Convergence of Stochastic Proximal Gradient Algorithm

Lorenzo Rosasco · Silvia Villa · Băng Công Vụ

Abstract
We study the extension of the proximal gradient algorithm where only a stochastic gradient estimate is available and a relaxation step is allowed. We establish convergence rates for function values in the convex case, as well as almost sure convergence and convergence rates for the iterates under further convexity assumptions. Our analysis avoid averaging the iterates and error summability assumptions which might not be satisfied in applications, e.g. in machine learning. Our proofing technique extends classical ideas from the analysis of deterministic proximal gradient algorithms.

Keywords Proximal methods · Forward–backward splitting algorithm · Stochastic optimization

1 Introduction
First order methods are widely applied to solve optimization problems in a variety of areas including machine learning [2,32] and signal processing [15]. In particular, proximal gradient algorithms, also known as forward–backward splitting algorithms [4], are natural in the common situation where the function to be minimized is the sum of a data fitting term and a regularizer. These approaches separate the contribution of each component: at each iteration the proximal operator defined by the
nonsmooth term is applied to the gradient descent step for the smooth term. In practice, it is relevant to consider situations where these operations cannot be performed exactly. For example, the case where the proximal operator is known only up-to an error have been considered see [43,50] and references therein. The other example, of interest for this paper, is when only stochastic estimates of the gradients are available. This setting is relevant in large scale statistical learning [10] and more generally in stochastic approximation [27]. The study of stochastic gradient methods goes back to [26,39] and this algorithm has been studied extensively for smooth, strongly convex objective functions [5,19,20,27,34,36]. In this context, a classic idea is averaging the iterates to allow for longer step-sizes in [35,37]. Stochastic gradient methods have also been extensively studied in the context of online learning [11]. Here, the analysis of the so called regret allows to derive convergence rates for the average iterates through a proofing technique known as online-to-batch conversion, see e.g. [44] and references therein. For nonsmooth problems, stochastic extension of projected subgradients methods [33,46,48] and the mirror descent algorithm [35] have been considered. Compared to these approaches, the advantage of proximal methods is on the one hand, that the proximal step enforces structure in the solution, for example sparsity, and, on the other hand, that no subdifferentiability of the convex, nonsmooth term is needed. The first results considering proximal methods with stochastic gradients are based on the aforementioned online-to-batch approach [18]. Here, $O(1/\sqrt{n})$ convergence rates in expectation for the function value of the average iterate are given. These results require the iterates and the stochastic approximations to be bounded, and the function to be Lipschitz continuous. More recently, basic and accelerated stochastic proximal gradient methods were studied in [1] in the convex case, requiring a sort of Lipschitz continuity of the nonsmooth component and boundedness of the domain of the objective function. Bounds in expectation on the convergence of the function values of the averaged iterates are provided and almost sure convergence of the iterates is proved. Here, the strongly convex case is not considered. Convergence of the iterates of stochastic proximal gradient has been also obtained using the stochastic fixed point algorithms [12,13], but only under summability assumptions on the errors of the stochastic estimates. Stochastic forward–backward algorithms for monotone inclusions have also been studied in [8] using techniques from dynamical systems. Finally, we note that stochastic variants of accelerated proximal gradient descent have been studied [22,25,28,29]. In contrast to the deterministic case, these results show that accelerated methods do not improve convergence rates but have only a second order effect.

In this paper, we analyze a natural stochastic extension of proximal gradient descent, where we allow for a relaxation step. First, we derive a worst case $O(1/\sqrt{n})$ convergence rate for the expected function values in the convex case. Second, assuming strong convexity, we prove almost sure convergence and non asymptotic bounds in expectation for the iterates. Our analysis extends to the composite setting the one in [3], which holds in the smooth case. Compared to previous results, as those in [1], our study focuses on the last, rather than the average, iterate and does not require boundedness of the iterates themselves. Thus, our paper complements previous results, and in particular those in [1], which do not address the strongly convex setting, and that require either boundedness of the iterates or Lipschitz continuity of the nonsmooth
component. We note that, avoiding averaging is of importance to preserve structural properties of the solution such as sparsity [31, 51] and also to prevent a possible negative impact on the convergence rate [38, 47]. Further, our analysis does not require summability of the errors or boundedness of the variance of the stochastic estimates, made in the previously mentioned papers [12, 13]. The latter assumption is replaced by a more general condition, allowing to deal with multiplicative noise, see [24]. This is relevant, since this is the usual setting e.g. in machine learning, where summability assumptions are hardly ever satisfied. We assume that the stochastic estimate of the gradient is unbiased. This is an assumption which, on the one hand may be not satisfied in challenging machine learning problems, but on the other hand is very often satisfied [10]. We note that this assumption is not made in [1, 12, 13]. Our results derive from a stochastic extension of the classic analysis of proximal methods in a deterministic setting. As shown in [40, 41] the ideas in the paper can be further generalized to find zeros of sums of maximal monotone operators.

The rest of the paper is organized as follows: In Sect. 2 we introduce composite optimization and the stochastic proximal gradient algorithm, along with some relevant special cases. In Sect. 3, we study convergence in expectation of the function values in the strongly convex case, generalizing the results in [3, Sect. 3] to the nonsmooth case. Moreover, we state the main results about convergence in expectation and almost surely of the iterates, that we prove in Appendix. Section 4 describes some numerical tests comparing the stochastic projected gradient algorithm with state of the art stochastic first order methods.

**Notation and basic definitions** Throughout, \((\mathcal{E}, \mathcal{A}, P)\) is a probability space, \(\mathbb{N}^* = \mathbb{N} \setminus \{0\}\), and \(\mathcal{H}\) is a real separable Hilbert space. We use the notation \(\langle \cdot | \cdot \rangle\) and \(\| \cdot \|\) for the scalar product and the associated norm in \(\mathcal{H}\). The symbols \(\rightarrow\) and \(\rightharpoonup\) denote, respectively, weak and strong convergence. The class of lower semicontinuous convex functions \(f : \mathcal{H} \rightarrow [-\infty, +\infty]\) such that \(\text{dom } f = \{ x \in \mathcal{H} | f(x) < +\infty \} \neq \emptyset\), is denoted by \(\Gamma_0(\mathcal{H})\). The proximity operator of \(f \in \Gamma_0(\mathcal{H})\) is

\[
\text{prox}_f : \mathcal{H} \rightarrow \mathcal{H}, \quad \text{prox}_f(w) = \arg\min_{v \in \mathcal{H}} f(v) + \frac{1}{2} \| w - v \|^2. \tag{1.1}
\]

Throughout this paper, we assume implicitly that the closed-form expressions of the proximity operators to be available. We refer to [4, 14] for the closed-form expression of a wide class of functions, see [32] for examples in machine learning. Given a random variable \(X\), we denote by \(\mathbb{E}[X]\) its expected value, and by \(\sigma(X)\) the \(\sigma\)-field generated by \(X\). The conditional expectation of \(X\) given a \(\sigma\)-algebra \(\mathcal{A} \subset \mathcal{F}\) is denoted by \(\mathbb{E}[X|\mathcal{A}]\). The conditional expectation of \(X\) given \(Y\) is denoted by \(\mathbb{E}[X|Y]\). A filtration of \(\mathcal{A}\) is an increasing sequence \((\mathcal{A}_n)_{n \in \mathbb{N}^*}\) of sub-\(\sigma\)-algebras of \(\mathcal{A}\). A \(\mathcal{H}\)-valued random process is a sequence of random variables \((X_n)_{n \in \mathbb{N}^*}\) taking values in \(\mathcal{H}\). The shorthand notation ‘a.s.’ stands for ‘almost sure’. We denote by \(\ell_+^1(\mathbb{N})\) the set of summable sequences in \([0, +\infty[\).
2 Problem Setting and Examples

We begin introducing the composite convex optimization problem and the stochastic proximal method we study, and then discuss some special cases of the framework we consider.

Composite optimization problems are defined as the problem of minimizing the sum of a smooth convex function and a possibly nonsmooth convex function. Here we assume that the latter is proximable, that is the proximity operator (1.1) is available in closed form or can be easily computed.

Problem 2.1 Let $G \in \Gamma_0(\mathcal{H})$, let $\beta \in ]0, \infty[$, and let $H : \mathcal{H} \to \mathbb{R}$ be convex and differentiable, with a $\beta$-Lipschitz continuous gradient. The problem is to

$$\min_{w \in \mathcal{H}} \Phi(w) = H(w) + G(w),$$

(2.1)

under the assumption that the set of solutions to (2.1) is non-empty.

Problems with this structure has been recently extensively studied in convex optimization. The class of splitting methods, which decouple the contribution of the smooth term and the nonsmooth one, received a lot of attention [4]. Within this class, in this paper we study the following stochastic proximal gradient (SPG) algorithm.

Algorithm 2.2 (SPG) Let $(\gamma_n)_{n \in \mathbb{N}^*}$ be a strictly positive sequence, let $(\lambda_n)_{n \in \mathbb{N}^*}$ be a sequence in $[0, 1]$, and let $(\mathcal{G}_n)_{n \in \mathbb{N}^*}$ be a $\mathcal{H}$-valued random process such that $(\forall n \in \mathbb{N}^*)$ $E[\|\mathcal{G}_n\|^2] < +\infty$. Fix $w_1$ a $\mathcal{H}$-valued integrable vector with $E[\|w_1\|^2] < +\infty$ and set

$$\begin{align*}
(z_n &= w_n - \gamma_n \mathcal{G}_n \\
y_n &= \text{prox}_{\gamma_n G} z_n \\
w_{n+1} &= (1 - \lambda_n) w_n + \lambda_n y_n.
\end{align*}$$

(2.2)

Algorithm 2.2 is a stochastic version of the forward–backward algorithm [15], where we replace the exact gradient by a stochastic estimate. More specifically, if $\mathcal{G}_n = \nabla H(w_n)$, our algorithm reduces to the one in [15]. A stochastic forward–backward splitting (FOBOS) was firstly proposed in [18] for minimizing the sum of two functions where one of them is proximable, and the other is convex and subdifferentiable. Algorithm 2.2 generalizes the FOBOS algorithm, by including a relaxation step (the third one in (2.2)), while assuming the first component in (2.1) to be smooth.

As it is the standard, to ensure convergence of the proposed algorithm, we need additional conditions on the random process $(\mathcal{G}_n)_{n \in \mathbb{N}^*}$ as well as on the sequence of step-sizes $(\gamma_n)_{n \in \mathbb{N}^*}$.  

Condition 2.3 The following conditions will be considered for the filtration $(\mathcal{A}_n)_{n \in \mathbb{N}^*}$ with $\mathcal{A}_n = \sigma(w_1, \ldots, w_n)$.

