we assume \( \|a_\omega\| = 1 \) for all \( \omega \). Let us define the matrix \( A = [a_1 \cdots a_m]^T \) and the subset of indexes selected at the current iteration \( J_k = \{ \omega_k^1, \ldots, \omega_k^N \} \subset \mathcal{A} \). We also denote \( J_k^+ = \{ \omega \in J_k : a_\omega^T v_k + b_\omega > 0 \} \) and denote \( A_{J_k}^+ \) the submatrix of \( A \) having the rows indexed in the set \( J_k^+ \). With these notations and using that \( \|a_\omega\| = 1 \) for all \( \omega \), then \( L_N^k \) can be written explicitly as (assuming that \( |J_k^+| \geq 1 \):

\[
L_N^k = \left[ \frac{1}{N} \sum_{\omega \in J_k^+} \frac{g_\omega^+(v_k)}{\|d_k^\omega\|^2} d_k^\omega \right]^2 / \left[ \frac{1}{N} \sum_{\omega \in J_k^+} \frac{(g_\omega^+(v_k))^2}{\|d_k^\omega\|^2} \right]
\]

\[
= \left[ \sum_{\omega \in J_k^+} (a_\omega v_k + b_\omega)a_\omega \right] / \left[ \sum_{\omega \in J_k^+} (a_\omega v_k + b_\omega)^2 \right]
\]

\[
= \left[ A_k^T (A_k^+ v_k + b_k^+) \right]^2 / \left[ N \|A_k^+ v_k + b_k^+\|^2 \right]
\]

\[
\leq \frac{\lambda_{\max}(A_k^+ A_k^T)}{N} \leq \frac{\lambda_{\max}(A_J A_J^T)}{N} < \frac{\lambda_{\max}(A_J A_J^T)}{N} = 1 \quad \forall k,
\]

where the first inequality follows from the definition of the maximal eigenvalue \( \lambda_{\max} \) of a matrix, the second inequality follows from the fact that \( J_k^+ \subseteq J_k \), and the third inequality holds strictly provided that the submatrix \( A_{J_k}^+ \) has at least rank two. In conclusion, if the matrix \( A \) has e.g. full row rank and consider a sampling of \( J_k \) based on a given probability \( P \), then \( L_N \) satisfies:

\[
L_N = \max_{J \sim P, |J| = N, J \neq J} \frac{\lambda_{\max}(A_J A_J^T)}{N} < 1. \tag{27}
\]

Note that for particular sampling rules we can compute \( L_N \) efficiently, such as when we consider a uniform distribution over a fixed partition of \( \mathcal{A} = \bigcup_{i=1}^N J_i \) of equal size. The reader may find other examples of functional constraints satisfying \( L_N < 1 \) and we believe that this paper opens a window of opportunities for algorithmic research in this direction.

### 3 Sequential random minibatch projections algorithm

In this section we consider a sequential variant of the algorithm defined in terms of the following iterative process:

\[
v_k = \Pi_Y [x_{k-1} - \alpha_{k-1} s_f(x_{k-1})], \tag{28a}
\]

\[
z_{k}^0 = v_k, \ z_{k}^i = \Pi_Y \left[ z_{k-1}^i - \beta \frac{g^+_k(z_{k-1}^{i-1})}{\|d_k^i\|^2} d_k^i \right] \quad \text{for } i = 1, \ldots, N, \tag{28b}
\]

\[
x_k = z_k^N. \tag{28c}
\]
This method takes, as for the parallel variant, one subgradient step for the objective function, followed by \( N \) sequential feasibility updates. As before, the vector \( d_k^i \) is chosen as \( d_k^i \in \partial g_{\omega_k^i}(z_k^{i-1}) \) if \( g_{\omega_k^i}(v_k) > 0 \), and \( d_k^i = d \) for some \( d \neq 0 \) if \( g_{\omega_k^i}(z_k^{i-1}) = 0 \). Note that in this variant, the feasibility updates use the projection on \( Y \) in order to confine the intermittent iterates \( z_k^i \) and \( x_k \) to the set \( Y \), where \( g_{\omega} \)'s and \( f \) (for the last step) are assumed to have uniformly bounded subgradients.

