# Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization 

Saeed Ghadimi • Guanghui Lan • Hongchao Zhang

Received: 24 August 2013 / Accepted: 19 November 2014
© Springer-Verlag Berlin Heidelberg and Mathematical Optimization Society 2014


#### Abstract

This paper considers a class of constrained stochastic composite optimization problems whose objective function is given by the summation of a differentiable (possibly nonconvex) component, together with a certain non-differentiable (but convex) component. In order to solve these problems, we propose a randomized stochastic projected gradient (RSPG) algorithm, in which proper mini-batch of samples are taken at each iteration depending on the total budget of stochastic samples allowed. The RSPG algorithm also employs a general distance function to allow taking advantage of the geometry of the feasible region. Complexity of this algorithm is established in a unified setting, which shows nearly optimal complexity of the algorithm for convex stochastic programming. A post-optimization phase is also proposed to significantly reduce the variance of the solutions returned by the algorithm. In addition, based on the RSPG algorithm, a stochastic gradient free algorithm, which only uses the stochastic zeroth-order information, has been also discussed. Some preliminary numerical results are also provided.


[^0]Keywords Constrained stochastic programming • Mini-batch of samples .
Stochastic approximation • Nonconvex optimization • Stochastic programming . First-order method • Zeroth-order method

Mathematics Subject Classification 90C25 •90C06 - 90C22 • 49M37

## 1 Introduction

In this paper, we consider the following problem

$$
\begin{equation*}
\Psi^{*}:=\min _{x \in X}\{\Psi(x):=f(x)+h(x)\}, \tag{1.1}
\end{equation*}
$$

where $X$ is a closed convex set in Euclidean space $\mathbb{R}^{n}, f: X \rightarrow \mathbb{R}$ is continuously differentiable, but possibly nonconvex, and $h$ is a simple convex function with known structure, but possibly nonsmooth [e.g., $h(x)=\|x\|_{1}$ or $h(x) \equiv 0$ ]. We also assume that the gradient of $f$ is $L$-Lipschitz continuous for some $L>0$, i.e.,

$$
\begin{equation*}
\|\nabla f(y)-\nabla f(x)\| \leq L\|y-x\|, \quad \text { for any } \quad x, y \in X \tag{1.2}
\end{equation*}
$$

and $\Psi$ is bounded below over $X$, i.e., $\Psi^{*}$ is finite. Although $f$ is Lipschitz continuously differentiable, we assume that only the noisy gradient of $f$ is available via subsequent
 the input $x_{k} \in X, \mathcal{S F} \mathcal{O}$ would output a stochastic gradient $G\left(x_{k}, \xi_{k}\right)$, where $\xi_{k}$ is a random variable whose distribution is supported on $\Xi_{k} \subseteq \mathbb{R}^{d}$. Throughout the paper, we make the following assumptions for the Borel functions $G\left(x_{k}, \xi_{k}\right)$.

A1 For any $k \geq 1$, we have

$$
\begin{align*}
& \text { a) }\left[G\left(x_{k}, \xi_{k}\right)\right]=\nabla f\left(x_{k}\right),  \tag{1.3}\\
& \text { b) }\left[\left\|G\left(x_{k}, \xi_{k}\right)-\nabla f\left(x_{k}\right)\right\|^{2}\right] \leq \sigma^{2} \tag{1.4}
\end{align*}
$$

where $\sigma>0$ is a constant. Note that part b ) of A 1 is slightly weaker than the usual assumption that $E\left[\|G(x, \xi)\|^{2}\right]$ is bounded in general stochastic optimization. For some examples which fit our setting, one may refer the problems in references $[1,12$, $13,16,17,23-25,34]$.

Stochastic programming (SP) problems have been the subject of intense studies for more than 50 years. In the seminal 1951 paper, Robbins and Monro [32] proposed a classical stochastic approximation (SA) algorithm for solving SP problems. Although their method has "asymptotically optimal" rate of convergence for solving a class of strongly convex SP problems, the practical performance of their method is often poor (e.g., [36, Section 4.5.3]). Later, Polyak [30] and Polyak and Juditsky [31] proposed important improvements to the classical SA algorithms, where larger stepsizes were allowed in their methods. Recently, there have been some important developments of SA algorithms for solving convex SP problems [i.e., $\Psi$ in (1.1) is a convex function].

Motivated by the complexity theory in convex optimization [27], these studies focus on the convergence properties of SA-type algorithms in a finite number of iterations. For example, Nemirovski et al. [26] presented a mirror descent SA approach for solving general nonsmooth convex stochastic programming problems. They showed that the mirror descent SA exhibits an optimal $\mathcal{O}\left(1 / \epsilon^{2}\right)$ iteration complexity for solving these problems with an essentially unimprovable constant factor. Lan [21] presented a unified optimal method for smooth, nonsmooth and stochastic optimization. This unified optimal method also leads to optimal methods for strongly convex problems [14,15]. Duchi et al. [10] also presented modified mirror descent method for solving convex stochastic composite problems. Recently, some stochastic gradient type methods have been developed for solving (strongly) convex "finite-batch" problems which can achieve faster convergence rates (see e.g., [35]). However, all of the above mentioned methods need the convexity of the problem to establish their convergence and cannot deal with the situations where the objective function is not necessarily convex.

When problem (1.1) is nonconvex, the research on SP algorithms so far is very limited and still far from mature. For the deterministic case, i.e., $\sigma=0$ in (1.4), the complexity of the gradient descent method for solving problem (1.1) has been studied in [6,28] (see also [37] for a proximal splitting algorithm for deterministic nonconvex composite problems). Very recently, Ghadimi and Lan [16] proposed an SA-type algorithm coupled with a randomization scheme, namely, a randomized stochastic gradient (RSG) method, for solving the unconstrained nonconvex SP problem, i.e., problem (1.1) with $h \equiv 0$ and $X=\mathbb{R}^{n}$. In their algorithm, a trajectory $\left\{x_{1}, \ldots, x_{N}\right\}$ is generated by a stochastic gradient descent method, and a solution $\bar{x}$ is randomly selected from this trajectory according to a certain probability distribution. They showed that the number of calls to the $\mathcal{S F} \mathcal{F}$ required by this algorithm to find an $\epsilon$-solution, i.e., a point $\bar{x}$ such that $\mathbb{E}\left[\|\nabla f(\bar{x})\|_{2}^{2}\right] \leq \epsilon$, is bounded by $\mathcal{O}\left(\sigma^{2} / \epsilon^{2}\right)$. They also presented a variant of the RSG algorithm, namely, a two-phase randomized stochastic gradient (2-RSG) algorithm to improve the large-deviation results of the RSG algorithm. Specifically, they showed that the complexity of the 2-RSG algorithm for computing an $(\epsilon, \Lambda)$-solution, i.e., a point $\bar{x}$ satisfying $\operatorname{Prob}\left\{\|\nabla f(\bar{x})\|_{2}^{2} \leq \epsilon\right\} \geq 1-\Lambda$, for some $\epsilon>0$ and $\Lambda \in(0,1)$, can be bounded by

$$
\mathcal{O}\left\{\frac{\log (1 / \Lambda) \sigma^{2}}{\epsilon}\left[\frac{1}{\epsilon}+\frac{\log (1 / \Lambda)}{\Lambda}\right]\right\}
$$

They also specialized the RSG algorithm and presented a randomized stochastic gradient free (RSGF) algorithm for the situations where only noisy function values are available. It is shown that the expected complexity of this RSGF algorithm is $\mathcal{O}\left(n \sigma^{2} / \epsilon^{2}\right)$.

While the RSG algorithm and its variants can handle the unconstrained nonconvex SP problems, their convergence cannot be guaranteed for stochastic composite optimization problems in (1.1) where $X \neq \mathbb{R}^{n}$ and/or $h(\cdot)$ is non-differentiable. Our contributions in this paper mainly consist of developing variants of the RSG algorithm by taking a mini-batch of samples at each iteration of our algorithm to deal with the constrained composite problems while preserving the complexity results. More specifically, we first modify the scheme of the RSG algorithm to propose a randomized stochastic projected gradient (RSPG) algorithm to solve constrained nonconvex
stochastic composite problems. Unlike the RSG algorithm, at each iteration of the RSPG algorithm, we take multiple samples such that the total number of calls to the $\mathcal{S F O}$ to find a solution $\bar{x} \in X$ such that $\mathbb{E}\left[\left\|g_{X}(\bar{x})\right\|^{2}\right] \leq \epsilon$, is still $\mathcal{O}\left(\sigma^{2} / \epsilon^{2}\right)$, where $g_{X}(\bar{x})$ is a generalized projected gradient of $\Psi$ at $\bar{x}$ over X. In addition, our RSPG algorithm is in a more general setting depending on a general distance function rather than Euclidean distance [16]. This would be particularly useful for special structured constrained set (e.g., $X$ being a standard simplex). Secondly, we present a two-phase randomized stochastic projected gradient (2-RSPG) algorithm, the RSPG algorithm with a post-optimization phase, to improve the large-deviation results of the RSPG algorithm. And we show that the complexity of this approach can be further improved under a light-tail assumption about the $\mathcal{S F O}$. Thirdly, under the assumption that the gradient of $f$ is also bounded on $X$, we specialize the RSPG algorithm to give a randomized stochastic projected gradient free (RSPGF) algorithm, which only uses the stochastic zeroth-order information. Finally, we present some numerical results to show the effectiveness of the aforementioned randomized stochastic projected gradient algorithms, including the RSPG, 2-RSPG and RSPGF algorithms. Some practical improvements of these algorithms are also discussed.

The remaining part of this paper is organized as follows. We first describe some properties of the projection based on a general distance function in Sect. 2. In Sect. 3, a deterministic first-order method for problem (1.1) is proposed, which mainly provides a basis for our stochastic algorithms developed in later sections. Then, by incorporating a randomized scheme, we present the RSPG and 2-RSPG algorithms for solving the SP problem (1.1) in Sect. 4. In Sect. 5, we discuss how to generalize the RSPG algorithm to the case when only zeroth-order information is available. Some numerical results and discussions from implementing our algorithms are presented in Sect. 6. Finally, in Sect. 7, we give some concluding remarks.

Notation We use $\|\cdot\|$ to denote a general norm without specific mention. Also, for any $p \geq 1,\|\cdot\|_{p}$ denote the standard $p$-norm in $\mathbb{R}^{n}$, i.e.,

$$
\|x\|_{p}^{p}=\sum_{i=1}^{n}\left|x_{i}\right|^{p}, \quad \text { for any } \quad x \in \mathbb{R}^{n}
$$

For any convex function $h, \partial h(x)$ is the subdifferential set at $x$. Given any $\Omega \subseteq \mathbb{R}^{n}$, we say $f \in \mathcal{C}_{L}^{1,1}(\Omega)$, if $f$ is Lipschitz continuously differentiable with Lipschitz constant $L>0$, i.e.,

$$
\begin{equation*}
\|\nabla f(y)-\nabla f(x)\| \leq L\|y-x\|, \quad \text { for any } \quad x, y \in \Omega, \tag{1.5}
\end{equation*}
$$

which clearly implies

$$
\begin{equation*}
|f(y)-f(x)-\langle\nabla f(x), y-x\rangle| \leq \frac{L}{2}\|y-x\|^{2}, \quad \text { for any } \quad x, y \in \Omega \tag{1.6}
\end{equation*}
$$

For any real number $r,\lceil r\rceil$ and $\lfloor r\rfloor$ denote the nearest integer to $r$ from above and below, respectively. $\mathbb{R}_{+}$denotes the set of nonnegative real numbers.