(A1) For every $n \in \mathbb{N}^*$, $E[\mathcal{G}_n | \mathcal{A}_n] = \nabla H(w_n)$.  

 Springer
(A2) For every \( n \in \mathbb{N}^* \), there exist \( \sigma \in ]0, +\infty[ \) and \( \alpha_n \in ]0, +\infty[ \) such that

\[
E[\| G_n - \nabla H(w_n) \|^2 | A_n ] \leq \sigma^2 (1 + \alpha_n \| \nabla H(w_n) \|^2) \quad \text{almost surely.} \tag{2.3}
\]

(A3) There exists \( \epsilon \in ]0, 1[ \) such that \( (\forall n \in \mathbb{N}^*) \ 0 < \gamma_n \leq \frac{1 - \epsilon}{\beta (1 + 2\sigma^2 \alpha_n)} \).

(A4) Let \( \bar{w} \) be a solution of the problem (2.1), and set \( (\forall n \in \mathbb{N}^*) \ \chi_n^2 = \lambda_n \gamma_n^2 (1 + 2\alpha_n \| \nabla H(\bar{w}) \|^2) \). Assume that

\[
\sum_{n \in \mathbb{N}^*} \lambda_n \gamma_n = +\infty \quad \text{and} \quad \sum_{n \in \mathbb{N}^*} \chi_n^2 < +\infty. \tag{2.4}
\]

Condition (A1) means that, at each iteration \( n \), \( \mathcal{G}_n \) is an unbiased estimate of the gradient of the smooth term. While this is a classical assumption in the study of stochastic gradient methods, it has been relaxed in recent works, both to the case where the bias is vanishing [12] and to the more challenging case where it is not [1]. Unbiasedness of the estimate is of course a restriction, however the most popular stochastic algorithms in the context of machine learning are unbiased [10]. In condition (A2), \( \alpha_n \) and \( \sigma \) are not random, and they are defined so that with probability one, for any \( n \), the control of the conditional variance holds. Condition (A2) is weaker than typical conditions used in the analysis of stochastic (sub)gradient algorithms, namely boundedness of the sequence \( (E[\| \mathcal{G}_n \|^2 | A_n ])_{n \in \mathbb{N}^*} \) (see [34]) or even boundedness of \( (\| \mathcal{G}_n \|^2)_{n \in \mathbb{N}^*} \), see [18]. We note that this last requirement on the entire space is not compatible with the assumption of strong convexity, because the gradient is necessarily not uniformly bounded, therefore the use of the more general condition (A2) is needed in this case. Moreover, this allows to deal with, so called, multiplicative noise [24], see also Example 2.5. We note that when \( (\lambda_n)_{n \in \mathbb{N}^*} \) is bounded away from zero, and \( (\alpha_n)_{n \in \mathbb{N}^*} \) is bounded, (A4) implies (A3) for \( n \) large enough. The condition \( \sum_{n \in \mathbb{N}^*} \chi_n^2 < +\infty \) does not depend on the chosen solution, since \( \nabla H \) is constant on the solution set. Assumption (A4) is satisfied if \( (\lambda_n \gamma_n^2 (1 + 2\alpha_n))_{n \in \mathbb{N}^*} \) is summable. Moreover, if \( G = 0 \), it reduces to \( \sum_{n \in \mathbb{N}^*} \lambda_n \gamma_n^2 < +\infty \). The step-size is required to converge to zero, while it is typically bounded away from zero in the study of deterministic proximal forward–backward splitting algorithm [15] and in the stochastic extension considered in [13]. We remark that this is the classical summability assumption on the stepsizes in the stochastic optimization literature. However, we do not assume that \( \sum_n \lambda_n \sqrt{E[\| \mathcal{G}_n - \nabla H(w_n) \|^2]} \) is summable and therefore we do not impose summability of the errors. Our analysis is thus different from the one proposed e.g. in [13].

### 2.1 Special Cases

Problem 2.1 covers a wide class of deterministic as well as stochastic convex optimization problems, especially from machine learning and signal processing, see e.g. [15,32,34,49] and references therein. The simplest case is when \( G \) is identically equal to 0, so that Problem 2.1 reduces to the classic problem of finding a minimizer of a convex differentiable function from unbiased estimates of its gradients. In the case when
G is the indicator function of a nonempty, convex, closed set \( C \), i.e. then problem (2.1) reduces to a constrained minimization problem, which is well studied in the literature, as mentioned in the introduction. Below, we discuss some special cases of interest.

**Example 2.4 (Minimization of an Expectation).** Let \( \xi \) be a random vector with probability distribution \( P \) supported on \( E \) and \( F : \mathcal{H} \times E \rightarrow \mathbb{R} \). Stochastic gradient descent methods are usually studied in the case where \( H \) is an Euclidean space and
\[
    H(w) = \mathbb{E}[F(w, \xi)] = \int_E F(w, \xi) dP(\xi),
\]
under the assumption that \((\forall \xi \in E) F(\cdot, \xi)\) is a convex differentiable function with bounded gradient [34]. Let \((\xi_n)_{n \in \mathbb{N}^*}\) be independent copies of the random vector \( \xi \). Assume that there is an oracle that, for each \((w, \xi) \in \mathcal{H} \times E\), returns a vector \( G(w, \xi) \) such that
\[
    \nabla H(w) = \mathbb{E}[G(w, \xi)].
\]
By setting \((\forall n \in \mathbb{N}^*) \vartheta_n = G(w_n, \xi_n)\) and \( A_n = \sigma(\xi_1, \ldots, \xi_n) \), then (A1) holds. This latter assertion follows from standard properties of conditional expectation.

**Example 2.5 (Minimization of a Sum of Functions)** Let \( G \in \Gamma_0(\mathcal{H}) \), let \( m \) be a strictly positive integer. For every \( i \in \{1, \ldots, m\} \), let \( H_i \) be convex and differentiable, such that \( \sum_{i=1}^m H_i \) has a \( \beta \)-Lipschitz continuous gradient, for some \( \beta \in ]0, +\infty[ \). The problem is to
\[
    \min_{w \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m H_i(w) + G(w). \tag{2.5}
\]
This problem is a special case of Problem 2.1 with \( H = \sum_{i=1}^m H_i \), and is especially of interest when \( m \) is very large and we know the exact gradient of each component \( H_i \). The stochastic estimate of the gradient of \( H \) is then defined as
\[
    (\forall n \in \mathbb{N}) \quad \vartheta_n = \nabla H_{i(n)}(w_n), \tag{2.6}
\]
where \((i(n))_{n \in \mathbb{N}^*}\) is a random process of independent random variables uniformly distributed on \( \{1, \ldots, m\} \), see [6]. Clearly (A1) holds. Assumption (A2) specializes in this case to
\[
    (\forall n \in \mathbb{N}^*) \quad \frac{1}{m} \sum_{i=1}^m \left( \|\nabla H_i(w_n) - \frac{1}{m} \sum_{i=1}^m \nabla H_i(w_n)\|^2 \right) \leq \sigma^2 (1 + \alpha_n \|\nabla H(w_n)\|^2). \tag{2.7}
\]
If the latter is satisfied, then SPG algorithm can be applied with a suitable choice of the stepsize. In particular, (2.7) is satisfied if \( H_i \) has a \( \beta_i \)-Lipschitz continuous gradient for some \( \beta_i > 0 \), \( H \) is \( \mu \)-strongly convex for some \( \mu > 0 \) and \( \inf \alpha_n > \underline{\alpha} \) for some \( \underline{\alpha} > 0 \). Indeed, in this case, for any \( n \in \mathbb{N}^* \)
\[
    \|\nabla H_i(w_n)\|^2 \leq 2\|\nabla H_i(w_n) - \nabla H_i(0)\|^2 + 2\|\nabla H_i(0)\|^2 \leq 2\beta_i^2 \|w_n\|^2 + 2\|\nabla H_i(0)\|^2
\]
\[
\leq \frac{2\beta_i^2}{\mu^2} \| \nabla H(w_n) - \nabla H(0) \|^2 + 2\| \nabla H_i(0) \|^2 \\
\leq \frac{4\beta_i^2}{\mu^2} \| \nabla H(w_n) \|^2 + \frac{4\beta_i^2}{\mu^2} \| \nabla H(0) \|^2 + 2\| \nabla H_i(0) \|^2.
\]

Therefore, if we define

\[
\sigma^2 = \max \left\{ \alpha^{-1} \left( \frac{8}{m\mu^2} \sum_{i=1}^{m} \beta_i^2 + 2 \right), \left( \frac{8}{m\mu^2} \sum_{i=1}^{m} \beta_i^2 \right) \| \nabla H(0) \|^2 + \frac{4}{m} \sum_{i=1}^{m} \| \nabla H_i(0) \|^2 \right\},
\]

we derive from the previous inequality that

\[
\frac{1}{m} \sum_{i=1}^{m} \left( \| \nabla H_i(w_n) \| - \frac{1}{m} \sum_{i=1}^{m} \nabla H_i(w_n) \right)^2 \\
\leq \frac{2}{m} \sum_{i=1}^{m} \| \nabla H_i(w_n) \|^2 + 2\| \nabla H(w_n) \|^2 \\
\leq \left( \frac{8}{m\mu^2} \sum_{i=1}^{m} \beta_i^2 + 2 \right) \| \nabla H(w_n) \|^2 + \sigma^2 \\
\leq \sigma^2 \left( 1 + \alpha_n \| \nabla H(w_n) \|^2 \right).
\]

This shows that assumptions (A2) and (A4) are satisfied for instance in regularized least squares problems, for which classical boundedness assumptions are not satisfied, unless one assumes compactness of the domain.

We note that for solving the (deterministic) problem in (2.5), variance reduction techniques have been recently developed, allowing to derive linear convergence in the strongly convex case, at the price of a higher memory requirement [9,17,42]. Moreover, in this setting, averaging can be in general avoided using a “best selection rule” as it is standard in the literature on subgradient methods [48]. However, this strategy requires the additional computation of the function values, which could be computationally expensive in the setting where the application of an incremental method is advantageous, namely when the number of summands is large.

In the next subsection, we discuss how the above setting specializes to the context of machine learning.