In this section we analyze the convergence properties of this new algorithm \( \text{(28)} \). Given \( x_{k-1} \), the update of \( v_k \) is the same as in the parallel method \( \text{(2)} \), thus Lemma \( \text{2} \) still applies here. We need an analog of Lemma \( \text{5} \).

**Lemma 7** Let Assumptions \( \text{1(a)} \) and \( \text{1(c)} \) hold. Let \( x_k \) be generated by algorithm \( \text{(28)} \) with \( \beta \in (0, 2) \). Then, the following relations are valid:

\[
\text{dist}^2(z_k^i, X) \leq \text{dist}^2(z_k^{i-1}, X) - \frac{\beta(2 - \beta)}{M_g^2} (g_{\omega_k^i}(z_k^{i-1}))^2 \quad \text{for all} \quad i = 1, \ldots, N,
\]

\[
\|x_k - y\|^2 \leq \|v_k - y\|^2 - \frac{\beta(2 - \beta)}{M_g^2} \sum_{i=1}^{N} (g_{\omega_k^i}(z_k^{i-1}))^2 \quad \text{for all} \quad y \in X,
\]

\[
\text{dist}^2(x_k, X) \leq \text{dist}^2(v_k, X) - \frac{\beta(2 - \beta)}{M_g^2} \sum_{i=1}^{N} (g_{\omega_k^i}(z_k^{i-1}))^2 \quad \text{for all} \quad k \geq 1.
\]

**Proof** We start with the definition of \( z_k^i \) in \( \text{(28)} \) and Lemma \( \text{3} \) with \( Z = Y \). Thus, we obtain for all \( y \in X \) (which satisfies \( g_{\omega_k^i}(y) = 0 \) for any realization of \( \omega_k^i \)) and for all \( i = 1, \ldots, N \),

\[
\|z_k^i - y\|^2 \leq \|z_k^{i-1} - y\|^2 - \beta(2 - \beta) \frac{(g_{\omega_k^i}(z_k^{i-1}))^2}{\|d_k^i\|^2}.
\]

By using \( \|d_k^i\|^2 \leq M_g^2 \), we have for all \( i = 1, \ldots, N \),

\[
\|z_k^i - y\|^2 \leq \|z_k^{i-1} - y\|^2 - \frac{\beta(2 - \beta)}{M_g^2} (g_{\omega_k^i}(z_k^{i-1}))^2.
\]

The distance relation for \( z \)-iterates follows by taking the minimum over \( y \in X \) on both sides of inequality \( \text{(29)} \). By summing relations \( \text{(29)} \) over \( i = 1, \ldots, N \), and by using \( z_k^0 = v_k \) and \( z_k^N = x_k \), we obtain for any \( y \in X \),

\[
\|x_k - y\|^2 \leq \|v_k - y\|^2 - \frac{\beta(2 - \beta)}{M_g^2} \sum_{i=1}^{N} (g_{\omega_k^i}(z_k^{i-1}))^2.
\]

The distance relation follows by taking the minimum over \( y \in X \) on both sides of the preceding inequality. \( \square \)
Taking $\rho = 1/2$ in Lemma 2 we get:

$$\|v_k+1 - x^*\|^2 + \alpha_k (f(HX[x_k]) - f^*) \leq \left(1 - \frac{\alpha_k \mu}{2}\right) \|x_k - x^*\|^2 + \alpha_k M_f \|HX[x_k] - x_k\| + \alpha_k^2 M_f^2,$$

and using the inequality for $\|x_k - y\|^2$ from Lemma 7 in $y = x^*$, yields:

$$\|v_k+1 - x^*\|^2 + \alpha_k (f(HX[x_k]) - f^*) \leq \left(1 - \frac{\alpha_k \mu}{2}\right) \|v_k - x^*\|^2$$

(30)

$$- \left(1 - \frac{\alpha_k \mu}{2}\right) \frac{\beta(2 - \beta)}{M_f^2} \sum_{i=1}^N (g^+ \omega_k(z_k^{i-1}))^2 + \alpha_k M_f \|HX[x_k] - x_k\| + \alpha_k^2 M_f^2.$$