## 2 Some properties of generalized projection

In this section, we review the concept of projection in a general sense as well as its important properties. This section consists of two subsections. We first discuss the concept of prox-function and its associated projection in Sect. 2.1. Then, in Sect. 2.2, we present some important properties of the projection, which will play a critical role for the proofs in our later sections.

### 2.1 Prox-function and projection

It is well-known that using a generalized distance generating function, instead of the usual Euclidean distance function, would lead to algorithms that can be adjusted to the geometry of the feasible set and/or efficient solutions of the projection [2,3,5,21,26, 38]. Hence, in this paper we would like to set up the projection based on the so-called prox-function.

A function $\omega: X \rightarrow \mathbb{R}$ is said to be a distance generating function with modulus $\alpha>0$ with respect to $\|\cdot\|$, if $\omega$ is continuously differentiable and strongly convex satisfying

$$
\begin{equation*}
\langle x-z, \nabla \omega(x)-\nabla \omega(z)\rangle \geq \alpha\|x-z\|^{2}, \quad \forall \quad x, z \in X \tag{2.1}
\end{equation*}
$$

Then, the prox-function associated with $\omega$ is defined as

$$
\begin{equation*}
V(x, z)=\omega(x)-[\omega(z)+\langle\nabla \omega(z), x-z\rangle] . \tag{2.2}
\end{equation*}
$$

In this paper, we assume that the prox-function $V$ is chosen such that the generalized projection problem given by

$$
\begin{equation*}
x^{+}=\arg \min _{u \in X}\left\{\langle g, u\rangle+\frac{1}{\gamma} V(u, x)+h(u)\right\} \tag{2.3}
\end{equation*}
$$

is easily solvable for any $\gamma>0, g \in \mathbb{R}^{n}$ and $x \in X$. Apparently, different choices of $\omega$ can be used in the definition of prox-function. One simple example would be $\omega(x)=\|x\|_{2}^{2} / 2$, which gives $V(x, z)=\|x-z\|_{2}^{2} / 2$. And in this case, if $h(x) \equiv 0, x^{+}$ is just the usual Euclidean projection. Some less trivial examples can be found, e.g., in [2,4, 8, 19, 27].

### 2.2 Properties of projection

In this subsection, we discuss some important properties of the generalized projection defined in (2.3). Let us first define

$$
\begin{equation*}
P_{X}(x, g, \gamma)=\frac{1}{\gamma}\left(x-x^{+}\right), \tag{2.4}
\end{equation*}
$$

where $x^{+}$is given in (2.3). We can see that $P_{X}(x, \nabla f(x), \gamma)$ can be viewed as a generalized projected gradient of $\Psi$ at $x$. Indeed, if $X=\mathbb{R}^{n}$ and $h$ vanishes, we would have $P_{X}(x, \nabla f(x), \gamma)=\nabla f(x)=\nabla \Psi(x)$.

The following lemma provides a bound for the size of $P_{X}(x, g, \gamma)$.
Lemma 1 Let $x^{+}$be given in (2.3). Then, for any $x \in X, g \in \mathbb{R}^{n}$ and $\gamma>0$, we have

$$
\begin{equation*}
\left\langle g, P_{X}(x, g, \gamma)\right\rangle \geq \alpha\left\|P_{X}(x, g, \gamma)\right\|^{2}+\frac{1}{\gamma}\left[h\left(x^{+}\right)-h(x)\right] . \tag{2.5}
\end{equation*}
$$

Proof By the optimality condition of (2.3) and the definition of prox-function in (2.2), there exists a $p \in \partial h\left(x^{+}\right)$such that

$$
\left\langle g+\frac{1}{\gamma}\left[\nabla \omega\left(x^{+}\right)-\nabla \omega(x)\right]+p, u-x^{+}\right\rangle \geq 0, \quad \text { for any } \quad u \in X .
$$

Letting $u=x$ in the above inequality, by the convexity of $h$ and (2.1), we obtain

$$
\begin{aligned}
\left\langle g, x-x^{+}\right\rangle & \geq \frac{1}{\gamma}\left\langle\nabla \omega\left(x^{+}\right)-\nabla \omega(x), x^{+}-x\right\rangle+\left\langle p, x^{+}-x\right\rangle \\
& \geq \frac{\alpha}{\gamma}\left\|x^{+}-x\right\|^{2}+\left[h\left(x^{+}\right)-h(x)\right]
\end{aligned}
$$

which in the view of (2.4) and $\gamma>0$ clearly imply (2.5).

It is well-known [33] that the Euclidean projection is Lipschitz continuous. Below, we show that this property also holds for the general projection.

Lemma 2 Let $x_{1}^{+}$and $x_{2}^{+}$be given in (2.3) with $g$ replaced by $g_{1}$ and $g_{2}$ respectively. Then,

$$
\begin{equation*}
\left\|x_{2}^{+}-x_{1}^{+}\right\| \leq \frac{\gamma}{\alpha}\left\|g_{2}-g_{1}\right\| . \tag{2.6}
\end{equation*}
$$

Proof By the optimality condition of (2.3), for any $u \in X$, there exist $p_{1} \in \partial h\left(x_{1}^{+}\right)$ and $p_{2} \in \partial h\left(x_{2}^{+}\right)$such that

$$
\begin{equation*}
\left\langle g_{1}+\frac{1}{\gamma}\left[\nabla \omega\left(x_{1}^{+}\right)-\nabla \omega(x)\right]+p_{1}, u-x_{1}^{+}\right\rangle \geq 0 \tag{2.7}
\end{equation*}
$$

and

$$
\begin{equation*}
\left\langle g_{2}+\frac{1}{\gamma}\left[\nabla \omega\left(x_{2}^{+}\right)-\nabla \omega(x)\right]+p_{2}, u-x_{2}^{+}\right\rangle \geq 0 \tag{2.8}
\end{equation*}
$$

Letting $u=x_{2}^{+}$in (2.7), by the convexity of $h$, we have

$$
\begin{align*}
\left\langle g_{1}, x_{2}^{+}-x_{1}^{+}\right\rangle \geq & \frac{1}{\gamma}\left\langle\nabla \omega(x)-\nabla \omega\left(x_{1}^{+}\right), x_{2}^{+}-x_{1}^{+}\right\rangle \\
& +\left\langle p_{1}, x_{1}^{+}-x_{2}^{+}\right\rangle \geq \frac{1}{\gamma}\left\langle\nabla \omega\left(x_{2}^{+}\right)-\nabla \omega\left(x_{1}^{+}\right), x_{2}^{+}-x_{1}^{+}\right\rangle \\
& +\frac{1}{\gamma}\left\langle\nabla \omega(x)-\nabla \omega\left(x_{2}^{+}\right), x_{2}^{+}-x_{1}^{+}\right\rangle+h\left(x_{1}^{+}\right)-h\left(x_{2}^{+}\right) . \tag{2.9}
\end{align*}
$$

Similarly, letting $u=x_{1}^{+}$in (2.8), we have

$$
\begin{align*}
\left\langle g_{2}, x_{1}^{+}-x_{2}^{+}\right\rangle & \geq \frac{1}{\gamma}\left\langle\nabla \omega(x)-\nabla \omega\left(x_{2}^{+}\right), x_{1}^{+}-x_{2}^{+}\right\rangle+\left\langle p_{2}, x_{2}^{+}-x_{1}^{+}\right\rangle \\
& \geq \frac{1}{\gamma}\left\langle\nabla \omega(x)-\nabla \omega\left(x_{2}^{+}\right), x_{1}^{+}-x_{2}^{+}\right\rangle+h\left(x_{2}^{+}\right)-h\left(x_{1}^{+}\right) . \tag{2.10}
\end{align*}
$$

Summing up (2.9) and (2.10), by the strong convexity (2.1) of $\omega$, we obtain

$$
\left\|g_{1}-g_{2}\right\|\left\|x_{2}^{+}-x_{1}^{+}\right\| \geq\left\langle g_{1}-g_{2}, x_{2}^{+}-x_{1}^{+}\right\rangle \geq \frac{\alpha}{\gamma}\left\|x_{2}^{+}-x_{1}^{+}\right\|^{2},
$$

which gives (2.6).
As a consequence of the above lemma, we have $P_{X}(x, \cdot, \gamma)$ is Lipschitz continuous.

Proposition 1 Let $P_{X}(x, g, \gamma)$ be defined in (2.4). Then, for any $g_{1}$ and $g_{2}$ in $\mathbb{R}^{n}$, we have

$$
\begin{equation*}
\left\|P_{X}\left(x, g_{1}, \gamma\right)-P_{X}\left(x, g_{2}, \gamma\right)\right\| \leq \frac{1}{\alpha}\left\|g_{1}-g_{2}\right\| . \tag{2.11}
\end{equation*}
$$

Proof Noticing (2.4), (2.7) and (2.8), we have

$$
\begin{aligned}
\left\|P_{X}\left(x, g_{1}, \gamma\right)-P_{X}\left(x, g_{2}, \gamma\right)\right\| & =\left\|\frac{1}{\gamma}\left(x-x_{1}^{+}\right)-\frac{1}{\gamma}\left(x-x_{2}^{+}\right)\right\| \\
& =\frac{1}{\gamma}\left\|x_{2}^{+}-x_{1}^{+}\right\| \leq \frac{1}{\alpha}\left\|g_{1}-g_{2}\right\|,
\end{aligned}
$$

where the last inequality follows from (2.6).
The following lemma (see e.g., Lemma 1 of [21] and Lemma 2 of [14]) characterizes the solution of the generalized projection.

Lemma 3 Let $x^{+}$be given in (2.3). Then, for any $u \in X$, we have

$$
\begin{equation*}
\left\langle g, x^{+}\right\rangle+h\left(x^{+}\right)+\frac{1}{\gamma} V\left(x^{+}, x\right) \leq\langle g, u\rangle+h(u)+\frac{1}{\gamma}\left[V(u, x)-V\left(u, x^{+}\right)\right] . \tag{2.12}
\end{equation*}
$$

## 3 Deterministic first-order methods

In this section, we consider the problem (1.1) with $f \in \mathcal{C}_{L}^{1,1}(X)$, and for each input $x_{k} \in X$, we assume that the exact gradient $\nabla f\left(x_{k}\right)$ is available. Using the exact gradient information, we give a deterministic projected gradient (PG) algorithm for solving (1.1), which mainly provides a basis for us to develop the stochastic first-order algorithms in the next section.
A projected gradient $(P G)$ algorithm
Input: Given initial point $x_{1} \in X$, total number of iterations $N$, and the stepsizes $\left\{\gamma_{k}\right\}$ with $\gamma_{k}>0, k \geq 1$.
Step $k=1, \ldots, N$. Compute

$$
\begin{equation*}
x_{k+1}=\arg \min _{u \in X}\left\{\left\langle\nabla f\left(x_{k}\right), u\right\rangle+\frac{1}{\gamma_{k}} V\left(u, x_{k}\right)+h(u)\right\} . \tag{3.1}
\end{equation*}
$$

Output: $x_{R} \in\left\{x_{k}, \ldots, x_{N}\right\}$ such that

$$
\begin{equation*}
R=\arg \min _{k \in\{1, \ldots, N\}}\left\|g_{X, k}\right\| \tag{3.2}
\end{equation*}
$$

where the $g_{X, k}$ is given by

$$
\begin{equation*}
g_{X, k}=P_{X}\left(x_{k}, \nabla f\left(x_{k}\right), \gamma_{k}\right) . \tag{3.3}
\end{equation*}
$$

We can see that the above algorithm outputs the iterate with the minimum norm of the generalized projected gradients. In practice, one may choose the solution with the minimum function value as the output of the algorithm. However, since $f$ may not be a convex function, we cannot provide theoretical performance guarantee for such a selection of the output solution. In the above algorithm, we have not specified the selection of the stepsizes $\left\{\gamma_{k}\right\}$. We will return to this issue after establishing the following convergence results.