### 2.2 Application to Machine Learning

Consider two measurable spaces $\mathcal{X}$ and $\mathcal{Y}$ and assume there is a probability measure $\rho$ on $\mathcal{X} \times \mathcal{Y}$. The measure $\rho$ is fixed but known only through a training set $z = (x_i, y_i)_{1 \leq i \leq m} \in (\mathcal{X} \times \mathcal{Y})^m$ of samples i.i.d with respect to $\rho$. Consider a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty]$ and a hypothesis space $\mathcal{H}$ of functions from $\mathcal{X}$ to $\mathcal{Y}$, e.g. a reproducing kernel Hilbert space. A key problem in this context is (regularized) empirical risk minimization,
minimize \( w \in H \sum_{i=1}^{m} \ell(y_i, w(x_i)) + G(w), \) \hspace{1cm} (2.8)

The above problem can be seen as an approximation of the problem,

\[ \minimize_{w \in H} \int_{\mathcal{X} \times \mathcal{Y}} \ell(y, w(x)) \, d \rho + G(w). \] \hspace{1cm} (2.9)

The analysis, in this paper, can be adapted to the machine learning setting in two different ways. The first, following Example 2.4, is to apply the SPG algorithm to directly solve the regularized expected loss minimization problem (2.9). The second, following Example 2.5, is to apply the SPG algorithm to solve the regularized empirical risk minimization problem (2.8). In either one of the above two problems, the first term is differentiable if the loss functions is differentiable with respect to its second argument, examples being the square or the logistic loss. The second term \( G \) can be seen as a regularizer/penalty encoding some prior information about the learning problem. Examples of convex, non-differentiable penalties include sparsity inducing penalties such as the \( \ell_1 \) norm, as well as more complex structured sparsity penalties [32]. Stronger convexity properties can be obtained considering an elastic net penalty [16,52], that is adding a small strongly convex term to the sparsity inducing penalty. Clearly, the latter term would not be necessary if the risk in Problem 2.9 (or the empirical risk in (2.8)) is strongly convex. However, this latter requirement depends on the probability measure \( \rho \) and is typically not satisfied when considering high (possibly infinite) dimensional settings.

3 Main Results and Discussion

In this section, we state and discuss the main results of the paper. We derive convergence rates of the proximal gradient algorithm (with relaxation) for stochastic minimization. The section is divided in two parts. In the first one, Sect. 3.2, we focus on convergence in expectation. In the second one, Sect. 3.3, we study almost sure convergence of the sequence of iterates. In both cases, additional convexity conditions on the objective function are required to derive convergence results. The proofs are deferred to Sect. 1.

3.1 Convergence of Function Values

In this section we consider convergence results for the function values in the convex case. We first prove a new result, establishing a convergence rate (in expectation) of the objective value corresponding to the last iterate. The convergent behavior of the objective function on the last iterate in practice was already observed in [45], and proved for the stochastic subgradient method in [38,47]. The proof is based on the adaptation to the stochastic case of a recent result proved in [30] to characterize convergence of a wide class of optimization algorithms.
Theorem 3.1 Suppose that the conditions \((A_1), (A_2), (A_3), \text{ and } (A_4)\) are satisfied and let \((w_n)_{n \in \mathbb{N}^*}\) be a sequence generated by Algorithm 2.2 with \(\sup \alpha_n < +\infty\), \((\gamma_n)_{n \in \mathbb{N}^*}\) decreasing, and \(\lambda_n = 1\) for every \(n \in \mathbb{N}^*\). Then, for every minimizer \(\bar{w}\) of \(\Phi\),

\[
E[\Phi(w_{n+1}) - \min \Phi] \leq \frac{\|w_1 - \bar{w}\|^2 + \sigma^2 (1 + \alpha_1 E[\|\nabla L(w_1)\|^2])}{n\gamma_n} + \frac{1}{\gamma_n} \sum_{i=1}^{n-1} \frac{\sigma^2 \gamma_i^2 (1 + \alpha_i E[\|\nabla L(w_i)\|^2])}{n-i+1}.
\]

In particular, suppose that for every \(n \in \mathbb{N}^*\), \(\gamma_n = cn^{-\theta}\) for some \(\theta \in ]1/2, 1[\) and \(c \in ]0, +\infty[\). Then, there exists \(C(\theta) \in ]0, +\infty[\) such that

\[
E[\Phi(w_{n+1}) - \min \Phi] \leq \frac{C(\theta)}{n^{1-\theta}}.
\] (3.1)

Remark 3.2 Choosing \(\theta\) arbitrarily close to \(1/2\) we can derive a convergence rate arbitrarily close to \(O(1/\sqrt{n})\). The upper bound on the convergence rate can also be computed for other choices of the stepsize, specializing (3.1) accordingly.

Remark 3.3 The proof of the second part of Theorem 3.1 is derived by adapting the proof of [30, Theorem 1] to the stochastic case.

For completeness, we establish in this section also ergodic convergence, i.e. convergence of the expectation of the function values corresponding to the averaged sequence. This is essentially known, but usually proved under different assumptions, such as Lipschitz continuity of the objective function (see e.g. [18]), which is not assumed here.

Theorem 3.4 Suppose that the conditions \((A_1), (A_2), (A_3), \text{ and } (A_4)\) are satisfied and let \((w_n)_{n \in \mathbb{N}^*}\) be a sequence generated by Algorithm 2.2 with \(\lambda_n = 1\) for every \(n \in \mathbb{N}^*\). Define \(\bar{w}_t = (\sum_{n=1}^t \gamma_n w_{n+1}) / (\sum_{n=1}^t \gamma_n)\). Then there exists \(C \in ]0, +\infty[\) such that

\[
E[\Phi(\bar{w}_n) - \min \Phi] \leq \frac{C}{\sum_{i=1}^n \gamma_i}.
\] (3.2)

In particular, \(\lim_{n \to +\infty} E[\Phi(\bar{w}_n) - \min \Phi] = 0\). Moreover suppose that, for every \(i \in \mathbb{N}^*\), \(\gamma_i = ci^{-\theta}\) for some \(\theta \in ]1/2, 1[\) and \(c \in ]0, +\infty[\). Then, there exists \(\bar{C} \in ]0, +\infty[\)

\[
E[\Phi(\bar{w}_n) - \min \Phi] \leq \frac{\bar{C}}{n^{1-\theta}}.
\] (3.3)

Comparing Eqs. (3.1) and (3.3), we see that the convergence rate of the objective values corresponding to the last iterate is arbitrarily close to the one on the averaged sequence if the stepsize is arbitrarily close to \(1/2\).
3.2 Convergence in Expectation of SPG Algorithm

In this section, we denote by \( \overline{w} \) a solution of Problem 2.1 and provide an explicit non-asymptotic bound on \( E[\|w_n - \overline{w}\|^2] \). Our convergence results consider convergence of the iterates with no averaging and hold in an infinite dimensional setting, without boundedness assumptions.

**Theorem 3.5** Under conditions (A1), (A2), (A3), assume that \( H \) is \( \mu \)-strongly convex and \( G \) is \( v \)-strongly convex, for some \( \mu \in [0, +\infty[ \) and \( v \in [0, +\infty[ \), with \( \mu + v > 0 \). Suppose that there exist \( \bar{\lambda} \in ]0, 1[ \) and \( \bar{\alpha} \in [0, +\infty[ \) such that

\[
\inf_{n \in \mathbb{N}^*} \lambda_n \geq \lambda \quad \text{and} \quad \sup_{n \in \mathbb{N}^*} \alpha_n \leq \bar{\alpha}.
\]

Let \( c_1 \in ]0, +\infty[ \) and let \( \theta \in ]0, 1[ \). Suppose that, for every \( n \in \mathbb{N}^* \), \( \gamma_n = c_1 n^{-\theta} \). Set

\[
t = 1 - 2^\theta - 1, \quad c = \frac{2 c_1 \lambda (v + \mu \epsilon)}{(1 + v)^2}, \quad \text{and} \quad \tau = \frac{2 \sigma^2 c_1^2 (1 + \bar{\alpha} \|\nabla H(\overline{w})\|^2)}{c^2}.
\]

Let \( n_0 \) be the smallest integer such that \( n_0 > 1 \), and \( \max\{c, c_1\} n_0^{-\theta} \leq 1 \). Let \((w_n)_{n \in \mathbb{N}^*}\) be the sequence generated by the SPG algorithm. Then, for every \( n \geq 2n_0 \), we have

\[
E[\|w_n - \overline{w}\|^2] = \begin{cases} O(n^{-\theta}) & \text{if } \theta \in ]0, 1[, \\ O(n^{-c}) + O(n^{-1}) & \text{if } \theta = 1. \end{cases}
\]

Thus, if \( \theta = 1 \) and \( c_1 \) is chosen such that \( c > 1 \), then \( E[\|w_n - \overline{w}\|^2] = O(n^{-1}) \). In particular, if \( \theta = 1 \), \( \lambda_n = \lambda = 1 \) for every \( n \in \mathbb{N}^* \), and \( c_1 = (1 + v)^2/(v + \mu \epsilon) \), then

\[
E[\|w_n - \overline{w}\|^2] \leq \frac{n_0^2 E[\|w_{n_0} - \overline{w}\|^2]}{(n + 1)^2} + \frac{8 \sigma^2 (1 + \bar{\alpha} \|\nabla H(\overline{w})\|^2)(1 + v)^4}{(\mu \epsilon + v)^2 (n + 1)^2} \quad (3.7)
\]

The proof of Theorem 3.5 immediately follows from Theorem A.3. Theorem A.3 is the extension to the nonsmooth case of [3, Theorem 1], in particular, when \( G = 0 \), we obtain the same bounds. Note however, that the assumptions on the stochastic approximations of the gradient of the smooth part are different. In particular, we replace the boundedness condition at the solution and the Lipschitz continuity assumption on \((\mathcal{G}_n)_{n \in \mathbb{N}^*}\) with assumption (A2).

As can be seen from Theorem 3.5, the fastest asymptotic rate corresponds to \( \theta = 1 \) and it is the same obtained in the smooth case in [3, Theorem 2]. Note that this rate depends on the asymptotic behavior of the step-size, but also on the constant \( c \), which in turns depends on \( c_1 \). As pointed out in [34], see also in [3], this choice is critical, because too small choices of \( c_1 \) affect the convergence rates, and too big choices influence significantly the value of the constants in the first term of (A.23), indeed the choice is determined by the strong convexity constants. Moreover, the dependence on
the strong convexity constant shown in Theorem 3.5 is of the same type of the one obtained in the regret minimization framework by [23].

There are other stochastic first order methods achieving the same rate of convergence for the iterates in the strongly convex case, see e.g. [1, 21, 23, 25, 31, 51]. Indeed, the rate we obtain is the rate that can be obtained by the optimal (in the sense of [35]) convergence rate on the function values. Among the mentioned methods those in [1, 21, 31] belong to the class of accelerated proximal gradient methods. In addition, if sparsity is the main interest, we highlight that many of the algorithms discussed above (e.g. [1, 21, 23, 51]) involve some form of averaging or linear combination which prevents sparsity of the iterates, as it is discussed in [31]. Our result shows that in this case averaging is not needed, since the iterates themselves are convergent. Note that, even when $\lambda_n < 1$, the relaxation step $w_{n+1} = (1 - \lambda_n)w_n + \lambda_n y_n$ does not critically affect sparsity, since it involves the sum of two sparse vectors.