Taking the conditional expectation on $\mathcal{F}_{k-1}$ and $z_k^{i-1}$, and using Assumption 2 give

$$E\left[g^+ \omega_k(z_k^{i-1})^2 | \mathcal{F}_{k-1}, z_k^{i-1}\right] \geq \frac{1}{\epsilon} \text{dist}(z_k^{i-1}, X).$$

Using the iterated expectation rule, we obtain

$$E\left[g^+ \omega_k(z_k^{i-1})^2 | \mathcal{F}_{k-1}\right] = E\left[E\left[g^+ \omega_k(z_k^{i-1})^2 | \mathcal{F}_{k-1}, z_k^{i-1}\right] | \mathcal{F}_{k-1}\right] \geq \frac{1}{\epsilon} E\left[\text{dist}(z_k^{i-1}, X) | \mathcal{F}_{k-1}\right],$$

(31)

which, when combined with the distance relation of Lemma 4 gives for all $i = 1, \ldots, N$

$$E\left[\text{dist}^2(z_k^{i}, X) | \mathcal{F}_{k-1}\right] \leq \left(1 - \frac{\beta(2 - \beta)}{\epsilon M_f^2}\right) E\left[\text{dist}^2(z_k^{i-1}, X) | \mathcal{F}_{k-1}\right].$$

Hence, using the definition of $x_k$, i.e., $x_k = z_k^N$, and letting $q = \frac{\beta(2 - \beta)}{\epsilon M_f^2} \in (0, 1)$ (since we assume $\epsilon M_f^2 > 1$ and $\beta \in (0, 2)$), we have for all $i = 1, \ldots, N$,

$$E\left[\text{dist}^2(x_k, X) | \mathcal{F}_{k-1}\right] \leq (1 - q)^{N-i+1} E\left[\text{dist}^2(z_k^{i-1}, X) | \mathcal{F}_{k-1}\right],$$

implying that for all $i = 1, \ldots, N$,

$$E\left[\text{dist}^2(z_k^{i-1}, X) | \mathcal{F}_{k-1}\right] \geq \frac{1}{(1 - q)^{N-i+1}} E\left[\text{dist}^2(x_k, X) | \mathcal{F}_{k-1}\right].$$

(32)

From (31) and (32) for all $i = 1, \ldots, N$,

$$E\left[g^+ \omega_k(z_k^{i-1})^2 | \mathcal{F}_{k-1}\right] \geq \frac{1}{c (1 - q)^{N-i+1}} E\left[\text{dist}^2(x_k, X) | \mathcal{F}_{k-1}\right].$$

By summing over $i$

$$\sum_{i=1}^N E\left[g^+ \omega_k(z_k^{i-1})^2 | \mathcal{F}_{k-1}\right] \geq \frac{1}{c} \left(\sum_{i=1}^N \frac{1}{(1 - q)^{N-i+1}}\right) E\left[\text{dist}^2(x_k, X) | \mathcal{F}_{k-1}\right].$$
However,
\[
\sum_{i=1}^{N} \frac{1}{(1-q)^{N-i+1}} = \frac{1}{(1-q)^N} \sum_{i=1}^{N} (1-q)^i = \frac{1}{q(1-q)^N}.
\]

Finally, we get
\[
\sum_{i=1}^{N} E\left[ g_{\omega_k}(z_k^{i-1})^2 \mid \mathcal{F}_{k-1} \right] \geq \frac{1}{c} \frac{(1-(1-q)^N)}{q(1-q)^N} E\left[ \text{dist}^2(x_k, X) \mid \mathcal{F}_{k-1} \right]
\]
and consequently
\[
\frac{\beta(2-\beta)}{M^2} \sum_{i=1}^{N} E\left[ (g_{\omega_k}(z_k^{i-1}))^2 \mid \mathcal{F}_{k-1} \right] \geq \frac{\beta(2-\beta)}{cM^2} \frac{(1-(1-q)^N)}{q(1-q)^N} E\left[ \text{dist}^2(x_k, X) \mid \mathcal{F}_{k-1} \right]
\]
\[
= \frac{q}{q(1-q)^N} \frac{(1-(1-q)^N)}{q(1-q)^N} E\left[ \text{dist}^2(x_k, X) \mid \mathcal{F}_{k-1} \right] \geq \frac{(1-(1-q)^N)}{(1-q)^N} E\left[ \text{dist}^2(x_k, X) \mid \mathcal{F}_{k-1} \right] \geq (1-q)^{-N} - 1 \text{ E}[\text{dist}^2(x_k, X) \mid \mathcal{F}_{k-1}].
\]