Theorem 1 Suppose that the stepsizes $\left\{\gamma_{k}\right\}$ in the PG algorithm are chosen such that $0<\gamma_{k} \leq 2 \alpha / L$ with $\gamma_{k}<2 \alpha / L$ for at least one $k$. Then, we have

$$
\begin{equation*}
\left\|g_{X, R}\right\|^{2} \leq \frac{L D_{\Psi}^{2}}{\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2} / 2\right)} \tag{3.4}
\end{equation*}
$$

where

$$
\begin{equation*}
g_{X, R}=P_{X}\left(x_{R}, \nabla f\left(x_{R}\right), \gamma_{R}\right) \quad \text { and } \quad D_{\Psi}:=\left[\frac{\left(\Psi\left(x_{1}\right)-\Psi^{*}\right)}{L}\right]^{\frac{1}{2}} . \tag{3.5}
\end{equation*}
$$

Proof Since $f \in \mathcal{C}_{L}^{1,1}(X)$, it follows from (1.6), (2.4), (3.1) and (3.3) that for any $k=1, \ldots, N$, we have

$$
\begin{align*}
f\left(x_{k+1}\right) & \leq f\left(x_{k}\right)+\left\langle\nabla f\left(x_{k}\right), x_{k+1}-x_{k}\right\rangle+\frac{L}{2}\left\|x_{k+1}-x_{k}\right\|^{2} \\
& =f\left(x_{k}\right)-\gamma_{k}\left\langle\nabla f\left(x_{k}\right), g_{X, k}\right\rangle+\frac{L}{2} \gamma_{k}^{2}\left\|g_{X, k}\right\|^{2} . \tag{3.6}
\end{align*}
$$

Then, by Lemma 1 with $x=x_{k}, \gamma=\gamma_{k}$ and $g=\nabla f\left(x_{k}\right)$, we obtain

$$
f\left(x_{k+1}\right) \leq f\left(x_{k}\right)-\left[\alpha \gamma_{k}\left\|g_{X, k}\right\|^{2}+h\left(x_{k+1}\right)-h\left(x_{k}\right)\right]+\frac{L}{2} \gamma_{k}^{2}\left\|g_{X, k}\right\|^{2},
$$

which implies

$$
\begin{equation*}
\Psi\left(x_{k+1}\right) \leq \Psi\left(x_{k}\right)-\left(\alpha \gamma_{k}-\frac{L}{2} \gamma_{k}^{2}\right)\left\|g_{X, k}\right\|^{2} . \tag{3.7}
\end{equation*}
$$

Summing up the above inequalities for $k=1, \ldots, N$, by (3.2) and $\gamma_{k} \leq 2 \alpha / L$, we have

$$
\begin{align*}
\left\|g_{X, R}\right\|^{2} \sum_{k=1}^{N}\left(\alpha \gamma_{k}-\frac{L}{2} \gamma_{k}^{2}\right) & \leq \sum_{k=1}^{N}\left(\alpha \gamma_{k}-\frac{L}{2} \gamma_{k}^{2}\right)\left\|g_{X, k}\right\|^{2} \\
& \leq \Psi\left(x_{1}\right)-\Psi\left(x_{k+1}\right) \leq \Psi\left(x_{1}\right)-\Psi^{*} . \tag{3.8}
\end{align*}
$$

By our assumption, $\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2} / 2\right)>0$. Hence, dividing both sides of the above inequality by $\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2} / 2\right)$, we obtain (3.4).

The following corollary shows a specialized complexity result for the PG algorithm with one proper constant stepsize policy.

Corollary 1 Suppose that in the PG algorithm the stepsizes $\gamma_{k}=\alpha / L$ for all $k=$ $1, \ldots, N$. Then, we have

$$
\begin{equation*}
\left\|g_{X, R}\right\|^{2} \leq \frac{2 L^{2} D_{\Psi}^{2}}{\alpha^{2} N} \tag{3.9}
\end{equation*}
$$

Proof With the constant stepsizes $\gamma_{k}=\alpha / L$ for all $k=1, \ldots, N$, we have

$$
\begin{equation*}
\frac{L D_{\Psi}^{2}}{\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2} / 2\right)}=\frac{2 L^{2} D_{\Psi}^{2}}{N \alpha^{2}} \tag{3.10}
\end{equation*}
$$

which together with (3.4), clearly imply (3.9).

## 4 Stochastic first-order methods

In this section, we consider problem (1.1) with $f \in \mathcal{C}_{L}^{1,1}(X)$, but its exact gradient is not available. We assume that only noisy first-order information of $f$ is available via subsequent calls to the stochastic first-order oracle $\mathcal{S F} \mathcal{O}$. In particular, given the $k$-th iteration $x_{k} \in X$ of our algorithm, the $\mathcal{S F O}$ will output the stochastic gradient $G\left(x_{k}, \xi_{k}\right)$, where $\xi_{k}$ is a random vector whose distribution is supported on $\Xi_{k} \subseteq \mathbb{R}^{d}$. We assume the stochastic gradient $G\left(x_{k}, \xi_{k}\right)$ satisfies Assumption A1.

This section also consists of two subsections. In Sect. 4.1, we present a stochastic variant of the PG algorithm in Sect. 3 incorporated with a randomized stopping criterion, called the RSPG algorithm. Then, in Sect. 4.2, we describe a two phase RSPG algorithm, called the 2-RSPG algorithm, which can significantly reduce the largedeviations resulted from the RSPG algorithm. We assume throughout this section that the norm $\|\cdot\|$ is associated with the inner product $\langle\cdot, \cdot\rangle$.

### 4.1 A randomized stochastic projected gradient method

Convexity of the objective function often plays an important role on establishing the convergence results for the current SA algorithms [14, 15,21,22,26]. In this subsection, we give an SA-type algorithm which does not require the convexity of the objective function. Moreover, this weaker requirement enables the algorithm to deal with the case in which the random noises $\left\{\xi_{k}\right\}, k \geq 1$ could depend on the iterates $\left\{x_{k}\right\}$.
A randomized stochastic projected gradient ( $R S P G$ ) algorithm
Input: Given initial point $x_{1} \in X$, iteration limit $N$, the stepsizes $\left\{\gamma_{k}\right\}$ with $\gamma_{k}>0, k \geq 1$, the batch sizes $\left\{m_{k}\right\}$ with $m_{k}>0, k \geq 1$, and the probability mass function $P_{R}$ supported on $\{1, \ldots, N\}$.
Step 0 . Let $R$ be a random variable with probability mass function $P_{R}$.
Step $k=1, \ldots, R-1$. Call the $\mathcal{S F} \mathcal{F} m_{k}$ times to obtain $G\left(x_{k}, \xi_{k, i}\right)$, $i=1, \ldots, m_{k}$, set

$$
\begin{equation*}
G_{k}=\frac{1}{m_{k}} \sum_{i=1}^{m_{k}} G\left(x_{k}, \xi_{k, i}\right) \tag{4.1}
\end{equation*}
$$

and compute

$$
\begin{equation*}
x_{k+1}=\arg \min _{u \in X}\left\{\left\langle G_{k}, u\right\rangle+\frac{1}{\gamma_{k}} V\left(u, x_{k}\right)+h(u)\right\} . \tag{4.2}
\end{equation*}
$$

Output: $x_{R}$.
Unlike many SA algorithms, in the RSPG algorithm we use a randomized iteration count to terminate the algorithm. In the RSPG algorithm, we also need to specify the stepsizes $\left\{\gamma_{k}\right\}$, the batch sizes $\left\{m_{k}\right\}$ and probability mass function $P_{R}$. We will again address these issues after presenting some convergence results of the RSPG algorithm.

Theorem 2 Suppose that the stepsizes $\left\{\gamma_{k}\right\}$ in the RSPG algorithm are chosen such that $0<\gamma_{k} \leq \alpha / L$ with $\gamma_{k}<\alpha / L$ for at least one $k$, and the probability mass function $P_{R}$ are chosen such that for any $k=1, \ldots, N$,

$$
\begin{equation*}
P_{R}(k):=\operatorname{Prob}\{R=k\}=\frac{\alpha \gamma_{k}-L \gamma_{k}^{2}}{\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right)} . \tag{4.3}
\end{equation*}
$$

Then, under Assumption Al,
(a) for any $N \geq 1$, we have

$$
\begin{equation*}
\mathbb{E}\left[\left\|\tilde{g}_{X, R}\right\|^{2}\right] \leq \frac{L D_{\Psi}^{2}+\left(\sigma^{2} / \alpha\right) \sum_{k=1}^{N}\left(\gamma_{k} / m_{k}\right)}{\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right)} \tag{4.4}
\end{equation*}
$$

where the expectation is taken with respect to $R$ and $\xi_{[N]}:=\left(\xi_{1}, \ldots, \xi_{N}\right), D_{\Psi}$ is defined in (3.5), and the stochastic projected gradient

$$
\begin{equation*}
\tilde{g}_{X, k}:=P_{X}\left(x_{k}, G_{k}, \gamma_{k}\right), \tag{4.5}
\end{equation*}
$$

with $P_{X}$ defined in (2.4);
(b) if, in addition, $f$ in problem (1.1) is convex with an optimal solution $x^{*}$, and the stepsizes $\left\{\gamma_{k}\right\}$ are non-decreasing, i.e.,

$$
\begin{equation*}
0 \leq \gamma_{1} \leq \gamma_{2} \leq \cdots \leq \gamma_{N} \leq \frac{\alpha}{L} \tag{4.6}
\end{equation*}
$$

we have

$$
\begin{equation*}
\mathbb{E}\left[\Psi\left(x_{R}\right)-\Psi\left(x^{*}\right)\right] \leq \frac{\left(\alpha-L \gamma_{1}\right) V\left(x^{*}, x_{1}\right)+\left(\sigma^{2} / 2\right) \sum_{k=1}^{N}\left(\gamma_{k}^{2} / m_{k}\right)}{\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right)} \tag{4.7}
\end{equation*}
$$

where the expectation is taken with respect to $R$ and $\xi_{[N]}$. Similarly, if the stepsizes $\left\{\gamma_{k}\right\}$ are non-increasing, i.e.,

$$
\begin{equation*}
\frac{\alpha}{L} \geq \gamma_{1} \geq \gamma_{2} \geq \cdots \geq \gamma_{N} \geq 0 \tag{4.8}
\end{equation*}
$$

we have

$$
\begin{equation*}
\mathbb{E}\left[\Psi\left(x_{R}\right)-\Psi\left(x^{*}\right)\right] \leq \frac{\left(\alpha-L \gamma_{N}\right) \bar{V}\left(x^{*}\right)+\left(\sigma^{2} / 2\right) \sum_{k=1}^{N}\left(\gamma_{k}^{2} / m_{k}\right)}{\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right)} \tag{4.9}
\end{equation*}
$$

where $\bar{V}\left(x^{*}\right):=\max _{u \in X} V\left(x^{*}, u\right)$.