### 3.3 Almost Sure Convergence of SPG Algorithm

In this section, we focus on almost sure convergence of SPG algorithm. This kind of convergence of the iterates is the one traditionally studied in the stochastic optimization literature. Depending on the convexity properties of the function $H$, we get two different convergence properties. The first theorem requires uniform convexity of $H$ at the solution.

**Theorem 3.6** Suppose that the conditions (A1), (A2), (A3), and (A4) are satisfied. Let $(w_n)_{n\in\mathbb{N}^*}$ be a sequence generated by Algorithm 2.2 and assume that $H$ is uniformly convex at $\bar{w}$. Then $w_n \to \bar{w}$ a.s.

If we relax the strong convexity assumption, we can still prove weak convergence of a subsequence in the strictly convex case, provided an additional regularity assumption holds.

**Theorem 3.7** Suppose that the conditions (A1), (A2), (A3), and (A4) are satisfied. Let $(w_n)_{n\in\mathbb{N}^*}$ be a sequence generated by Algorithm 2.2. Assume that $H$ is strictly convex, and let $\bar{w}$ be the unique solution of Problem 2.1. If $\nabla H$ is weakly continuous, then there exists a subsequence $(w_{n_k})_{n\in\mathbb{N}^*}$ such that $w_{n_k} \to \bar{w}$ a.s.

With respect to the previous section, here we make the additional assumption (A4) on the summability of the sequence of step-sizes multiplied by the relaxation parameters. For stochastic gradient algorithm without relaxation, i.e, $G = 0$ and, for every $n \in \mathbb{N}^*$ $\lambda_n = 1$, assumption (A4) coincides with the classical step-size condition $\sum_{n\in\mathbb{N}^*} \gamma_n = +\infty$ and $\sum_{n\in\mathbb{N}^*} \gamma_n^2 < +\infty$ which guarantees a sufficient but not too fast decrease of the step-size (see e.g. [7]). As mentioned in the introduction, the study of almost sure convergence is classical. Our approach is based on random quasi-Fejér sequences, and on probabilistic quasi martingale techniques.

**Remark 3.8** (Comparison to previous work) In the convex setting, almost sure convergence of the iterates has been proved using a similar approach in [12], but under a summability assumptions on the variance of the stochastic estimates. The same result
has been proved in [1], assuming boundedness of the domain and requiring a sort of Lipschitz continuity of $G$, but allowing a biased estimate of the gradient.

**Remark 3.9** If $H$ is assumed to be only strictly convex and its gradient is not weakly continuous, Theorem 3.7 does not ensure weak convergence of any subsequence of $(w_n)_{n \in \mathbb{N}^*}$. However, if the sequence of function values $(\Phi(w_n))_{n \in \mathbb{N}^*}$ converges to the minimum of $\Phi$ a.s., then $w_n \rightharpoonup \overline{w}$ a.s.

### 4 Numerical Experiments

In this section we first present numerical experiments aimed at studying the computational performance of the SPG algorithm (see Algorithm 2.2), with respect to the step-size, the strong convexity constant, and the noise level. Then we compare the proposed method with other state-of-the-art stochastic first order methods: an accelerated stochastic proximal gradient method, called SAGE [28, Theorem 2] and the FOBOS algorithm [18].

#### 4.1 Properties of SPG

In order to study the behavior of the SPG algorithm with respect to the relevant parameters of the optimization problem, we focus on a toy example, where the exact solution is known. More specifically, we consider the following minimization problem on the real line:

$$
\text{minimize } \phi(w) := \frac{\mu}{2} |w - 10|^2 + 0.02|w - 10|.
$$

(4.1)

It is clear that $\phi$ is $\mu$-strongly convex function with $w_{\text{opt}} = \arg\min \phi = \{10\}$ and the optimal value $\Phi = 0$. We consider a stochastic perturbation of the exact gradient of the function $H = \frac{1}{2} |\cdot - 10|^2$ of the form

$$
\mathcal{G}_n = \nabla H(w_n) + s_n,
$$

(4.2)

where $s_n$ is a realization of a Gaussian random variable with 0 mean and $\sigma^2$ variance. We apply SPG one hundred times for 100 independent realizations of the random process $(s_n)_{n \in \mathbb{N}^*}$ to problem (4.1) with $(\forall n \in \mathbb{N}^*) \lambda_n = 1$ and $\gamma_n = C/n$ for some constant $C > 0$. We evaluate the average performance of SPG over the first 100 iterations for different values of the strong convexity parameter $\mu$, and several values of $\sigma$ and $C$, and by measuring $|w_n - 10|$. The results are displayed in Fig. 1. As can be seen by visual inspection, the convergence is faster when $\mu$ is bigger and when the noise variance is smaller. Moreover, the constant $C$ in the step-size heavily influences the convergence behavior. The latter is a well-known phenomenon in the context of stochastic optimization [34].

#### 4.2 Comparison with Other Methods

In this section we compare SPG with the SAGE algorithm [28, Algorithm 1] and the FOBOS algorithm in [18]. We note that the main difference between SPG (with...
Fig. 1 Performance evaluation of Algorithm 2.2 with respect to different choices of $\mu$, setting $\gamma_n = 0.8/n$ and $\sigma = 0.01$ (left), with respect to different choices of $C$, with $\gamma_n = C/n$, $\mu = 0.05$ and $\sigma = 0.01$ (center), and with respect to different choices of $\sigma$, for $\mu = 0.05$ and $\gamma_n = 1/n$ (right).

$\lambda_n = 1$ for every $n \in \mathbb{N}^*$) and FOBOS is that the latter takes the average of the previous iterates. More precisely the sequence generated by the FOBOS iteration is the following

$$
\bar{w}_n = \left( \sum_{k=1}^{n} \eta_k \right)^{-1} \sum_{k=1}^{n} \eta_k w_k, \quad \eta_k = C_1/k,
$$

(4.3)

where $(w_k)_{k \in \mathbb{N}^*}$ is the sequence generated by the SPG algorithm. In [18] it is assumed that the gradient of the smooth term is bounded on the whole space. In our experiments this assumption is not satisfied, but since the sequence of iterates is bounded, the algorithm can be applied and its convergence is guaranteed. One advantage of the SAGE algorithm is that it does not require any parameter tuning. SPG and FOBOS instead require the choice of the stepsize. We check the accuracy of the three algorithms on different elastic net regularized problems with respect to the number of iterations, since the cost per iteration is the same for the three procedures.

4.2.1 Toy Example

We first consider the toy example presented in the previous section (see Eq. (4.1)), where we set $\mu = 1$. Moreover, we assume that in (4.2) $s_n$ is a realization of a Gaussian random variable with 0 mean and 0.1 variance. We run SPG, SAGE, and FOBOS one hundred times for one hundred independent realizations of the random process $(s_n)_{n \in \mathbb{N}^*}$. In SPG, we chose $\gamma_n = 1/n$, and $\lambda_n = 1$. Finally, after testing the FOBOS algorithm for different choices of the constant $C_1$ defining the stepsize, we got that $\eta_k = 1/k$ for every $k \in \mathbb{N}^*$ gave the best results. The average behavior of the sequences $|w_n - 10|$ corresponding to 100 independent realizations of the random process for the three algorithms on the first 1000 iterations is presented in Fig. 2. SPG and SAGE have a similar behavior, while FOBOS is slower.

4.2.2 Regression Problems with Random Design

Let $N$ and $p$ be strictly positive integers. Concerning the data generation protocol, the input points $(x_i)_{1 \leq i \leq N}$ are uniformly drawn in the interval $[a, b]$ (to be specified later
in the two cases we consider). For a suitably chosen finite dictionary of real valued functions $(\phi_k)_{1 \leq k \leq p}$ defined on $[a, b]$, the labels are computed using a noise-corrupted regression function, namely

$$(\forall i \in \{1, \ldots, N\}) \quad y_i = \sum_{k=1}^{p} \overline{w}_k \phi_k(x_i) + \epsilon_i,$$  
(4.4)

where $(\overline{w}_k)_{1 \leq k \leq p} \in \mathbb{R}^p$ and $\epsilon_i$ is an additive noise $\epsilon_i \sim \mathcal{N}(0, 0.3)$.

We will consider two different choices for the dictionary of functions: polynomials, i.e. $(\forall k \in \{1, \ldots, p\}) \phi_k : [-1, 1] \to \mathbb{R}, \phi_k(x) = x^{k-1}$ and trigonometric functions, i.e. $p = 2q + 1$ and $(\forall k \in \{1, \ldots, q\}) \phi_k : [0, 2\pi] \to \mathbb{R}, \phi_k(x) = \cos((k-1)x)$ and $(\forall k \in \{q+1, \ldots, 2q+1\}) \phi_k : [0, 2\pi] \to \mathbb{R}, \phi_k(x) = \sin(kx).$ The training set and the regression function for the two examples are presented in Fig. 3. We estimate $\overline{w}$ by solving the following regularized minimization problem

$$\minimize_{(w_k)_{1 \leq k \leq p} \in \mathbb{R}^p} \frac{1}{2N} \sum_{i=1}^{N} \left( y_i - \sum_{k=1}^{p} w_k \phi_k(x_i) \right)^2 + \frac{1}{2} \sum_{k=1}^{p} (\mu |w_k|^2 + \omega |w_k|),$$  
(4.5)
where $\mu$ and $\omega$ are strictly positive parameters. Problem (4.5) is a special case of Example 2.5, and hence it can be solved by using SPG, SAGE, and FOBOS in an incremental fashion. For the polynomial dictionary, we set

$$p = 6, \ N = 9, \ \gamma_n = 15/(n + 100), \ \eta_n = 15/(n + 100), \ \mu = 0.1, \ \omega = 0.01, \ w = [3, 2, 1, 0, 1, 0].$$

(4.6)

For the trigonometric dictionary, we set

$$p = 21, \ N = 32, \ \gamma_n = \eta_n = 10/(n + 100), \ \mu = 0.01, \ \omega = 0.01, \ \bar{w} = [0, 0.2, 0, 0.5, 1, -1, 0, 1, 2, 0.5, 0, 0, -0.1, -2.5, 1, 0, 0, -1, 0.9, -0.5, 0].$$

(4.7)

The resulting regression functions using the three algorithms are shown in Fig. 4. As can be seen from visual inspection, the three methods provide almost undistinguishable solutions. Finally, we computed an approximate solution of (4.5) by running the forward–backward splitting method in [15] for 50000 iterations. The convergence of the iterations to the solution of (4.5) is displayed in Fig. 5. On the regression problem with the polynomial dictionary, SAGE is performing the best, while on the trigonometric dictionary, SPG is the fastest. The oscillating behavior is mitigated by the averaging procedure at the expenses of a slower convergence rate, as the more regular behaviour of FOBOS clearly shows.
4.2.3 Deconvolution Problems