Let us denote \( b_N^* = (1-q)^{-N} - 1 \). It is clear that \( b_N^* \rightarrow \infty \) as \( N \rightarrow \infty \). Taking expectation in (30) and using the previous inequality we get an analog of Lemma 6:
\[
E[\|v_{k+1} - x^*\|^2 \mid \mathcal{F}_{k-1}] + \alpha_k E[(f(HX[x_k]) - f^*) \mid \mathcal{F}_{k-1}] \leq (1 - \frac{\alpha_k \mu}{2}) \|v_k - x^*\|^2 - \left( (1 - \frac{\alpha_k \mu}{2}) b_N^* - \frac{\eta}{2} \right) E[\text{dist}^2(x_k, X) \mid \mathcal{F}_{k-1}]
\]
\[+ \beta_k \left( 1 + \frac{1}{2\eta} \right) M^2,
\]
for any \( \eta > 0 \). Let us consider the same stepsize as for the parallel scheme, i.e. \( \alpha_k = \frac{2}{\mu} \gamma_k \), choose \( \eta = (1 - \frac{\alpha_k \mu}{2}) b_N^* > 0 \), and take the full expectation, to get the following recurrence (analog to Theorem 1):
\[
E[\|v_{k+1} - x^*\|^2] + \frac{2}{\mu} \gamma_k E[(f(HX[x_k]) - f^*)] \leq (1 - \gamma_k) E[\|v_k - x^*\|^2]
\]
\[= \frac{1}{2} (1 - \gamma_k) b_N^* E[\text{dist}^2(x_k, X)] + \frac{4}{\mu^2} \gamma_k^2 \left( 1 + \frac{1}{(2 - 2\gamma_k)b_N} \right) M^2.
\]

Using now \( \gamma_k = \frac{2}{\mu} \), then \( 1 - \gamma_k \geq \frac{1}{4} \) and we get:
\[
E[\|v_{k+1} - x^*\|^2] + \frac{2}{\mu} \gamma_k E[(f(HX[x_k]) - f^*)] \leq (1 - \gamma_k) E[\|v_k - x^*\|^2]
\]
\[= \frac{1}{6} b_N^* E[\text{dist}^2(x_k, X)] + \frac{4}{\mu^2} \gamma_k^2 \left( 1 + \frac{2}{b_N^*} \right) M^2.
\]
Since, $\frac{1}{\gamma_k} \leq \frac{1}{\gamma_{k-1}}$ for all $k \geq 1$, dividing (33) by $\gamma_k^2$ and using the preceding inequality we have for all $k \geq 1$:

$$
\gamma_k^{-2} E[\|v_{k+1} - x^*\|^2] + \frac{2}{\mu} \gamma_k^{-1} E[(f(\Pi_X[x_k])) - f^*)]
\leq \frac{b_N}{6} E[\text{dist}^2(x_k, X)] \leq \gamma_k^{-2} E[\|v_k - x^*\|^2] + \frac{4}{\mu^2} \left(1 + \frac{2}{b_N}\right) M_f^2.
$$

Summing these over $k = 1, \ldots, t$, for some $t > 0$, we obtain the following recurrence relation for the algorithm (28):

$$
\gamma_t^{-2} E[\|v_{t+1} - x^*\|^2] + \frac{2}{\mu} \sum_{k=1}^t \gamma_k^{-1} E[(f(\Pi_X[x_k])) - f^*)]
\leq \gamma_0^{-2} E[\|v_1 - x^*\|^2] + \frac{4}{\mu^2} \left(1 + \frac{2}{b_N}\right) M_f^2.
$$

Using the same definition for the weighted averages $\hat{w}_t$ and $\hat{x}_t$ from (23) and $\gamma_k = \frac{2}{k+1}$ in (34), we get the main recurrence for the sequential variant (28):

$$
\frac{(t+1)^2}{4} E[\|v_{t+1} - x^*\|^2] + \frac{S_t}{(t+1)\mu} E[(f(\hat{w}_t)) - f^*)] + \frac{b_N^2 S_t}{24} E[\|\hat{w}_t - \hat{x}_t\|^2]
\leq \frac{1}{4} E[\|v_1 - x^*\|^2] + \frac{4t}{\mu^2} \left(1 + \frac{2}{b_N}\right) M_f^2.
$$

Next theorem summarizes the convergence rates that follow from the recurrence relation (35) of the sequential algorithm (28).