Proof Let $\delta_{k} \equiv G_{k}-\nabla f\left(x_{k}\right), k \geq 1$. Since $f \in \mathcal{C}_{L}^{1,1}(X)$, it follows from (1.6), (2.4), (4.2) and (4.5) that, for any $k=1, \ldots, N$, we have

$$
\begin{align*}
f\left(x_{k+1}\right) & \leq f\left(x_{k}\right)+\left\langle\nabla f\left(x_{k}\right), x_{k+1}-x_{k}\right\rangle+\frac{L}{2}\left\|x_{k+1}-x_{k}\right\|^{2} \\
& =f\left(x_{k}\right)-\gamma_{k}\left\langle\nabla f\left(x_{k}\right), \tilde{g}_{X, k}\right\rangle+\frac{L}{2} \gamma_{k}^{2}\left\|\tilde{g}_{X, k}\right\|^{2} \\
& =f\left(x_{k}\right)-\gamma_{k}\left\langle G_{k}, \tilde{g}_{X, k}\right\rangle+\frac{L}{2} \gamma_{k}^{2}\left\|\tilde{g}_{X, k}\right\|^{2}+\gamma_{k}\left\langle\delta_{k}, \tilde{g}_{X, k}\right\rangle . \tag{4.10}
\end{align*}
$$

So, by Lemma 1 with $x=x_{k}, \gamma=\gamma_{k}$ and $g=G_{k}$, we obtain

$$
\begin{aligned}
f\left(x_{k+1}\right) \leq & f\left(x_{k}\right)-\left[\alpha \gamma_{k}\left\|\tilde{g}_{X, k}\right\|^{2}+h\left(x_{k+1}\right)-h\left(x_{k}\right)\right]+\frac{L}{2} \gamma_{k}^{2}\left\|\tilde{g}_{X, k}\right\|^{2} \\
& +\gamma_{k}\left\langle\delta_{k}, g_{X, k}\right\rangle+\gamma_{k}\left\langle\delta_{k}, \tilde{g}_{X, k}-g_{X, k}\right\rangle,
\end{aligned}
$$

where the projected gradient $g_{X, k}$ is defined in (3.3). Then, from the above inequality, (3.3) and (4.5), we obtain

$$
\begin{aligned}
\Psi\left(x_{k+1}\right) & \leq \Psi\left(x_{k}\right)-\left(\alpha \gamma_{k}-\frac{L}{2} \gamma_{k}^{2}\right)\left\|\tilde{g}_{X, k}\right\|^{2}+\gamma_{k}\left\langle\delta_{k}, g_{X, k}\right\rangle+\gamma_{k}\left\|\delta_{k}\right\|\left\|\tilde{g}_{X, k}-g_{X, k}\right\| \\
& \leq \Psi\left(x_{k}\right)-\left(\alpha \gamma_{k}-\frac{L}{2} \gamma_{k}^{2}\right)\left\|\tilde{g}_{X, k}\right\|^{2}+\gamma_{k}\left\langle\delta_{k}, g_{X, k}\right\rangle+\frac{\gamma_{k}}{\alpha}\left\|\delta_{k}\right\|^{2},
\end{aligned}
$$

where the last inequality follows from Proposition 1 with $x=x_{k}, \gamma=\gamma_{k}, g_{1}=G_{k}$ and $g_{2}=\nabla f\left(x_{k}\right)$. Summing up the above inequalities for $k=1, \ldots, N$ and noticing that $\gamma_{k} \leq \alpha / L$, we obtain

$$
\begin{align*}
\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right)\left\|\tilde{g}_{X, k}\right\|^{2} & \leq \sum_{k=1}^{N}\left(\alpha \gamma_{k}-\frac{L}{2} \gamma_{k}^{2}\right)\left\|\tilde{g}_{X, k}\right\|^{2} \\
& \leq \Psi\left(x_{1}\right)-\Psi\left(x_{k+1}\right)+\sum_{k=1}^{N}\left\{\gamma_{k}\left\langle\delta_{k}, g_{X, k}\right\rangle+\frac{\gamma_{k}}{\alpha}\left\|\delta_{k}\right\|^{2}\right\} \\
& \leq \Psi\left(x_{1}\right)-\Psi^{*}+\sum_{k=1}^{N}\left\{\gamma_{k}\left\langle\delta_{k}, g_{X, k}\right\rangle+\frac{\gamma_{k}}{\alpha}\left\|\delta_{k}\right\|^{2}\right\} . \tag{4.11}
\end{align*}
$$

Notice that the iterate $x_{k}$ is a function of the history $\xi_{[k-1]}$ of the generated random process and hence is random. By part a) of Assumption A1, we have $\mathbb{E}\left[\left\langle\delta_{k}, g_{X, k}\right\rangle \mid \xi_{[k-1]}\right]=0$. In addition, denoting $\delta_{k, i} \equiv G\left(x_{k}, \xi_{k, i}\right)-\nabla f\left(x_{k}\right)$, $i=1, \ldots, m_{k}, k=1, \ldots, N, S_{j}=\sum_{i=1}^{j} \delta_{k, i}, j=1, \ldots, m_{k}$, and $S_{0}=0$, and noting that $\mathbb{E}\left[\left\langle S_{i-1}, \delta_{k, i}\right\rangle \mid S_{i-1}\right]=0$ for all $i=1, \ldots, m_{k}$, we have

$$
\begin{aligned}
\mathbb{E}\left[\left\|S_{m_{k}}\right\|^{2}\right] & =\mathbb{E}\left[\left\|S_{m_{k}-1}\right\|^{2}+2\left\langle S_{m_{k}-1}, \delta_{k, m_{k}}\right\rangle+\left\|\delta_{k, m_{k}}\right\|^{2}\right] \\
& =\mathbb{E}\left[\left\|S_{m_{k}-1}\right\|^{2}\right]+\mathbb{E}\left[\left\|\delta_{k, m_{k}}\right\|^{2}\right]=\cdots=\sum_{i=1}^{m_{k}} \mathbb{E}\left\|\delta_{k, i}\right\|^{2}
\end{aligned}
$$

which, in view of (4.1) and Assumption A1(b), then implies that
$\mathbb{E}\left[\left\|\delta_{k}\right\|^{2}\right]=\mathbb{E}\left[\left\|\frac{1}{m_{k}} \sum_{i=1}^{m_{k}} \delta_{k, i}\right\|^{2}\right]=\frac{1}{m_{k}^{2}} \mathbb{E}\left[\left\|S_{m_{k}}\right\|^{2}\right]=\frac{1}{m_{k}^{2}} \sum_{i=1}^{m_{k}} \mathbb{E}\left[\left\|\delta_{k, i}\right\|^{2}\right] \leq \frac{\sigma^{2}}{m_{k}}$.

With these observations, now taking expectations with respect to $\xi_{[N]}$ on both sides of (4.11), we get

$$
\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right) \mathbb{E}\left\|\tilde{g}_{X, k}\right\|^{2} \leq \Psi\left(x_{1}\right)-\Psi^{*}+\left(\sigma^{2} / \alpha\right) \sum_{k=1}^{N}\left(\gamma_{k} / m_{k}\right)
$$

Then, since $\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right)>0$ by our assumption, dividing both sides of the above inequality by $\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right)$ and noticing that

$$
\mathbb{E}\left[\left\|\tilde{g}_{X, R}\right\|^{2}\right]=\frac{\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right) \mathbb{E}\left\|\tilde{g}_{X, k}\right\|^{2}}{\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right)}
$$

we have (4.4) holds.
We now show part (b) of the theorem. By Lemma 3 with $x=x_{k}, \gamma=\gamma_{k}, g=G_{k}$ and $u=x^{*}$, we have

$$
\begin{aligned}
& \left\langle G_{k}, x_{k+1}\right\rangle+h\left(x_{k+1}\right)+\frac{1}{\gamma_{k}} V\left(x_{k+1}, x_{k}\right) \leq\left\langle G_{k}, x^{*}\right\rangle+h\left(x^{*}\right) \\
& \quad+\frac{1}{\gamma_{k}}\left[V\left(x^{*}, x_{k}\right)-V\left(x^{*}, x_{k+1}\right)\right]
\end{aligned}
$$

which together with (1.6) and definition of $\delta_{k}$ give

$$
\begin{aligned}
& f\left(x_{k+1}\right)+\left\langle\nabla f\left(x_{k}\right)+\delta_{k}, x_{k+1}\right\rangle+h\left(x_{k+1}\right)+\frac{1}{\gamma_{k}} V\left(x_{k+1}, x_{k}\right) \\
& \quad \leq f\left(x_{k}\right)+\left\langle\nabla f\left(x_{k}\right), x_{k+1}-x_{k}\right\rangle+\frac{L}{2}\left\|x_{k+1}-x_{k}\right\|^{2} \\
& \quad+\left\langle\nabla f\left(x_{k}\right)+\delta_{k}, x^{*}\right\rangle+h\left(x^{*}\right)+\frac{1}{\gamma_{k}}\left[V\left(x^{*}, x_{k}\right)-V\left(x^{*}, x_{k+1}\right)\right] .
\end{aligned}
$$

Simplifying the above inequality, we have

$$
\begin{aligned}
\Psi\left(x_{k+1}\right) \leq & f\left(x_{k}\right)+\left\langle\nabla f\left(x_{k}\right), x^{*}-x_{k}\right\rangle+h\left(x^{*}\right)+\left\langle\delta_{k}, x^{*}-x_{k+1}\right\rangle+\frac{L}{2}\left\|x_{k+1}-x_{k}\right\|^{2} \\
& -\frac{1}{\gamma_{k}} V\left(x_{k+1}, x_{k}\right)+\frac{1}{\gamma_{k}}\left[V\left(x^{*}, x_{k}\right)-V\left(x^{*}, x_{k+1}\right)\right] .
\end{aligned}
$$