As a last experiment, we focus on the problem of recovering an ideal signal $\overline{w}$ from a noisy observation of the form

$$\mathbb{R}^{1024} \ni y = h \ast \overline{w} + s,$$

where $s \sim \mathcal{N}(0, 0.06)$ and $h$ is a Gaussian kernel. To find an approximation of the ideal signal, we solve the following variational problem

$$\min_{w \in \mathbb{R}^{1024}} \Phi(w), \quad \Phi(w) = \frac{1}{2} \| y - h \ast w \|^2 + \| w \|_1 + \frac{0.02}{2} \| w \|_2^2.$$

An approximation $\overline{w}$ of the exact solution is found by running the forward–backward splitting method in [15] for 10,000 iterations. Then, we run SPG, SAGE, and FOBOS with the same initialization, for 5000 iterations using at the $n$-th iteration a stochastic gradient of the form

$$\mathcal{G}_n = \nabla H(w_n) + s_n,$$
Fig. 8 Number of zero components of the vector \((w_n - \bar{w})_{n \in \mathbb{N}^*}\) with the same initial point 0 for SPG and SAGE.

where \(s_n \sim \mathcal{N}(0, 0.01)\) and \(H(w) = \frac{1}{2}\|y - h \ast w\|^2 + \frac{0.02}{2}\|w\|^2_2\). FOBOS is run with \(\eta_n = 3/(n + 100)\). This is not the theoretically optimal choice, but gave better results in practice. In SPG we set \(\lambda_n = 1\) and \(\gamma_n = 3/(n + 100)\). Convergence of \((\|w_n - \bar{w}\|)_{n \in \mathbb{N}^*}\) for the three algorithms is presented in Fig. 7. In this case SAGE is the fastest, and SPG shows slightly worse convergence. FOBOS is again slower. Finally, we address the problem of the iterations’ sparsity. We generate the data according to the model in (4.8), starting from an original signal with 993 zero components. In Fig. 8 we display the number of zero components of the iterates. As it can be readily seen by visual inspection, after few iteratons both SAGE ans SPG generate sparse iterations. On this example this does not hold for the FOBOS algorithm, for which the sparsity of the iterates is a decreasing function of the number of iterations. The number of zero components of the last iterate of SPG, SAGE, and FOBOS is 937, 937, and 438, respectively.

Acknowledgements This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216 and H2020-MSCA-RISE-2017 Project 777826. L. R. acknowledges the financial support of the Italian Ministry of Education, University and Research FIRB project RBFR12M3AC, and S. V. the ones of the Gruppo Nazionale per l’Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM). The work of Bang Cong Vũ is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under Grant Number 102.01-2017.05

Appendix: Proofs

We start by stating a simple lemma, whose proof is straightforward.

Lemma A.1 Let \((a_i)_{i \in \mathbb{N}}\) be a sequence in \(\mathbb{R}\). Let \(n \in \mathbb{N}, n > 1, k \in \{1, \ldots, n - 1\}\), and let \(s_k = \sum_{i=n-k}^n a_i\). Then

\[
a_n = \frac{s_{n-1}}{n} + \sum_{k=1}^{n-1} \frac{1}{k(k+1)} (s_{k-1} - ka_{n-k}).
\]
Proof of Theorem 3.1} Since \( \gamma_n \leq 1/\beta \) by condition (A3), from \([50, \text{Lemma 4.1}]\), for every \((u, v) \in \mathcal{H}^2\), for every \(w \in \text{dom } G\), and \(\eta \in \partial G(w)\) we have

\[
\Phi(u) \geq \Phi(w) + \langle u - w \mid \nabla H(v) + \eta \rangle - \frac{1}{2\gamma_n} \|w - v\|^2. \tag{A.1}
\]

Let \(u \in \mathcal{H}\). Applying the previous inequality with \(w = w_{n+1}\), \(\eta = \gamma_n^{-1}(z_n - w_{n+1})\), and \(v = w_n\) we obtain

\[
\Phi(u) \geq \Phi(w_{n+1}) + \left(\langle u - w_{n+1} \mid \nabla H(w_n) - \mathcal{G}_n + \gamma_n^{-1}(w_n - w_{n+1}) \rangle - \frac{1}{2\gamma_n} \|w_{n+1} - w_n\|^2 \right) \tag{A.2}
\]

and thus, setting \((\forall n \in \mathbb{N}^+) \tilde{w}_{n+1} = \text{prox}_{\gamma_n G}(w_n - \gamma_n \nabla H(w_n))\),

\[
\Phi(w_{n+1}) - \Phi(u) \leq \langle u - w_{n+1} \mid \mathcal{G}_n - \nabla H(w_n) - \gamma_n^{-1}(w_n - w_{n+1}) \rangle + \frac{1}{2\gamma_n} \|w_{n+1} - w_n\|^2 \\
= \langle u - w_{n+1} \mid \mathcal{G}_n - \nabla H(w_n) \rangle - \frac{1}{2\gamma_n} \left(2 \langle u - w_{n+1} \mid (w_n - w_{n+1}) \rangle + \|w_{n+1} - w_n\|^2\right) \\
= \langle u - \tilde{w}_{n+1} \mid \mathcal{G}_n - \nabla H(w_n) \rangle + \langle \tilde{w}_{n+1} - w_{n+1} \mid \mathcal{G}_n - \nabla H(w_n) \rangle + \frac{1}{2\gamma_n} \|w_n - u\|^2 - \|w_{n+1} - u\|^2 \tag{A.3}
\]

Since \(\text{prox}_{\gamma_n G} : \mathcal{H} \rightarrow \mathcal{H}\) is continuous, it is measurable with respect to the Borel \(\sigma\)-algebra. Therefore, taking into account that \(w_n\) and \(\nabla H(w_n)\) are \(\mathcal{F}_n\) measurable, we get by assumption (A1)

\[
\mathbb{E}[\langle u - \tilde{w}_{n+1} \mid \mathcal{G}_n - \nabla H(w_n) \rangle \mid \mathcal{F}_n] = \langle u - \tilde{w}_{n+1} \mid \mathbb{E}[\mathcal{G}_n - \nabla H(w_n) \mid \mathcal{F}_n] \rangle = 0. \tag{A.4}
\]

Moreover, by the nonexpansiveness of the proximity operator we derive

\[
\langle \tilde{w}_{n+1} - w_{n+1} \mid \mathcal{G}_n - \nabla H(w_n) \rangle \leq \gamma_n \|\mathcal{G}_n - \nabla H(w_n)\|^2. \tag{A.5}
\]

Multiplying by \(2\gamma_n\) and taking the expectation in (A.3), from (A.4) and (A.5) we obtain

\[
2\gamma_n \mathbb{E}[\Phi(w_{n+1}) - \Phi(u)] \leq 2\gamma_n^2 \mathbb{E}[\|\mathcal{G}_n - \nabla H(w_n)\|^2] + \mathbb{E}[\|w_n - u\|^2] - \mathbb{E}[\|w_{n+1} - u\|^2]. \tag{A.6}
\]

and thus, by (A2)

\[
\gamma_n \mathbb{E}[\Phi(w_{n+1}) - \Phi(u)] \leq \gamma_n^2 \alpha^2 (1 + \alpha_n \mathbb{E}[\|\nabla H(w_n)\|^2]) + \frac{1}{2} \mathbb{E}[\|w_n - u\|^2] - \frac{1}{2} \mathbb{E}[\|w_{n+1} - u\|^2]. \tag{A.7}
\]
Let $\overline{w} \in \mathcal{H}$ be a minimizer of $\Phi$. Applying Lemma A.1 to the sequence ($\forall n \in \mathbb{N}^*$) $a_n = \gamma_n E[\Phi(w_{n+1}) - \Phi(\overline{w})]$ we derive:

$$
\gamma_n E[\Phi(w_{n+1}) - \Phi(\overline{w})] = \frac{1}{n} \sum_{i=1}^{n} \gamma_i E[\Phi(w_{i+1}) - \Phi(\overline{w})] \\
+ \sum_{k=1}^{n-1} \frac{1}{k(k+1)} \sum_{i=n-k+1}^{n} \gamma_i E[\Phi(w_{i+1}) - \Phi(w_{n-k+1})] \\
+ \sum_{k=1}^{n-1} \frac{1}{k+1} \left[ \left( \frac{1}{k} \sum_{i=n-k+1}^{n} \gamma_i \right) - \gamma_{n-k} \right] E[\Phi(w_{n-k+1}) - \Phi(\overline{w})].
$$

(A.7)

Since $\{\gamma_n\}_{n \in \mathbb{N}^*}$ is decreasing and $\Phi(w_{n-k+1}) - \Phi(\overline{w}) \geq 0$, we get

$$
\gamma_n E[\Phi(w_{n+1}) - \Phi(\overline{w})] \leq \frac{1}{n} \sum_{i=1}^{n} \gamma_i E[\Phi(w_{i+1}) - \Phi(\overline{w})] \\
+ \sum_{i=1}^{n-1} \frac{1}{k(k+1)} \sum_{i=n-k+1}^{n} \gamma_i E[\Phi(w_{i+1}) - \Phi(w_{n-k+1})].
$$

(A.8)

Define, for every $n \in \mathbb{N}^*$, $\xi_n = \gamma_n^2 \sigma^2 (1 + \alpha_n E[\nabla H(w_n) ||^2])$. Let $j \in \{1, \ldots, n\}$. Taking the sum in (A.7) from $j$ to $n$, we obtain:

$$
\sum_{i=j}^{n} \gamma_i E[\Phi(w_{i+1}) - \Phi(u)] \leq \sum_{i=j}^{n} \xi_i + \frac{1}{2} E[||w_j - u||^2].
$$

The above inequality with $u = \overline{w}$ and $j = 1$ implies

$$
\frac{1}{n} \sum_{i=1}^{n} \gamma_i E[\Phi(w_{i+1}) - \Phi(\overline{w})] \leq \frac{1}{2n} ||w_1 - \overline{w}||^2 + \frac{1}{n} \sum_{i=1}^{n} \xi_i.
$$

(A.9)

Inequality (A.7) with $u = w_{n-k+1}$, which satisfies (A.4), and $j = n - k + 1$ yields

$$
\sum_{i=n-k+1}^{n} \gamma_i E[\Phi(w_{i+1}) - \Phi(w_{n-k+1})] \leq \sum_{i=n-k+1}^{n} \xi_i.
$$

(A.10)