**Theorem 3** Let Assumption 1 and Assumption 2 hold and the stepsizes $\beta \in (0, 2)$ and $\alpha_k = \frac{4}{\mu(k+1)}$. Let also $q = \frac{\beta(2-\beta)}{cM_f^2} < 1$ and $b_N = (1-q)^{-N} - 1$. Then, the following sublinear rates for suboptimality and feasibility violation hold for the average sequence $\hat{x}_t$ from (23) generated by the sequential algorithm (28):

$$
E[(f(\hat{x}_t)) - f^*)] \leq O\left(\frac{1}{t + \frac{1}{\sqrt{b_N} t}}\right), \quad E[\text{dist}_X(\hat{x}_t)] \leq O\left(\frac{1}{\sqrt{b_N} t}\right).
$$

**Proof** Denoting the same average sequences $\hat{w}_t$ and $\hat{x}_t$ as in (23), we get the following convergence rates (omitting the constants but keeping the terms depending on $b_N$):

$$
E[f(\hat{w}_t) - f^*)] \leq O\left(\frac{1}{t + \frac{1}{b_N t}}\right), \quad E[\|\hat{w}_t - \hat{x}_t\|^2] \leq O\left(\frac{1}{b_N t^2}\right).
$$

Hence, we get the following convergence rate for the feasibility violation of the constraints that depends explicitly on the minibatch size $N$ via the term $b_N$:

$$
E[\text{dist}_X^2(\hat{x}_t)] \leq O\left(\frac{1}{b_N t^2}\right).
$$
Using the same reasoning as in the proof of Theorem 2, we also get the following convergence rate for suboptimality:

$$E[|f(\hat{x}_t) - f^*|] \leq O\left(\frac{1}{t} + \frac{1}{\sqrt{b_N t}}\right),$$

which proves the statements of the theorem.

We observe that also for the sequential algorithm (28), the convergence estimate for the feasibility violation depends explicitly on the minibatch size $N$ via the term $b_N$ (recall that $b_N \to \infty$ as $N \to \infty$). Since $b_N$ is an increasing sequence in $N$, it follows that the larger is the minibatch size $N$ the better is also the complexity of the sequential algorithm (28) in terms of constraints feasibility. In conclusion, for the sequential variant our rates prove that minibatching always helps and the feasibility estimate depends exponentially on the minibatch size. On the other hand, the suboptimality estimate contains a term which does not depend on the minibatch size $N$ as it happens for feasibility violation estimate.

4 Conclusions

In this paper we have considered non-smooth convex optimization problems with (possibly) infinite intersection of constraints. For solving this general class of convex problems we have proposed subgradient algorithms with random minibatch feasibility steps. At each iteration, our algorithms take first a step for minimizing the objective function and then a subsequent step minimizing the feasibility violation of the observed minibatch of constraints. The feasibility updates were performed based on either parallel or sequential random observations of several constraint components. For a diminishing stepsize and for restricted strongly convex objective functions, we have proved sublinear convergence rates for the expected distances of the weighted averages of the iterates from the constraint set, as well as for the expected suboptimality of the function values along the weighted averages. Our convergence rates are optimal for subgradient methods with random feasibility steps for solving this class of non-smooth convex problems. Moreover, the rates depend explicitly on the minibatch size. From our knowledge, this work is the first proving that subgradient methods with random minibatch feasibility updates have better complexity than their non-minibatch variants.