Then, it follows from the convexity of $f,(2.1)$ and (2.2) that

$$
\begin{aligned}
\Psi\left(x_{k+1}\right) \leq & f\left(x^{*}\right)+h\left(x^{*}\right)+\left\langle\delta_{k}, x^{*}-x_{k+1}\right\rangle+\left(\frac{L}{2}-\frac{\alpha}{2 \gamma_{k}}\right)\left\|x_{k+1}-x_{k}\right\|^{2} \\
& +\frac{1}{\gamma_{k}}\left[V\left(x^{*}, x_{k}\right)-V\left(x^{*}, x_{k+1}\right)\right] \\
= & \Psi\left(x^{*}\right)+\left\langle\delta_{k}, x^{*}-x_{k}\right\rangle+\left\langle\delta_{k}, x_{k}-x_{k+1}\right\rangle+\frac{L \gamma_{k}-\alpha}{2 \gamma_{k}}\left\|x_{k+1}-x_{k}\right\|^{2} \\
& +\frac{1}{\gamma_{k}}\left[V\left(x^{*}, x_{k}\right)-V\left(x^{*}, x_{k+1}\right)\right] \\
\leq & \Psi\left(x^{*}\right)+\left\langle\delta_{k}, x^{*}-x_{k}\right\rangle+\left\|\delta_{k}\right\|\left\|x_{k}-x_{k+1}\right\|-\frac{\alpha-L \gamma_{k}}{2 \gamma_{k}}\left\|x_{k+1}-x_{k}\right\|^{2} \\
& +\frac{1}{\gamma_{k}}\left[V\left(x^{*}, x_{k}\right)-V\left(x^{*}, x_{k+1}\right)\right] \\
\leq & \Psi\left(x^{*}\right)+\left\langle\delta_{k}, x^{*}-x_{k}\right\rangle+\frac{\gamma_{k}}{2\left(\alpha-L \gamma_{k}\right)}\left\|\delta_{k}\right\|^{2}+\frac{1}{\gamma_{k}}\left[V\left(x^{*}, x_{k}\right)-V\left(x^{*}, x_{k+1}\right)\right]
\end{aligned}
$$

where the last inequality follows from the fact that $a x-b x^{2} / 2 \leq a^{2} /(2 b)$. Noticing $\gamma_{k} \leq \alpha / L$, multiplying both sides of the above inequality by $\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right)$ and summing them up for $k=1, \ldots, N$, we obtain

$$
\begin{align*}
\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right)\left[\Psi\left(x_{k+1}\right)-\Psi\left(x^{*}\right)\right] \leq & \sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right)\left\langle\delta_{k}, x^{*}-x_{k}\right\rangle \\
& +\sum_{k=1}^{N} \frac{\gamma_{k}^{2}}{2}\left\|\delta_{k}\right\|^{2}+\sum_{k=1}^{N}\left(\alpha-L \gamma_{k}\right) \\
& \times\left[V\left(x^{*}, x_{k}\right)-V\left(x^{*}, x_{k+1}\right)\right] \tag{4.13}
\end{align*}
$$

Now, if the increasing stepsize condition (4.6) is satisfied, we have from $V\left(x^{*}, x_{N+1}\right)$ $\geq 0$ that

$$
\begin{aligned}
& \sum_{k=1}^{N}\left(\alpha-L \gamma_{k}\right)\left[V\left(x^{*}, x_{k}\right)-V\left(x^{*}, x_{k+1}\right)\right] \\
& \quad=\left(\alpha-L \gamma_{1}\right) V\left(x^{*}, x_{1}\right)+\sum_{k=2}^{N}\left(\alpha-L \gamma_{k}\right) V\left(x^{*}, x_{k}\right)-\sum_{k=1}^{N}\left(\alpha-L \gamma_{k}\right) V\left(x^{*}, x_{k+1}\right)
\end{aligned}
$$

$$
\begin{aligned}
& \leq\left(\alpha-L \gamma_{1}\right) V\left(x^{*}, x_{1}\right)+\sum_{k=2}^{N}\left(\alpha-L \gamma_{k-1}\right) V\left(x^{*}, x_{k}\right)-\sum_{k=1}^{N}\left(\alpha-L \gamma_{k}\right) V\left(x^{*}, x_{k+1}\right) \\
& =\left(\alpha-L \gamma_{1}\right) V\left(x^{*}, x_{1}\right)-\left(\alpha-L \gamma_{N}\right) V\left(x^{*}, x_{N+1}\right) \\
& \leq\left(\alpha-L \gamma_{1}\right) V\left(x^{*}, x_{1}\right)
\end{aligned}
$$

Taking expectation on both sides of (4.13) with respect to $\xi_{[N]}$, again using the observations that $\mathbb{E}\left[\left\|\delta_{k}^{2}\right\|\right] \leq \sigma^{2} / m_{k}$ and $\mathbb{E}\left[\left\langle\delta_{k}, g_{X, k}\right\rangle \mid \xi_{[k-1]}\right]=0$, then it follows from the above inequality that
$\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right) \mathbb{E}_{\xi_{[N]}}\left[\Psi\left(x_{k+1}\right)-\Psi\left(x^{*}\right)\right] \leq\left(\alpha-L \gamma_{1}\right) V\left(x^{*}, x_{1}\right)+\frac{\sigma^{2}}{2} \sum_{k=1}^{N}\left(\gamma_{k}^{2} / m_{k}\right)$.
Finally, (4.7) follows from the above inequality and the arguments similar to the proof in part (a). Now, if the decreasing stepsize condition (4.8) is satisfied, we have from the definition $\bar{V}\left(x^{*}\right):=\max _{u \in X} V\left(x^{*}, u\right) \geq 0$ and $V\left(x^{*}, x_{N+1}\right) \geq 0$ that

$$
\begin{aligned}
& \sum_{k=1}^{N}\left(\alpha-L \gamma_{k}\right)\left[V\left(x^{*}, x_{k}\right)-V\left(x^{*}, x_{k+1}\right)\right] \\
& \quad=\left(\alpha-L \gamma_{1}\right) V\left(x^{*}, x_{1}\right)+L \sum_{k=1}^{N-1}\left(\gamma_{k}-\gamma_{k+1}\right) V\left(x^{*}, x_{k+1}\right)-\left(\alpha-L \gamma_{N}\right) V\left(x^{*}, x_{N+1}\right) \\
& \quad \leq\left(\alpha-L \gamma_{1}\right) \bar{V}\left(x^{*}\right)+L \sum_{k=1}^{N-1}\left(\gamma_{k}-\gamma_{k+1}\right) \bar{V}\left(x^{*}\right)-\left(\alpha-L \gamma_{N}\right) V\left(x^{*}, x_{N+1}\right) \\
& \quad \leq\left(\alpha-L \gamma_{N}\right) \bar{V}\left(x^{*}\right)
\end{aligned}
$$

which together with (4.13) and similar arguments used above would give (4.9).
A few remarks about Theorem 2 are in place. Firstly, if $f$ is convex and the batch sizes $m_{k}=1$, then by properly choosing the stepsizes $\left\{\gamma_{k}\right\}$ (e.g., $\gamma_{k}=\mathcal{O}(1 / \sqrt{k})$ for $k$ large), we can still guarantee a nearly optimal rate of convergence for the RSPG algorithm (see $(4.7)$ or (4.9), and $[21,26]$ ). However, if $f$ is possibly nonconvex and $m_{k}=1$, then the right hand side of (4.4) is bounded from below by

$$
\frac{L D_{\Psi}^{2}+\left(\sigma^{2} / \alpha\right) \sum_{k=1}^{N} \gamma_{k}}{\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right)} \geq \frac{\sigma^{2}}{\alpha^{2}}
$$

which can not guarantee the convergence of the RSPG algorithm, no matter how the stepsizes $\left\{\gamma_{k}\right\}$ are specified. This is exactly the reason why we consider taking multiple samples $G\left(x_{k}, \xi_{k, i}\right), i=1, \ldots, m_{k}$, for some $m_{k}>1$ at each iteration of the RSPG method.

Secondly, we need to estimate $L$ to ensure the condition on the stepsize $\gamma_{k}$. In Sect. 6, we describe how to do it by taking a small number of samples before running
the algorithm. However, we do not need a very accurate estimation for $L$ (see the discussion after Corollary 2.2 in [16] for more details in the similar case).

Thirdly, from (4.11) in the proof of Theorem 2, we see that the stepsize policies can be further relaxed to get a similar result as (4.4). More specifically, we can have the following corollary.

Corollary 2 Suppose that the stepsizes $\left\{\gamma_{k}\right\}$ in the RSPG algorithm are chosen such that $0<\gamma_{k} \leq 2 \alpha / L$ with $\gamma_{k}<2 \alpha / L$ for at least one $k$, and the probability mass function $P_{R}$ are chosen such that for any $k=1, \ldots, N$,

$$
\begin{equation*}
P_{R}(k):=\operatorname{Prob}\{R=k\}=\frac{\alpha \gamma_{k}-L \gamma_{k}^{2} / 2}{\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2} / 2\right)} . \tag{4.14}
\end{equation*}
$$

Then, under Assumption Al, we have

$$
\begin{equation*}
\mathbb{E}\left[\left\|\tilde{g}_{X, R}\right\|^{2}\right] \leq \frac{L D_{\Psi}^{2}+\left(\sigma^{2} / \alpha\right) \sum_{k=1}^{N}\left(\gamma_{k} / m_{k}\right)}{\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2} / 2\right)} \tag{4.15}
\end{equation*}
$$

where the expectation is taken with respect to $R$ and $\xi_{[N]}:=\left(\xi_{1}, \ldots, \xi_{N}\right)$.
Based on the Theorem 2, we can establish the following complexity results of the RSPG algorithm with proper selection of stepsizes $\left\{\gamma_{k}\right\}$ and batch sizes $\left\{m_{k}\right\}$ at each iteration.

Corollary 3 Suppose that in the RSPG algorithm the stepsizes $\gamma_{k}=\alpha /(2 L)$ for all $k=1, \ldots, N$, and the probability mass function $P_{R}$ are chosen as (4.3). Also assume that the batch sizes $m_{k}=m, k=1, \ldots, N$, for some $m \geq 1$. Then under Assumption Al, we have

$$
\begin{equation*}
\mathbb{E}\left[\left\|g_{X, R}\right\|^{2}\right] \leq \frac{8 L^{2} D_{\Psi}^{2}}{\alpha^{2} N}+\frac{6 \sigma^{2}}{\alpha^{2} m} \text { and } \mathbb{E}\left[\left\|\tilde{g}_{X, R}\right\|^{2}\right] \leq \frac{4 L^{2} D_{\Psi}^{2}}{\alpha^{2} N}+\frac{2 \sigma^{2}}{\alpha^{2} m} \tag{4.16}
\end{equation*}
$$

where $g_{X, R}$ and $\tilde{g}_{X, R}$ are defined in (3.3) and (4.5), respectively. If, in addition, $f$ in the problem (1.1) is convex with an optimal solution $x^{*}$, then

$$
\begin{equation*}
\mathbb{E}\left[\Psi\left(x_{R}\right)-\Psi\left(x^{*}\right)\right] \leq \frac{2 L V\left(x^{*}, x_{1}\right)}{N \alpha}+\frac{\sigma^{2}}{2 L m} \tag{4.17}
\end{equation*}
$$

Proof By (4.4), we have

$$
\mathbb{E}\left[\left\|\tilde{g}_{X, R}\right\|^{2}\right] \leq \frac{L D_{\Psi}^{2}+\frac{\sigma^{2}}{m \alpha} \sum_{k=1}^{N} \gamma_{k}}{\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right)},
$$

which together with $\gamma_{k}=\alpha /(2 L)$ for all $k=1, \ldots, N$ imply that

$$
\mathbb{E}\left[\left\|\tilde{g}_{X, R}\right\|^{2}\right]=\frac{L D_{\Psi}^{2}+\frac{\sigma^{2} N}{2 m L}}{\frac{N \alpha^{2}}{4 L}}=\frac{4 L^{2} D_{\Psi}^{2}}{N \alpha^{2}}+\frac{2 \sigma^{2}}{m \alpha^{2}}
$$

Then, by Proposition 1 with $x=x_{R}, \gamma=\gamma_{R}, g_{1}=\nabla f\left(x_{R}\right), g_{2}=G_{k}$, we have from the above inequality and 4.12 that

$$
\begin{aligned}
\mathbb{E}\left[\left\|g_{X, R}\right\|^{2}\right] & \leq 2 \mathbb{E}\left[\left\|\tilde{g}_{X, R}\right\|^{2}\right]+2 \mathbb{E}\left[\left\|g_{X, R}-\tilde{g}_{X, R}\right\|^{2}\right] \\
& \leq 2\left(\frac{4 L^{2} D_{\Psi}^{2}}{N \alpha^{2}}+\frac{2 \sigma^{2}}{\alpha^{2} m}\right)+\frac{2}{\alpha^{2}} \mathbb{E}\left[\left\|G_{k}-\nabla f\left(x_{R}\right)\right\|^{2}\right] \\
& \leq \frac{8 L^{2} D_{\Psi}^{2}}{N \alpha^{2}}+\frac{6 \sigma^{2}}{\alpha^{2} m} .
\end{aligned}
$$

Moreover, since $\gamma_{k}=\alpha /(2 L)$ for all $k=1, \ldots, N$, the stepsize conditions (4.6) are satisfied. Hence, if the problem is convex, (4.17) can be derived in a similar way as (4.7).