Therefore,

$$
\sum_{k=1}^{n-1} \frac{1}{k(k+1)} \sum_{i=n-k+1}^{n} \gamma_i E[\Phi(w_{i+1}) - \Phi(w_{n-k+1})] \leq \sum_{k=1}^{n-1} \frac{1}{k(k+1)} \sum_{i=n-k+1}^{n} \xi_i.
$$

(A.11)
Exchanging the order in the sum, we obtain
\[
\sum_{k=1}^{n-1} \frac{1}{k(k+1)} \sum_{i=n-k+1}^{n} \xi_i = \sum_{i=2}^{n} \sum_{k=n-i+1}^{n-1} \frac{1}{k(k+1)} \xi_i = \sum_{i=2}^{n} \frac{1}{n-i+1} \xi_i - \frac{1}{n} \sum_{i=2}^{n} \xi_i.
\] (A.12)

Equation (3.1) follows by plugging (A.9), (A.11), and (A.12) into (A.8). Next, set, for every \(n \in \mathbb{N}^+\), \(\gamma_n = cn^{-\theta}\) for some \(\theta \in ]1/2, 1[\) and \(c \in ]0, +\infty[\). Since \(\sup \alpha_n < +\infty\), and \(\sup \mathbb{E}[\|L(w_n)\|^2] < +\infty\) by Proposition A.4(i), there exists \(c_1 \in ]0, +\infty[\) such that \(\sum_{i=1}^{n} \xi_i / (n - i + 1) \leq c_1 \sum_{i=1}^{n} i^{-2\theta} / (n - i + 1)\). Finally, (3.1) is derived by the upper bound \(\sum_{i=1}^{n} i^{-2\theta} / (n - i + 1) \leq (8/(2\theta - 1))n^{-1}\).

**Proof of Theorem 3.4** Taking the sum in (A.6) from 1 to \(n\) we get
\[
2 \sum_{i=1}^{n} \gamma_i \mathbb{E}(\Phi(w_{i+1}) - \Phi(u)) \leq 2 \sum_{i=1}^{n} \gamma_i^2 \mathbb{E}(\|\Phi_i - \nabla H(w_i)\|^2) + \|w_1 - u\|^2 - \mathbb{E}(\|w_{i+1} - u\|^2).
\] (A.13)

By (A2) we have \(\mathbb{E}(\|\Phi_i - \nabla H(w_i)\|^2) \leq \sigma^2(1 + \alpha_i \mathbb{E}(\|\nabla H(w_i)\|^2))\) and \(\mathbb{E}(\nabla H(w_i))\) is bounded by Proposition A.4(i). The sequence \((\gamma_i^2 \mathbb{E}(\|\Phi_i - \nabla H(w_i)\|^2))_{i \in \mathbb{N}^+}\) is summable by assumption (A4). This, together with (A.13) implies that \((\gamma_i \mathbb{E}(\Phi(w_{i+1}) - \Phi(u)))_{i \in \mathbb{N}^+}\) is summable. Dividing (A.13) by \(\sum_{i=1}^{n} \gamma_i\) and using convexity of \(\Phi\), we derive that there exists a constant \(C \in ]0, +\infty[\) such that
\[
\mathbb{E}(\Phi(\bar{w}_n) - \min \Phi) \leq \frac{C}{\sum_{i=1}^{n} \gamma_i}.
\] (A.14)

The last part of the statement easily follows from (A.14).

We next state a technical result, giving some bounds that will be repeatedly used.

**Proposition A.2** Consider the setting of the SPG algorithm and let \(\bar{w}\) be a solution of Problem 2.1. Suppose that conditions (A1), (A2), and (A3) are satisfied. Then the following hold:

(i) \((\forall n \in \mathbb{N}^+)\) \(\|w_{n+1} - \bar{w}\|^2 \leq (1 - \lambda_n)\|w_n - \bar{w}\|^2 + \lambda_n\|y_n - \bar{w}\|^2\).

(ii) Set
\[
(\forall n \in \mathbb{N}^+)\ u_n = w_n - y_n - \gamma_n(\Phi_n - \nabla H(\bar{w})).
\] (A.15)

Then, for every \(n \in \mathbb{N}^+\)
\[
\|y_n - \bar{w}\|^2 \leq \|w_n - \bar{w}\|^2 - 2\gamma_n \langle w_n - \bar{w} | \Phi_n - \nabla H(\bar{w}) \rangle + \gamma_n^2 \|\Phi_n - \nabla H(\bar{w})\|^2 - \|u_n\|^2.
\] (A.16)
(iii) For every \( n \in \mathbb{N}^* \)

\[
E[\|y_n - \overline{w}\|^2 | A_n] \leq \|w_n - \overline{w}\|^2 - 2\gamma_n (1 - \gamma_n \beta (1 + 2\sigma^2 \alpha_n)) \cdot \\
\langle w_n - \overline{w} | \nabla H(w_n) - \nabla H(\overline{w}) \rangle + 2\gamma_n^2 \sigma^2 (1 + \alpha_n \|\nabla H(\overline{w})\|^2).
\]

(A.17)

\textbf{Proof} (i): Follows from convexity of \( \| \cdot \|^2 \).

(ii): We have

\[
(\forall n \in \mathbb{N}^*) \quad \overline{w} = \text{prox}_{\gamma_n G}(w - \gamma_n \nabla H(\overline{w})).
\]

(A.18)

Moreover, since \( \text{prox}_{\gamma_n G} \) is firmly non-expansive by [15, Lemma 2.4]

\[
\|y_n - \overline{w}\|^2 \leq \|(w_n - \overline{w}) - \gamma_n (G_n - \nabla H(\overline{w}))\|^2 - \|u_n\|^2
\]

and the statement follows.

(iii): It follows from (ii) that, for every \( n \in \mathbb{N}^* \),

\[
E[\|y_n - \overline{w}\|^2 | A_n] \leq \|(w_n - \overline{w}) - \gamma_n (G_n - \nabla H(\overline{w}))\|^2 - \|u_n\|^2
\]

Using assumption (A1) and the fact that \( w_n \) is \( A_n \) measurable, we derive

\[
(\forall n \in \mathbb{N}^*) \quad E[\langle w_n - \overline{w} | G_n - \nabla H(\overline{w}) \rangle | A_n] = \langle w_n - \overline{w} | E[G_n - \nabla H(\overline{w}) | A_n] \rangle
\]

\[
= \langle w_n - \overline{w} | \nabla H(w_n) - \nabla H(\overline{w}) \rangle.
\]

(A.19)

Moreover, using the assumption (A2), we have

\[
E[\|G_n - \nabla H(\overline{w})\|^2 | A_n] \leq 2\|\nabla H(w_n) - \nabla H(\overline{w})\|^2 + 2E[\|G_n - \nabla H(w_n)\|^2 | A_n]
\]

\[
\leq 2\|\nabla H(w_n) - \nabla H(\overline{w})\|^2 + 2\sigma^2 (1 + \alpha_n \|\nabla H(w_n)\|^2)
\]

\[
\leq (2 + 4\sigma^2 \alpha_n) \beta \langle w_n - \overline{w} | \nabla H(w_n) - \nabla H(\overline{w}) \rangle
\]

\[
+ 2\sigma^2 (1 + 2\alpha_n \|\nabla H(\overline{w})\|^2), \quad \text{(A.20)}
\]

where the last inequality follows from the fact that \( \nabla H \) is cocoercive since it is Lipschitz-continuous (by the Baillon-Haddad Theorem, see e.g. [4, Theorem 18.15]). The statement then follows from (A.16), (A.19), and (A.20). \( \square \)

In the statement of the following theorem, we will use the family of functions \((\varphi_c)_{c \in \mathbb{R}}\) defined by setting, for every \( c \in \mathbb{R} \),

\[
\varphi_c : ]0, +\infty[ \to \mathbb{R} : t \mapsto \begin{cases} (t^c - 1)/c & \text{if } c \neq 0; \\
\log t & \text{if } c = 0.
\end{cases}
\]

(A.21)
This family of functions are useful to bound the sum of the stepsizes.

**Theorem A.3** Assume that conditions (A1), (A2), (A3) are satisfied. Suppose that $H$ is $\mu$-strongly convex and $G$ is $\nu$-strongly convex, for some $\mu \in [0, +\infty[$ and $\nu \in [0, +\infty[$, with $\mu + \nu > 0$. Suppose that there exist $\lambda \in ]0, 1[$ and $\tilde{\alpha} \in [0, +\infty[$ such that $\inf_{n \in \mathbb{N}^*} \lambda_n \geq \lambda$ and $\sup_{n \in \mathbb{N}^*} \alpha_n \leq \tilde{\alpha}$. Let $c_1 \in ]0, +\infty[$ and let $\theta \in ]0, 1[$. Suppose that, for every $n \in \mathbb{N}^*$, $\gamma_n = c_1 n^{-\theta}$. Set

$$t = 1 - 2^{\theta - 1}, \quad c = \frac{2c_1 \lambda (\nu + \mu \varepsilon)}{(1 + \nu)^2}, \quad \text{and} \quad \tau = \frac{2\sigma^2 c_1^2 (1 + \tilde{\alpha} \parallel \nabla H(w) \parallel^2)}{c^2}.$$  

(A.22)

Let $n_0$ be the smallest integer such that $n_0 > 1$, and $\max \{c, c_1\} n_0^{-\theta} \leq 1$. Let $(w_n)_{n \in \mathbb{N}^*}$ be the sequence generated by the SPG algorithm. Then, by setting $(\forall n \in \mathbb{N}^*) s_n = E[\|w_n - \bar{w}\|^2]$, we have, for every $n \geq 2n_0$,

$$s_{n+1} \leq \begin{cases} 
\left( \tau c^2 \varphi_1(n) + s_n \exp \left( \frac{t n_0}{\tilde{\alpha} - \theta} \right) \right) \exp \left( \frac{-ct (n + 1)^{1-\theta}}{1 - \theta} \right) + \frac{2^\theta \tau c}{(n - 2)^\theta} & \text{if } \theta \in ]0, 1[, \\
\left( \frac{n_0}{n+1} \right) c + \frac{2^\theta \tau c^2}{(n + 1)^c} \varphi_{c-1}(n) & \text{if } \theta = 1.
\end{cases}$$  

(A.23)

**Proof** Since $\mu + \nu > 0$, then $H + G$ is strongly convex. Hence, Problem (2.1) has a unique minimizer $\bar{w}$. Since $\gamma_n G$ is $\gamma_n \nu$-strongly convex, by [4, Proposition 23.11] prox$_{\gamma_n G}$ is $(1 + \gamma_n \nu)$-cocoercive, and then

$$(\forall n \in \mathbb{N}^*) \quad \|y_n - \bar{w}\|^2 \leq \frac{1}{(1 + \gamma_n \nu)^2} \|w_n - \bar{w}\|^2 - \gamma_n \|\xi_n - \nabla H(\bar{w})\|^2.$$  