Acknowledgements This research was supported by the National Science Foundation under CAREER grants CMMI 07-42538 and by the Executive Agency for Higher Education, Research and Innovation Funding (UEFISCDI), Romania, PNII-P4-PCE-2016-0731, project ScaleFreeNet, no. 39/2017.
References

1. D. Blatt and A.O. Hero, *Energy based sensor network source localization via projection onto convex sets*, IEEE Transactions on Signal Processing, 54(9): 3614–3619, 2006.
2. H. Bauschke and J. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Review 38(3): 367-376, 1996.
3. R.I. Bot and C. Hendrich, *A double smoothing technique for solving unconstrained non-differentiable convex optimization problems*, Computational Optimization and Applications, 54 (2): 239–262, 2013.
4. R.I. Bot and T. Hein, *Iterative regularization with general penalty term - theory and application to L1 and TV regularization*, Inverse Problems, 28, 2012.
5. J. Burke and M. Ferris, *Weak sharp minima in mathematical programming*, SIAM Journal of Control and Optimization, 31(6): 1340-1359, 1993.
6. P. Bianchi, W. Hachem, and A. Salim, *A constant step forward-backward algorithm involving random maximal monotone operators*, arXiv:1702.04144, 2017.
7. D.P. Bertsekas, *Incremental proximal methods for large scale convex optimization*, Mathematical Programming, 129(2). 2011.
8. K. Keys, H. Zhou and K. Lange, *Proximal Distance Algorithms: Theory and Practice*, ArXiv e-prints, 2016.
9. A. Kundu, F. Bach, and C. Bhattacharya, *Convex optimization over intersection of simple sets: improved convergence rate guarantees via an exact penalty approach*, arxiv.org/abs/1710.06465, 2017.
10. A. Lewis and J. Pang, *Error bounds for convex inequality systems*, in Generalized Convexity, Generalized Monotonicity (J. Crouzeix, J. Martinez-Legaz, and M. Volle, eds.), Cambridge University Press, 75–110, 1998.
11. E. Moulines and F.R. Bach, *Non-asymptotic analysis of stochastic approximation algorithms for machine learning*, Advances in Neural Information Process. Syst., 2011.
12. I. Necoara, Yu. Nesterov and F. Glineur, *Linear convergence of first order methods for non-strongly convex optimization*, SIAM Journal on Optimization, doi.org/10.1137/10107-018-1232-1, 2018.
13. I. Necoara, P. Richtarik and A. Patrascu, *Randomized projection methods for convex feasibility problems*, submitted to SIAM Journal on Optimization, 2018.
14. A. Nedić, *Random projection algorithms for convex set intersection Problems*, IEEE Conference on Decision and Control, 7655–7660, 2010.
15. A. Nedić, *Random algorithms for convex minimization problems*, Math. Program., 129(2): 225-273, 2011.
16. A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. on Optimization, 19(4):1574–1609, 2009.
17. A. Nemirovski and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*, John Wiley & Sons, 1983.
18. Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer, Boston, 2004.
19. H. Ouyang and A. Gray, *Stochastic smoothing for nonsmooth minimizations: Accelerating SGD by exploiting structure*, arXiv:1205.4481, 2012.
20. A. Patrascu and I. Necoara, *Nonasymptotic convergence of stochastic proximal point algorithms for constrained convex optimization*, Journal of Machine Learning Research, 18(198): 1–42, 2018.
21. B.T. Polyak, *Minimization of unsmooth functionals*, U.S.S.R. Comput. Math. and Math. Phys. 9: 14–29, 1969.
22. B.T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM Journal on Control and Optimization, 30(4): 838–855, 1992.
23. R. Shefi and M. Teboulle, *Rate of Convergence Analysis of Decomposition Methods Based on the Proximal Method of Multipliers for Convex Minimization*, SIAM Journal on Optimization, 24(1): 269–297, 2014.
24. Q. Tran-Dinh, O. Fecoq, and V. Cevher, *A smooth primal-dual optimization framework for nonsmooth composite convex minimization*, SIAM Journal on Optimization, 28 (1): 96–134, 2018.
25. V. Vapnik, *Statistical learning theory*, John Wiley, 1998.
26. M. Wang, Y. Chen, J. Liu and Y. Gu, *Random multiconstraint projection: stochastic gradient methods for convex optimization with many constraints*, arXiv:1511.03760, 2015.