Note that all the bounds in the above corollary depend on $m$. Indeed, if $m$ is set to some fixed positive integer constant, then the second terms in the above results will always majorize the first terms when $N$ is sufficiently large. Hence, the appropriate choice of $m$ should be balanced with the number of iterations $N$, which would eventually depend on the total computational budget given by the user. The following corollary shows an appropriate choice of $m$ depending on the total number of calls to the $\mathcal{S F O}$.

Corollary 4 Suppose that all the conditions in Corollary 3 are satisfied. Given a fixed total number of calls $\bar{N}$ to the $\mathcal{S F O}$, if the number of calls to the $\mathcal{S F O}$ (number of samples) at each iteration of the RSPG algorithm is

$$
\begin{equation*}
m=\left\lceil\min \left\{\max \left\{1, \frac{\sigma \sqrt{6 \bar{N}}}{4 L \tilde{D}}\right\}, \bar{N}\right\}\right\rceil, \tag{4.18}
\end{equation*}
$$

for some $\tilde{D}>0$, then we have $\left(\alpha^{2} / L\right) \mathbb{E}\left[\left\|g_{X, R}\right\|^{2}\right] \leq \mathcal{B}_{\bar{N}}$, where

$$
\begin{equation*}
\mathcal{B}_{\bar{N}}:=\frac{16 L D_{\Psi}^{2}}{\bar{N}}+\frac{4 \sqrt{6} \sigma}{\sqrt{\bar{N}}}\left(\frac{D_{\Psi}^{2}}{\tilde{D}}+\tilde{D} \max \left\{1, \frac{\sqrt{6} \sigma}{4 L \tilde{D} \sqrt{\bar{N}}}\right\}\right) . \tag{4.19}
\end{equation*}
$$

If, in addition, $f$ in problem (1.1) is convex, then $\mathbb{E}\left[\Psi\left(x_{R}\right)-\Psi\left(x^{*}\right)\right] \leq \mathcal{C}_{\bar{N}}$, where $x^{*}$ is an optimal solution and

$$
\begin{equation*}
\mathcal{C}_{\bar{N}}:=\frac{4 L V\left(x^{*}, x_{1}\right)}{\alpha \bar{N}}+\frac{\sqrt{6} \sigma}{\alpha \sqrt{\bar{N}}}\left(\frac{V\left(x^{*}, x_{1}\right)}{\tilde{D}}+\frac{\alpha \tilde{D}}{3} \max \left\{1, \frac{\sqrt{6} \sigma}{4 L \tilde{D} \sqrt{\bar{N}}}\right\}\right) . \tag{4.20}
\end{equation*}
$$

Proof Given the total number of calls to the stochastic first-order oracle $\bar{N}$ and the number $m$ of calls to the $\mathcal{S F O}$ at each iteration, the RSPG algorithm can perform at most $N=\lfloor\bar{N} / m\rfloor$ iterations. Obviously, $N \geq \bar{N} /(2 m)$. With this observation and (4.16), we have

$$
\begin{align*}
\mathbb{E}\left[\left\|g_{X, R}\right\|^{2}\right] & \leq \frac{16 m L^{2} D_{\Psi}^{2}}{\alpha^{2} \bar{N}}+\frac{6 \sigma^{2}}{\alpha^{2} m} \\
& \leq \frac{16 L^{2} D_{\Psi}^{2}}{\alpha^{2} \bar{N}}\left(1+\frac{\sigma \sqrt{6 \bar{N}}}{4 L \tilde{D}}\right)+\max \left\{\frac{4 \sqrt{6} L \tilde{D} \sigma}{\alpha^{2} \sqrt{\bar{N}}}, \frac{6 \sigma^{2}}{\alpha^{2} \bar{N}}\right\} \\
& =\frac{16 L^{2} D_{\Psi}^{2}}{\alpha^{2} \bar{N}}+\frac{4 \sqrt{6} L \sigma}{\alpha^{2} \sqrt{\bar{N}}}\left(\frac{D_{\Psi}^{2}}{\tilde{D}}+\tilde{D} \max \left\{1, \frac{\sqrt{6} \sigma}{4 L \tilde{D} \sqrt{\bar{N}}}\right\}\right) \tag{4.21}
\end{align*}
$$

which gives (4.19). The bound (4.20) can be obtained in a similar way.
We now would like add a few remarks about the above results in Corollary 4. Firstly, although we use the constant value for $m_{k}=m$ at each iteration, one can also choose it adaptively during the execution of the RSPG algorithm while monitoring the convergence. For example, in practice $m_{k}$ could adaptively depend on $\sigma_{k}^{2}:=$ $\mathbb{E}\left[\left\|G\left(x_{k}, \xi_{k}\right)-\nabla f\left(x_{k}\right)\right\|^{2}\right]$. Another example is to choose growing batch sizes where one uses a smaller number of samples in the beginning of the algorithm. In particular, by setting

$$
m_{k}=\left\lceil\min \left\{\frac{\sigma\left(k^{2} \bar{N}\right)^{\frac{1}{4}}}{L \tilde{D}}, \bar{N}\right\}\right\rceil,
$$

we can easily see that the RSPG algorithm still achieves the same rates of convergence as those obtained by using constant bath sizes in Corollary 4. Secondly, we need to specify the parameter $\tilde{D}$ in (4.18). It can be seen from (4.19) and (4.20) that when $\bar{N}$ is relatively large such that

$$
\begin{equation*}
\max \{1, \sqrt{6} \sigma /(4 L \tilde{D} \sqrt{\bar{N}})\}=1, \quad \text { i.e., } \quad \bar{N} \geq 3 \sigma^{2} /\left(8 L^{2} \tilde{D}^{2}\right) \tag{4.22}
\end{equation*}
$$

an optimal choice of $\tilde{D}$ would be $D_{\Psi}$ and $\sqrt{3 V\left(x^{*}, x_{1}\right) / \alpha}$ for solving nonconvex and convex SP problems, respectively. With this selection of $\tilde{D}$, the bounds in (4.19) and (4.20), respectively, reduce to

$$
\begin{equation*}
\frac{\alpha^{2}}{L} \mathbb{E}\left[\left\|g_{X, R}\right\|^{2}\right] \leq \frac{16 L D_{\Psi}^{2}}{\bar{N}}+\frac{8 \sqrt{6} D_{\Psi} \sigma}{\sqrt{\bar{N}}} \tag{4.23}
\end{equation*}
$$

and

$$
\begin{equation*}
\mathbb{E}\left[\Psi\left(x^{*}\right)-\Psi\left(x_{1}\right)\right] \leq \frac{4 L V\left(x^{*}, x_{1}\right)}{\alpha \bar{N}}+\frac{2 \sqrt{2 V\left(x^{*}, x_{1}\right)} \sigma}{\sqrt{\alpha \bar{N}}} \tag{4.24}
\end{equation*}
$$

Thirdly, the stepsize policy in Corollary 3 and the probability mass function (4.3) together with the number of samples (4.18) at each iteration of the RSPG algorithm provide a unified strategy for solving both convex and nonconvex SP problems. In
particular, the RSPG algorithm exhibits a nearly optimal rate of convergence for solving smooth convex SP problems, since the second term in (4.24) is unimprovable (see e.g., [27]), while the first term in (4.24) can be considerably improved [21].

### 4.2 A two-phase randomized stochastic projected gradient method

In the previous subsection, we present the expected complexity results over many runs of the RSPG algorithm. Indeed, we are also interested in the performance of a single run of RSPG. In particular, we want to establish the complexity results for finding an $(\epsilon, \Lambda)$-solution of the problem (1.1), i.e., a point $x \in X$ satisfying $\operatorname{Prob}\left\{\left\|g_{X}(x)\right\|^{2} \leq\right.$ $\epsilon\} \geq 1-\Lambda$, for some $\epsilon>0$ and $\Lambda \in(0,1)$. Noticing that by the Markov's inequality and (4.19), we can directly have

$$
\begin{equation*}
\operatorname{Prob}\left\{\left\|g_{X, R}\right\|^{2} \geq \frac{\lambda L \mathcal{B}_{\bar{N}}}{\alpha^{2}}\right\} \leq \frac{1}{\lambda}, \quad \text { for any } \quad \lambda>0 \tag{4.25}
\end{equation*}
$$

This implies that the total number of calls to the $\mathcal{S F \mathcal { O }}$ performed by the RSPG algorithm for finding an $(\epsilon, \Lambda)$-solution, after disregarding a few constant factors, can be bounded by

$$
\begin{equation*}
\mathcal{O}\left\{\frac{1}{\Lambda \epsilon}+\frac{\sigma^{2}}{\Lambda^{2} \epsilon^{2}}\right\} \tag{4.26}
\end{equation*}
$$

In this subsection, we present a approach to improve the dependence of the above bound on $\Lambda$. More specifically, we propose a variant of the RSPG algorithm which has two phases: an optimization phase and a post-optimization phase. The optimization phase consists of independent single runs of the RSPG algorithm to generate a list of candidate solutions, and in the post-optimization phase, we choose a solution $x^{*}$ from these candidate solutions generated by the optimization phase. For the sake of simplicity, we assume throughout this subsection that the norm $\|\cdot\|$ in $\mathbb{R}^{n}$ is the standard Euclidean norm.

## A two phase RSPG (2-RSPG) algorithm

Input: Given initial point $x_{1} \in X$, number of runs $S$, total $\bar{N}$ of calls to the $\mathcal{S F} \mathcal{O}$ in each run of the RSPG algorithm, and sample size $T$ in the post-optimization phase.

## Optimization phase:

For $s=1, \ldots, S$
Call the RSPG algorithm with initial point $x_{1}$, iteration limit $N=\lfloor\bar{N} / m\rfloor$ with $m$ given by (4.18), stepsizes $\gamma_{k}=\alpha /(2 L)$ for $k=1, \ldots, N$, batch sizes $m_{k}=m$, and probability mass function $P_{R}$ in (4.3).
Let $\bar{x}_{s}=x_{R_{s}}, s=1, \ldots, S$, be the outputs of this phase.

## Post-optimization phase:

Choose a solution $\bar{x}^{*}$ from the candidate list $\left\{\bar{x}_{1}, \ldots, \bar{x}_{S}\right\}$ such that

$$
\begin{equation*}
\left\|\bar{g}_{X}\left(\bar{x}^{*}\right)\right\|=\min _{s=1, \ldots, S}\left\|\bar{g}_{X}\left(\bar{x}_{S}\right)\right\|, \quad \bar{g}_{X}\left(\bar{x}_{s}\right):=P_{X}\left(\bar{x}_{s}, \bar{G}_{T}\left(\bar{x}_{s}\right), \gamma_{R_{s}}\right), \tag{4.27}
\end{equation*}
$$

where $\bar{G}_{T}(x)=\frac{1}{T} \sum_{k=1}^{T} G\left(x, \xi_{k}\right)$ and $P_{X}(x, g, \gamma)$ is defined in (2.4).
Output: $\bar{x}^{*}$.