Next, proceeding as in the proof of Proposition A.2, we get an inequality analogue to (A.17), that is

$$\begin{aligned}
\mathbb{E}[\|y_n - \bar{w}\|^2] &\leq \frac{1}{(1 + \gamma_n \nu)^2} \left( \mathbb{E}[\|w_n - \bar{w}\|^2] - 2\gamma_n \left( 1 - \gamma_n \beta (1 + 2\sigma^2 \alpha_n) \right) \right) \cdot \\
&\quad \cdot \mathbb{E}[\langle w_n - \bar{w} \mid \nabla H(w_n) - \nabla H(\bar{w}) \rangle + 2\gamma_n^2 \sigma^2 (1 + \alpha_n \parallel \nabla H(\bar{w}) \parallel^2)].
\end{aligned}$$  

(A.24)

Since $H$ is strongly convex of parameter $\mu$, it holds $\langle \nabla H(w_n) - \nabla H(\bar{w}) \mid w_n - \bar{w} \rangle \geq \mu \parallel w_n - \bar{w}\|^2$. Therefore, from (A.24), using the $\mu$-strong convexity of $H$ and (A3), we get

$$\begin{aligned}
\mathbb{E}[\|y_n - \bar{w}\|^2] &\leq \frac{1}{(1 + \gamma_n \nu)^2} \left( (1 - 2\gamma_n \mu \varepsilon) \mathbb{E}[\|w_n - \bar{w}\|^2] + 2\sigma^2 \lambda_n \lambda_n^{-1} \right).
\end{aligned}$$  

(A.25)
Hence, by definition of \( w_{n+1} \),
\[
E[\| w_{n+1} - \overline{w} \|^2] \leq \left( 1 - \frac{\lambda_n \gamma_n (2 \nu + \gamma_n \nu^2 + 2 \mu \epsilon)}{(1 + \gamma_n \nu)^2} \right) E[\| w_n - \overline{w} \|^2] + \frac{2 \sigma^2 \chi^2_n}{(1 + \gamma_n \nu)^2}.
\]  
(A.26)

Let \( \gamma_n = c_1 n^{-\theta} \) and fix \( n \geq n_0 \). Since \( \gamma_n \leq \gamma_{n_0} = c_1 n_0^{-\theta} \leq 1 \), we have
\[
\frac{\lambda_n \gamma_n (2 \nu + \gamma_n \nu^2 + 2 \mu \epsilon)}{(1 + \gamma_n \nu)^2} \geq \frac{2 \lambda (\nu + \mu \epsilon)}{(1 + \nu)^2} \gamma_n = cn^{-\theta},
\]  
(A.27)
where we set \( c = c_1 2 \lambda (\nu + \mu \epsilon)/(1 + \nu)^2 \). On the other hand,
\[
\frac{2 \sigma^2 \chi^2_n}{(1 + \gamma_n \nu)^2} \leq 2 \sigma^2 (1 + \nu) \| \nabla H(\overline{w}) \|^2) c_1^2 n^{-2\theta}.
\]  
(A.28)

Then, putting together (A.26), (A.27), and (A.28), we get
\[
E[\| w_{n+1} - \overline{w} \|^2] \leq (1 - \eta_n)E[\| w_n - \overline{w} \|^2] + \tau \eta_n^2,
\]  
with \( \tau = 2\sigma^2 c_1^2 (1 + \nu) \| \nabla H(\overline{w}) \|^2)/c^2 \) and \( \eta_n = cn^{-\theta} \). Finally, (A.23) follows from [40, Lemma 3.1].

We next collect some convergence results that will be useful in the proof of the main Theorem 3.6.

**Proposition A.4** Suppose that (A1), (A2), (A3), and (A4) are satisfied. Let \( (w_n)_{n \in \mathbb{N}^*} \) be a sequence generated by Algorithm 2.2. Then, for any solution \( \overline{w} \) of the problem (2.1), the following hold:

(i) The sequence \( (E[\| w_n - \overline{w} \|^2])_{n \in \mathbb{N}^*} \) converges to a finite value.

(ii) The sequence \( (\| w_n - \overline{w} \|^2)_{n \in \mathbb{N}^*} \) converges a.s to some integrable random variable \( \xi_{\overline{w}} \).

(iii) \( \sum_{n \in \mathbb{N}^*} \lambda_n E[\| w_n - \overline{w} \| \nabla H(w_n) - \nabla H(\overline{w})] < +\infty \). Consequently,
\[
\lim_{n \to \infty} E[\| w_n - \overline{w} \| \nabla H(w_n) - \nabla H(\overline{w})] = 0 \quad \text{and} \quad \lim_{n \to \infty} E[\| w_n - \overline{w} \| \nabla H(w_n) - \nabla H(\overline{w})] = 0.
\]

(iv) \( \sum_{n \in \mathbb{N}^*} \lambda_n E[\| w_n - y_n - \gamma_n (G_n - \nabla H(\overline{w}))\|^2] < +\infty \) and \( \sum_{n \in \mathbb{N}^*} \lambda_n E[\| w_n - y_n \|^2] < +\infty \).

**Proof** By Proposition A.2(i)–(iii), and by condition (A3), we get
\[
E[\| w_{n+1} - \overline{w} \|^2] \leq (1 - \lambda_n)E[\| w_n - \overline{w} \|^2] + \lambda_n E[\| y_n - \overline{w} \|^2] \\
\leq E[\| w_n - \overline{w} \|^2] - 2 \gamma_n \lambda_n E[\| w_n - \overline{w} \| \nabla H(w_n) - \nabla H(\overline{w})] \\
+ 2 \sigma^2 \chi^2_n - \lambda_n E[\| u_n \|^2] \\
\leq E[\| w_n - \overline{w} \|^2] + 2 \sigma^2 \chi^2_n,
\]  
(A.29)
where the last inequality follows by the monotonicity of \( \nabla H \).
(i): Since the sequence $(\chi_{\mathbb{N}}^2 n)_{n \in \mathbb{N}^*}$ is summable by assumption (A4), we derive from (A.29) that $(E[\|w_{n+1} - \bar{w}\|^2])_{n \in \mathbb{N}^*}$ converges to a finite value.

(ii): Using the definition of $w_{n+1}$ in SPG algorithm, and Proposition A.2(iii),
\[
E[\|w_{n+1} - \bar{w}\|^2 | \mathcal{A}_n] \leq (1 - \lambda_n) \|w_n - \bar{w}\|^2 + \lambda_n E[\|y_n - \bar{w}\|^2 | \mathcal{A}_n]
\]
\[
\leq \|w_n - \bar{w}\|^2 - 2\gamma_n \lambda_n (1 - \beta \gamma_n (1 + 2\sigma^2 \alpha_n)) \langle \nabla H(w_n) - \nabla H(\bar{w}) | w_n - \bar{w} \rangle + 2\sigma^2 \chi_n^2
\]
\[
\leq \|w_n - \bar{w}\|^2 - 2\varepsilon \gamma_n \lambda_n \langle \nabla H(w_n) - \nabla H(\bar{w}) | w_n - \bar{w} \rangle + 2\sigma^2 \chi_n^2
\]
\[
\sum_{n \in \mathbb{N}^*} \lambda_n E[\|w_n - \bar{w}\|^2] < +\infty.
\]
(A.30)

Hence, $(w_n)_{n \in \mathbb{N}^*}$ is a random quasi-Fejér sequence (see [20] for the definition) with respect to the nonempty closed and convex set $\text{Argmin} \Phi$.

Taking into account that $E[\|w_1\|^2] < +\infty$ by assumption, it follows from [12, Proposition 2.3(iii)] that $(\|w_n - \bar{w}\|^2)_{n \in \mathbb{N}^*}$ converges a.s to some integrable random variable $\zeta_w$.

(iii): We derive from (A.29) that
\[
\sum_{n \in \mathbb{N}^*} \gamma_n \lambda_n E[\langle w_n - \bar{w} | \nabla H(w_n) - \nabla H(\bar{w}) \rangle] < +\infty.
\]
(A.31)

Since $\sum_{n \in \mathbb{N}^*} \lambda_n \gamma_n = +\infty$, we obtain
\[
\lim_{n \to \infty} E[\langle w_n - \bar{w} | \nabla H(w_n) - \nabla H(\bar{w}) \rangle] = 0 \Rightarrow \lim_{n \to \infty} E[\|\nabla H(w_n) - \nabla H(\bar{w})\|^2] = 0,
\]
(A.32)

using again the cocoercivity of $\nabla H$.

(iv) We directly get from (A.29) that $\sum_{n \in \mathbb{N}^*} \lambda_n \gamma_n \|u_n\|^2 < +\infty$.

Since $\nabla H$ is Lipschitz-continuous, and $(E[\|w_n - \bar{w}\|^2])_{n \in \mathbb{N}^*}$ is convergent by (i), there exists $M \in [0, +\infty]$ such that
\[
(\forall n \in \mathbb{N}^*) \ E[\langle w_n - \bar{w} | \nabla H(w_n) - \nabla H(\bar{w}) \rangle] \leq \beta E[\|w_n - \bar{w}\|^2] \leq M < +\infty.
\]
(A.33)

Hence, we derive from (A.20) and (2.4) that
\[
\sum_{n \in \mathbb{N}^*} \lambda_n \gamma_n^2 E[\|\mathcal{G}_n - \nabla H(\bar{w})\|^2] < +\infty.
\]
(A.34)

Now, recalling the definition of $u_n$ in (A.15), using (A.34) and (A.29), we obtain
\[
\sum_{n \in \mathbb{N}^*} \lambda_n E[\|w_n - y_n\|^2] \leq 2 \sum_{n \in \mathbb{N}} \lambda_n E[\|u_n\|^2] + 2 \sum_{n \in \mathbb{N}^*} \lambda_n \gamma_n^2 E[\|\mathcal{G}_n - \nabla H(\bar{w})\|^2] < +\infty.
\]
(A.35)
Proof of Theorem 3.6  Since $H$ is uniformly convex at $\overline{w}$, there exists $\phi: [0, +\infty] \to [0, +\infty]$ increasing and vanishing only at 0 such that

$$
\langle \nabla H(w_n) - \nabla H(\overline{w}) \mid w_n - \overline{w} \rangle \geq \phi(\|w_n - \overline{w}\|).
$$

(A.36)

Therefore, we derive from Proposition A.4(iii) that

$$
\sum_{n \in \mathbb{N}^*} \lambda_n \gamma_n \mathbb{E}[\phi(\|w_n - \overline{w}\|)] < \infty, \text{ and hence}
$$

$$
\sum_{n \in \mathbb{N}^*} \lambda_n \gamma_n \phi(\|w_n - \overline{w}\|) < \infty \text{ a.s. (A.37)}
$$