In the 2-RSPG algorithm, the total number of calls of $\mathcal{S F O}$ in the optimization phase and post-optimization phase is bounded by $S \times \bar{N}$ and $S \times T$, respectively. In the next theorem, we provide certain bounds of $S, N$ and $T$ for finding an $(\epsilon, \Lambda)$-solution of problem (1.1).

We need the following well-known large deviation theorem of vector-valued martingales to derive the large deviation results of the 2-RSPG algorithm (see [18] for a general result using possibly non-Euclidean norm).

Lemma 4 Assume that we are given a Polish space with Borel probability measure $\mu$ and a sequence of $\mathcal{F}_{0}=\{\emptyset, \Omega\} \subseteq \mathcal{F}_{1} \subseteq \mathcal{F}_{2} \subseteq \ldots$ of $\sigma$-sub-algebras of Borel $\sigma$-algebra of $\Omega$. Let $\zeta_{i} \in \mathbb{R}^{n}, i=1, \ldots, \infty$, be a martingale-difference sequence of Borel functions on $\Omega$ such that $\zeta_{i}$ is $\mathcal{F}_{i}$ measurable and $\mathbb{E}\left[\zeta_{i} \mid i-1\right]=0$, where $\mathbb{E}[\cdot \mid i], i=1,2, \ldots$, denotes the conditional expectation w.r.t. $\mathcal{F}_{i}$ and $\mathbb{E} \equiv \mathbb{E}[\cdot \mid 0]$ is the expectation w.r.t. $\mu$.
(a) If $\mathbb{E}\left[\left\|\zeta_{i}\right\|^{2}\right] \leq \sigma_{i}^{2}$ for any $i \geq 1$, then $\mathbb{E}\left[\left\|\sum_{i=1}^{N} \zeta_{i}\right\|^{2}\right] \leq \sum_{i=1}^{N} \sigma_{i}^{2}$. As a consequence, we have

$$
\forall N \geq 1, \lambda \geq 0: \operatorname{Prob}\left\{\left\|\sum_{i=1}^{N} \zeta_{i}\right\|^{2} \geq \lambda \sum_{i=1}^{N} \sigma_{i}^{2}\right\} \leq \frac{1}{\lambda}
$$

(b) If $\mathbb{E}\left[\exp \left(\left\|\zeta_{i}\right\|^{2} / \sigma_{i}^{2}\right) \mid i-1\right] \leq \exp (1)$ almost surely for any $i \geq 1$, then

$$
\forall N \geq 1, \lambda \geq 0: \operatorname{Prob}\left\{\left\|\sum_{i=1}^{N} \zeta_{i}\right\| \geq \sqrt{2}(1+\lambda) \sqrt{\sum_{i=1}^{N} \sigma_{i}^{2}}\right\} \leq \exp \left(-\lambda^{2} / 3\right)
$$

We are now ready to state the main convergence properties for the 2-RSPG algorithm.

Theorem 3 Under Assumption A1, the following statements hold for the 2-RSPG algorithm applied to problem (1.1).
(a) Let $\mathcal{B}_{\bar{N}}$ be defined in (4.19). Then, for all $\lambda>0$

$$
\begin{equation*}
\operatorname{Prob}\left\{\left\|g_{X}\left(\bar{x}^{*}\right)\right\|^{2} \geq \frac{2}{\alpha^{2}}\left(4 L \mathcal{B}_{\bar{N}}+\frac{3 \lambda \sigma^{2}}{T}\right)\right\} \leq \frac{S}{\lambda}+2^{-S} ; \tag{4.28}
\end{equation*}
$$

(b) Let $\epsilon>0$ and $\Lambda \in(0,1)$ be given. If the parameters $(S, \bar{N}, T)$ are set to

$$
\begin{align*}
& S(\Lambda):=\left\lceil\log _{2}(2 / \Lambda)\right\rceil,  \tag{4.29}\\
& \bar{N}(\epsilon):=\left\lceil\max \left\{\frac{512 L^{2} D_{\Psi}^{2}}{\alpha^{2} \epsilon},\left[\left(\tilde{D}+\frac{D_{\Psi}^{2}}{\tilde{D}}\right) \frac{128 \sqrt{6} L \sigma}{\alpha^{2} \epsilon}\right]^{2}, \frac{3 \sigma^{2}}{8 L^{2} \tilde{D}^{2}}\right\}\right],  \tag{4.30}\\
& T(\epsilon, \Lambda):=\left\lceil\frac{24 S(\Lambda) \sigma^{2}}{\alpha^{2} \Lambda \epsilon}\right\rceil, \tag{4.31}
\end{align*}
$$

then the 2-RSPG algorithm computes an $(\epsilon, \Lambda)$-solution of the problem (1.1) after taking at most

$$
\begin{equation*}
S(\Lambda)[\bar{N}(\epsilon)+T(\epsilon, \Lambda)] \tag{4.32}
\end{equation*}
$$

calls of the stochastic first order oracle.
Proof We first show part (a). Let $g_{X}\left(\bar{x}_{s}\right)=P_{X}\left(\bar{x}_{s}, \nabla f\left(\bar{x}_{s}\right), \gamma_{R_{s}}\right)$. Then, it follows from the definition of $\bar{x}^{*}$ in (4.27) that

$$
\begin{aligned}
\left\|\bar{g}_{X}\left(\bar{x}^{*}\right)\right\|^{2} & =\min _{s=1, \ldots, S}\left\|\bar{g}_{X}\left(\bar{x}_{S}\right)\right\|^{2}=\min _{s=1, \ldots, S}\left\|g_{X}\left(\bar{x}_{S}\right)+\bar{g}_{X}\left(\bar{x}_{s}\right)-g_{X}\left(\bar{x}_{S}\right)\right\|^{2} \\
& \leq \min _{s=1, \ldots, S}\left\{2\left\|g_{X}\left(\bar{x}_{S}\right)\right\|^{2}+2\left\|\bar{g}_{X}\left(\bar{x}_{s}\right)-g_{X}\left(\bar{x}_{S}\right)\right\|^{2}\right\} \\
& \leq 2 \min _{s=1, \ldots, S}\left\|g_{X}\left(\bar{x}_{S}\right)\right\|^{2}+2 \max _{s=1, \ldots, S}\left\|\bar{g}_{X}\left(\bar{x}_{S}\right)-g_{X}\left(\bar{x}_{S}\right)\right\|^{2},
\end{aligned}
$$

which implies that

$$
\begin{align*}
\left\|g_{X}\left(\bar{x}^{*}\right)\right\|^{2} \leq & 2\left\|\bar{g}_{X}\left(\bar{x}^{*}\right)\right\|^{2}+2\left\|g_{X}\left(\bar{x}^{*}\right)-\bar{g}_{X}\left(\bar{x}^{*}\right)\right\|^{2} \\
\leq & 4 \min _{s=1, \ldots, S}\left\|g_{X}\left(\bar{x}_{s}\right)\right\|^{2}+4 \max _{s=1, \ldots, S}\left\|\bar{g}_{X}\left(\bar{x}_{s}\right)-g_{X}\left(\bar{x}_{s}\right)\right\|^{2} \\
& +2\left\|g_{X}\left(\bar{x}^{*}\right)-\bar{g}_{X}\left(\bar{x}^{*}\right)\right\|^{2} \\
\leq & 4 \min _{s=1, \ldots, S}\left\|g_{X}\left(\bar{x}_{s}\right)\right\|^{2}+6 \max _{s=1, \ldots, S}\left\|\bar{g}_{X}\left(\bar{x}_{s}\right)-g_{X}\left(\bar{x}_{s}\right)\right\|^{2} . \tag{4.33}
\end{align*}
$$

We now provide certain probabilistic bounds to the two terms in the right hand side of the above inequality. Firstly, from the fact that $\bar{x}_{s}, 1 \leq s \leq S$, are independent and (4.25) (with $\lambda=2$ ), we have
$\operatorname{Prob}\left\{\min _{s \in\{1,2, \ldots, S\}}\left\|g_{X}\left(\bar{x}_{s}\right)\right\|^{2} \geq \frac{2 L \mathcal{B}_{\bar{N}}}{\alpha^{2}}\right\}=\prod_{s=1}^{S} \operatorname{Prob}\left\{\left\|g_{X}\left(\bar{x}_{s}\right)\right\|^{2} \geq \frac{2 L \mathcal{B}_{\bar{N}}}{\alpha^{2}}\right\} \leq 2^{-S}$.

Moreover, denoting $\delta_{s, k}=G\left(\bar{x}_{s}, \xi_{k}\right)-\nabla f\left(\bar{x}_{s}\right), k=1, \ldots, T$, by Proposition 1 with $x=\bar{x}_{s}, \gamma=\gamma_{R_{s}}, g_{1}=\bar{G}_{T}\left(\bar{x}_{s}\right), g_{2}=\nabla f\left(\bar{x}_{s}\right)$, we have

$$
\begin{equation*}
\left\|\bar{g}_{X}\left(\bar{x}_{s}\right)-g_{X}\left(\bar{x}_{s}\right)\right\| \leq \frac{1}{\alpha}\left\|\sum_{k=1}^{T} \delta_{s, k} / T\right\| . \tag{4.35}
\end{equation*}
$$

From the above inequality, Assumption A1 and Lemma 4(a), for any $\lambda>0$ and any $s=1, \ldots, S$, we have

$$
\operatorname{Prob}\left\{\left\|\bar{g}_{X}\left(\bar{x}_{s}\right)-g_{X}\left(\bar{x}_{s}\right)\right\|^{2} \geq \frac{\lambda \sigma^{2}}{\alpha^{2} T}\right\} \leq \operatorname{Prob}\left\{\left\|\sum_{k=1}^{T} \delta_{s, k}\right\|^{2} \geq \lambda T \sigma^{2}\right\} \leq \frac{1}{\lambda}
$$

which implies

$$
\begin{equation*}
\text { Prob }\left\{\max _{s=1, \ldots, S}\left\|\bar{g}_{X}\left(\bar{x}_{s}\right)-g_{X}\left(\bar{x}_{s}\right)\right\|^{2} \geq \frac{\lambda \sigma^{2}}{\alpha^{2} T}\right\} \leq \frac{S}{\lambda} \tag{4.36}
\end{equation*}
$$

Then, the conclusion (4.28) follows from (4.33), (4.34) and (4.36).
We now show part (b). With the settings in part (b), it is easy to count the total number of calls of the $\mathcal{S F} \mathcal{O}$ in the 2-RSPG algorithm is bounded up by (4.32). Hence, we only need to show that the $\bar{x}^{*}$ returned by the 2-RSPG algorithm is indeed an $(\epsilon, \Lambda)$-solution of the problem (1.1). With the choice of $\bar{N}(\epsilon)$ in (4.30), we can see that (4.22) holds. So, we have from (4.19) and (4.30) that

$$
\mathcal{B}_{\bar{N}(\epsilon)}=\frac{16 L D_{\Psi}^{2}}{\bar{N}(\epsilon)}+\frac{4 \sqrt{6} \sigma}{\sqrt{\bar{N}(\epsilon)}}\left(\tilde{D}+\frac{D_{\Psi}^{2}}{\tilde{D}}\right) \leq \frac{\alpha^{2} \epsilon}{32 L}+\frac{\alpha^{2} \epsilon}{32 L}=\frac{\alpha^{2} \epsilon}{16 L} .
$$