Since $(\lambda_n \gamma_n)_{n \in \mathbb{N}^*}$ is not summable, we have \(\lim \phi(\|w_n - \overline{w}\|) = 0\) a.s. Consequently, there exists a subsequence $(k_n)_{n \in \mathbb{N}^*}$ such that $\phi(\|w_{k_n} - \overline{w}\|) \to 0$ a.s, which implies that $\|w_{k_n} - \overline{w}\| \to 0$ a.s. In view of Proposition A.4(ii), we get $w_n \to \overline{w}$ a.s.

\[\Box\]

Proof of Theorem 3.7  By Proposition A.4(i), $(\|w_n - \overline{w}\|^2)_{n \in \mathbb{N}^*}$ converges to an integrable random variable, hence it is uniformly bounded. Moreover, $\lim \mathbb{E}[\|\nabla H(w_n) - \nabla H(\overline{w})\|^2] = 0$, and hence there exists a subsequence $(k_n)_{n \in \mathbb{N}^*}$ such that $\lim_{n \to \infty} \mathbb{E}[\|\nabla H(w_{k_n}) - \nabla H(\overline{w})\|^2] = 0$. Thus, there exists a subsequence $(p_n)_{n \in \mathbb{N}^*}$ of $(k_n)_{n \in \mathbb{N}^*}$ such that

$$
\|\nabla H(w_{p_n}) - \nabla H(\overline{w})\|^2 \to 0 \text{ a.s. (A.38)}
$$

Let $\overline{z}$ be a weak cluster point of $(w_{p_n})_{n \in \mathbb{N}^*}$, then there exists a subsequence $(w_{q_{p_n}})_{n \in \mathbb{N}^*}$ such that for almost all $\omega$, $w_{q_{p_n}}(\omega) \rightharpoonup \overline{z}(\omega)$. Since $\nabla H$ is weakly continuous, for almost all $\omega$, $\nabla H(w_{q_{p_n}}(\omega)) \rightharpoonup \nabla H(\overline{z}(\omega))$. Therefore, for almost every $\omega$, by (A.38), $
abla H(\overline{w}) = \nabla H(\overline{z}(\omega))$, and hence

$$
\langle \nabla H(\overline{z}(\omega)) - \nabla H(\overline{w}) \mid \overline{z}(\omega) - \overline{w} \rangle = 0.
$$

Since $H$ is strictly convex, $\nabla H$ is strictly monotone, we obtain $\overline{w} = \overline{z}(\omega)$. This shows that $w_{q_{p_n}} \to \overline{w}$ a.s. \[\Box\]

Proof of Remark 3.9  Let $w$ be a weak cluster point of $(w_n)_{n \in \mathbb{N}^*}$, i.e., there exists a subsequence $(w_{k_n})_{n \in \mathbb{N}^*}$ such that $w_{k_n} \rightharpoonup w$ a.s. Since $\Phi = H + G$ is convex and lower semicontinous, it is weakly lower semicontinous, hence

$$
\Phi(w) \leq \lim \Phi(w_{k_n}) = \inf \Phi, \text{ (A.39)}
$$

which shows that $w \in \text{Argmin } \Phi$ a.s. We therefore conclude that $(w_n)_{n \in \mathbb{N}^*}$ converges weakly to an optimal solution a.s. \[\Box\]

References

1. Atchade, Y.F., Fort, G., Moulines, E.: On perturbed proximal gradient algorithms. J. Mach. Learn. Res. 18, 310–342 (2017)
2. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Structured sparsity through convex optimization. Stat. Sci. 27, 450–468 (2012)
3. Bach, F., Moulies, E.: Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In: Advances in Neural Information Processing Systems, vol. 24 (2011)
4. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, New York (2011)
5. Benveniste, A., Métévier, M., Priouret, P.: Adaptive Algorithms and Stochastic Approximations, vol. 22. Springer, Berlin (1990)
6. Bertsekas, D.P.: Incremental gradient, subgradient, and proximal methods for convex optimization: a survey. In: Optimization for Machine Learning, pp. 85–104 (2011)
7. Bertsekas, D.P., Tsitsiklis, J.N.: Gradient convergence in gradient methods with errors. SIAM J. Optim. 10, 627–642 (2000). (electronic)
8. Bianchi, P., Hachem, W.: Dynamical behavior of a stochastic forward-backward algorithm using random monotone operators. J. Optim. Theory Appl. 171, 90–120 (2016)
9. Blatt, D., Hero, A., Gauchman, H.: A convergent incremental gradient method with a constant step size. SIAM J. Optim. 18, 29–51 (2007)
10. Bottou, L., Curtis, F., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Rev. 60, 223–311 (2018)
11. Cesa-Bianchi, N., Lugosi, G.: Prediction, Learning, and Games. Cambridge University Press, Cambridge (2006)
12. Combettes, P.L., Pesquet, J.-C.: Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping. SIAM J. Optim. 25, 1221–1248 (2015)
13. Combettes, P.L., Pesquet, J.-C.: Stochastic approximations and perturbations in forward-backward splitting for monotone operators. Pure Appl. Funct. Anal. 1, 13–37 (2016)
14. Combettes, P.L., Pesquet, J.-C.: Proximal splitting methods in signal processing. In: Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 185–212. Springer, New York (2011)
15. Combettes, P.L., Wajs, Valérie R.: Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. 4, 1168–1200 (2005). (electronic)
16. De Mol, C., De Vito, E., Rosasco, L.: Elastic-net regularization in learning theory. J. Complex. 25, 201–230 (2009)
17. Defazio, A., Bach, F., Lacoste-Julien, S.: SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. Adv. Neural Inf. Process. Syst. 27, 1646–1654 (2014)
18. Duchi, J., Singer, Y.: Efficient online and batch learning using forward backward splitting. J. Mach. Learn. Res. 10, 2899–2934 (2009)
19. Ermol’ev, YuM: On the method of generalized stochastic gradients and quasi-fejér sequences. Cybernetics 5, 208–220 (1969)
20. Ermol’ev, YuM, Tuniev, A.D.: Random fejér and quasi-fejér sequences. Theory Optim. Solut. Akademiya Nauk Ukrainskoi SSR Kiev 2, 76–83 (1968)
21. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, I: a generic algorithmic framework. SIAM J. Optim. 22, 1469–1492 (2012)
22. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. SIAM J. Optim. 23, 2061–2089 (2013)
23. Hazan, E., Kale, S.: Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. J. Mach. Learn. Res. 15, 2489–2512 (2014)
24. Jofré, A., Thompson, P.: On variance reduction for stochastic smooth convex optimization with multiplicative noise. In: Mathematical Programming, Series B, Published Online, pp. 1–40 (2018)
25. Juditski, A., Nesterov, Y.: Deterministic and stochastic primal-dual subgradient methods for uniformly convex minimization. Stoch. Syst. 4, 44–80 (2014)
26. Kiefer, J., Wolfowitz, J.: Stochastic estimation of the maximum of a regression function. Ann. Math. Stat. 23, 462–466 (1952)
27. Kushner, H.J., Yin, G.G.: Stochastic Approximation Algorithms and Applications, vol. 35. Springer, New York (1997)
28. Kwok, J.T., Hu, C., Pan, W.: Accelerated gradient methods for stochastic optimization and online learning. Adv. Neural Inf. Process. Syst. 22, 781–789 (2009)
29. Lan, G.: An optimal method for stochastic composite optimization. Math. Progr. 133, 365–397 (2012)
30. Lin, J., Rosasco, L., Villa, S., Zhou, D.-X.: Modified Fejér sequences and applications. Comput. Optim. Appl. 71, 95–113 (2018)
31. Lin, Q., Chen, X., Peña, J.: A sparsity preserving stochastic gradient methods for sparse regression. Comput. Optim. Appl. 58, 388–402 (2014)
32. Mosci, S., Rosasco, L., Santoro, M., Verri, A., Villa, S.: Solving structured sparsity regularization with proximal methods. In: Machine Learning and Knowledge discovery in Databases European Conference, ECML PKDD, pp. 418–433. Springer, Berlin (2010)
33. Nekrylova, Z.V.: The rate of convergence of the stochastic gradient method. Cybernetics 11, 218–222 (1975)
34. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. 19, 1574–1609 (2008)
35. Nemirovski, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Wiley, New York (1983)
36. Polyak, B.: Introduction to Optimization. Optimization Software, New York (1987)
37. Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. SIAM J. Control Optim. 30, 838–855 (1992)
38. Rakhlin, A., Shamir, O., Sridaran, K.: Making gradient descent optimal for strongly convex stochastic optimization. In: Proceedings of the 29th International Conference on Machine Learning, pp. 449–456 (2012)
39. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. 22, 400–407 (1951)
40. Rosasco, L., Villa, S., Vu, B.C.: Stochastic forward-backward splitting for monotone inclusions. J. Optim. Theory Appl. 169, 1458–1466 (2011)
41. Rosasco, L., Villa, S., Vu, B.C.: A first-order stochastic primal-dual algorithm with correction step. Numer. Funct. Anal. Optim. 38, 602–626 (2017)
42. Schmidt, M., Le Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. Math. Progr. Ser. B 162, 83–112 (2017)
43. Schmidt, M.W., Le Roux, N., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization. Adv. Neural Inf. Process. Syst. 24, 1458–1466 (2011)
44. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, Cambridge (2014)
45. Shalev-Shwartz, S., Singer, Y., Srebro, N., Cotter, A.: Pegasos: primal estimated sub-gradient solver for SVM. Math. Progr. Ser. B 127, 3–30 (2011)
46. Shalev-Shwartz, S., Zhang, T.: Accelerated mini-batch stochastic dual coordinate ascent. In: Advances in Neural Information Processing Systems, vol. 26 (2013)
47. Shamir, O., Zhang, T.: Stochastic gradient descent for non-smooth optimization: convergence results and optimal averaging schemes. In: Proceedings of the 30th International Conference on Machine Learning, pp. 71–79 (2013)
48. Shor, N.Z.: Minimization Methods for Non-differentiable Functions. Springer, Berlin (1985)
49. Villa, S., Rosasco, L., Mosci, S., Verri, A.: Proximal methods for the latent group lasso penalty. Comput. Optim. Appl. 58, 381–407 (2014)
50. Villa, S., Salzo, S., Baldassarre, L., Verri, A.: Accelerated and inexact forward-backward algorithms. SIAM J. Optim. 23, 1607–1633 (2013)
51. Xiao, L.: Dual averaging methods for regularized stochastic learning and online optimization. J. Mach. Learn. Res. 11, 2543–2596 (2010)
52. Zou, Z., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67, 301–320 (2005)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.