By the above inequality and (4.31), setting $\lambda=2 S / \Lambda$ in (4.28), we have

$$
\frac{8 L \mathcal{B}_{\bar{N}(\epsilon)}}{\alpha^{2}}+\frac{6 \lambda \sigma^{2}}{\alpha^{2} T(\epsilon, \Lambda)} \leq \frac{\epsilon}{2}+\frac{\lambda \Lambda \epsilon}{4 S}=\epsilon
$$

which together with (4.28), (4.29) and $\lambda=2 S / \Lambda$ imply

$$
\operatorname{Prob}\left\{\left\|g_{X}\left(\bar{x}^{*}\right)\right\|^{2} \geq \epsilon\right\} \leq \frac{\Lambda}{2}+2^{-S} \leq \Lambda
$$

Hence, $\bar{x}^{*}$ is an $(\epsilon, \Lambda)$-solution of the problem (1.1).
Now, it is interesting to compare the complexity bound in (4.32) with the one in (4.26). In view of (4.29), (4.30) and (4.31), the complexity bound in (4.32) for finding an $(\epsilon, \Lambda)$-solution, after discarding a few constant factors, is equivalent to

$$
\begin{equation*}
\mathcal{O}\left\{\frac{1}{\epsilon} \log _{2} \frac{1}{\Lambda}+\frac{\sigma^{2}}{\epsilon^{2}} \log _{2} \frac{1}{\Lambda}+\frac{\sigma^{2}}{\Lambda \epsilon} \log _{2}^{2} \frac{1}{\Lambda}\right\} . \tag{4.37}
\end{equation*}
$$

When the second terms are the dominating terms in both bounds, the above bound (4.37) can be considerably smaller than the one in (4.26) up to a factor of $1 /\left[\Lambda^{2} \log _{2}(1 / \Lambda)\right]$.

The following theorem shows that under a certain "light-tail" assumption:
A2 For any $x_{k} \in X$, we have

$$
\begin{equation*}
\mathbb{E}\left[\exp \left\{\left\|G\left(x_{k}, \xi_{k}\right)-\nabla f(x)\right\|^{2} / \sigma^{2}\right\}\right] \leq \exp \{1\} \tag{4.38}
\end{equation*}
$$

the bound (4.32) in Theorem 3 can be further improved.
Corollary 5 Under Assumptions A1 and A2, the following statements hold for the 2-RSPG algorithm applied to problem (1.1).
(a) Let $\mathcal{B}_{\bar{N}}$ is defined in (4.19). Then, for all $\lambda>0$
$\operatorname{Prob}\left\{\left\|g_{X}\left(\bar{x}^{*}\right)\right\|^{2} \geq\left[\frac{8 L \mathcal{B}_{\bar{N}}}{\alpha^{2}}+\frac{12(1+\lambda)^{2} \sigma^{2}}{T \alpha^{2}}\right]\right\} \leq S \exp \left(-\frac{\lambda^{2}}{3}\right)+2^{-S} ;$
(b) Let $\epsilon>0$ and $\Lambda \in(0,1)$ be given. If $S$ and $\bar{N}$ are set to $S(\Lambda)$ and $\bar{N}(\epsilon)$ as in (4.29) and (4.30), respectively, and the sample size $T$ is set to

$$
\begin{equation*}
T^{\prime}(\epsilon, \Lambda):=\frac{24 \sigma^{2}}{\alpha^{2} \epsilon}\left[1+\left(3 \log _{2} \frac{2 S(\Lambda)}{\Lambda}\right)^{\frac{1}{2}}\right]^{2} \tag{4.40}
\end{equation*}
$$

then the 2-RSPG algorithm can compute an $(\epsilon, \Lambda)$-solution of the problem (1.1) after taking at most

$$
\begin{equation*}
S(\Lambda)\left[\bar{N}(\epsilon)+T^{\prime}(\epsilon, \Lambda)\right] \tag{4.41}
\end{equation*}
$$

calls to the stochastic first-order oracle.
Proof We only give a sketch of the proof for part (a). The proof of part (b) follows from part (a) and similar arguments for proving (b) part of Theorem 3. Now, denoting $\delta_{s, k}=G\left(\bar{x}_{s}, \xi_{k}\right)-\nabla f\left(\bar{x}_{s}\right), k=1, \ldots, T$, again by Proposition 1, we have (4.35) holds. Then, by Assumption A2 and Lemma 4(b), for any $\lambda>0$ and any $s=1, \ldots, S$, we have

$$
\begin{aligned}
& \text { Prob }\left\{\left\|\bar{g}_{X}\left(\bar{x}_{s}\right)-g_{X}\left(\bar{x}_{s}\right)\right\|^{2} \geq(1+\lambda)^{2} \frac{2 \sigma^{2}}{\alpha^{2} T}\right\} \\
& \quad \leq \operatorname{Prob}\left\{\left\|\sum_{k=1}^{T} \delta_{s, k}\right\| \geq \sqrt{2 T}(1+\lambda) \sigma\right\} \leq \exp \left(-\frac{\lambda^{2}}{3}\right),
\end{aligned}
$$

which implies that for any $\lambda>0$

$$
\begin{equation*}
\text { Prob }\left\{\max _{s=1, \ldots, S}\left\|\bar{g}_{X}\left(\bar{x}_{s}\right)-g_{X}\left(\bar{x}_{s}\right)\right\|^{2} \geq(1+\lambda)^{2} \frac{2 \sigma^{2}}{\alpha^{2} T}\right\} \leq S \exp \left(-\frac{\lambda^{2}}{3}\right), \tag{4.42}
\end{equation*}
$$

Then, the conclusion (4.39) follows from (4.33), (4.34) and (4.42).
In view of (4.29), (4.30) and (4.40), the bound in (4.41), after discarding a few constant factors, is equivalent to

$$
\begin{equation*}
\mathcal{O}\left\{\frac{1}{\epsilon} \log _{2} \frac{1}{\Lambda}+\frac{\sigma^{2}}{\epsilon^{2}} \log _{2} \frac{1}{\Lambda}+\frac{\sigma^{2}}{\epsilon} \log _{2}^{2} \frac{1}{\Lambda}\right\} \tag{4.43}
\end{equation*}
$$

Clearly, the third term of the above bound is smaller than the third term in (4.37) by a factor of $1 / \Lambda$.

In the remaining part of this section, we briefly discuss another variant of the 2-RSPG algorithm, namely, 2-RSPG-V algorithm which can improve the practical performance of the 2-RSPG algorithm (see Sect. 6). Similarly to the 2-RSPG algorithm, this variant also consists of two phases. The only difference exists in that the $S$ runs of the RSPG algorithm in the optimization phase are not independent of each other and the output of each run is used as the initial point of the next run, although the post-optimization phase of the 2-RSPG-V algorithm is the same as that of the 2-RSPG algorithm. We now formally state the optimization phase of the 2-RSPG-V algorithm as follows.

Optimization phase of 2-RSPG-V algorithm:
For $s=1, \ldots, S$
Call the RSPG algorithm with initial point $\bar{x}_{s-1}$ where $\bar{x}_{0}=x_{1}$ and $\bar{x}_{s}=x_{R_{s}}, s=$ $1, \ldots, S$, are the outputs of the $s$-th run of the RSPG algorithm, iteration limit $N=\lfloor\bar{N} / m\rfloor$ with $m$ given by (4.18), stepsizes $\gamma_{k}=\alpha /(2 L)$ for $k=1, \ldots, N$, batch sizes $m_{k}=m$, and probability mass function $P_{R}$ in (4.3).
As mentioned above, in the 2-RSPG-V algorithm, unlike the 2-RSPG algorithm, the $S$ candidate solutions are not independent and hence the analysis of Theorem 3 cannot be directly applied. However, by slightly modifying the proof of Theorem 3, we can show that the above 2-RSPG-V algorithm exhibits similar convergence behavior as the 2-RSPG algorithm under certain more restrictive conditions.

Corollary 6 Suppose that the feasible set $X$ is bounded and Assumption Al holds. Then, the complexity of the 2-RSPG-V algorithm to find an $(\epsilon, \Lambda)$-solution of problem (1.1) is bounded by (4.37). If in addition, Assumption A2 holds, then this complexity bound improves to (4.43).

Proof Denote $\bar{\Psi}=\max _{x \in X} \Psi(x)$ and let $E_{S}$ be the event that $\left\|g_{X}\left(\bar{x}_{s}\right)\right\|^{2} \geq \frac{2 L \hat{\mathcal{B}}_{\bar{N}}}{\alpha^{2}}$ where

$$
\hat{\mathcal{B}}_{\bar{N}}:=\frac{16\left(\bar{\Psi}-\Psi^{*}\right)}{\bar{N}}+\frac{4 \sqrt{6} \sigma}{\sqrt{\bar{N}}}\left(\frac{\bar{\Psi}-\Psi^{*}}{L \tilde{D}}+\tilde{D} \max \left\{1, \frac{\sqrt{6} \sigma}{4 L \tilde{D} \sqrt{\bar{N}}}\right\}\right)
$$

Now note that due to the boundedness of $X$ and continuity of $f, \bar{\Psi}$ is finite and therefore the bound $\hat{\mathcal{B}}_{\bar{N}}$ is valid. Also observe that by (4.25) (with $\lambda=2$ ) together with the fact that $\hat{\mathcal{B}}_{\bar{N}} \geq \mathcal{B}_{\bar{N}}$, we have

$$
\operatorname{Prob}\left\{E_{s} \mid \bigcap_{j=1}^{s-1} E_{j}\right\} \leq \frac{1}{2}, \quad s=1,2, \ldots, S
$$

which consequently implies that

$$
\begin{aligned}
\operatorname{Prob} & \left\{\min _{s \in\{1,2, \ldots, S\}}\left\|g_{X}\left(\bar{x}_{s}\right)\right\|^{2} \geq \frac{2 L \hat{\mathcal{B}}_{\bar{N}}}{\alpha^{2}}\right\} \\
& =\operatorname{Prob}\left\{\bigcap_{s=1}^{S} E_{S}\right\}=\prod_{s=1}^{S} \operatorname{Prob}\left\{E_{S} \mid \bigcap_{j=1}^{s-1} E_{j}\right\} \leq 2^{-S} .
\end{aligned}
$$

Observing that the above inequality is similar to (4.34), the rest of proof is almost identical to those of Theorem 3 and Corollary 5 and hence we skip the details.

## 5 Stochastic zeroth-order methods

In this section, we discuss how to specialize the RSPG algorithm to deal with the situations where only noisy function values of the problem (1.1) are available. More specifically, we assume that we can only access the noisy zeroth-order information of $f$ by a stochastic zeroth-order oracle $(\mathcal{S Z O})$. For any input $x_{k}$ and $\xi_{k}$, the $\mathcal{S Z O}$ would output a quantity $F\left(x_{k}, \xi_{k}\right)$, where $x_{k}$ is the $k$-th iterate of our algorithm and $\xi_{k}$ is a random variable whose distribution is supported on $\Xi \in \mathbb{R}^{d}$ (noting that $\Xi$ does not depend on $x_{k}$ ). Throughout this section, we assume $F\left(x_{k}, \xi_{k}\right)$ is an unbiased estimator of $f\left(x_{k}\right)$, that is

A3 For any $k \geq 1$, we have

$$
\begin{equation*}
\mathbb{E}\left[F\left(x_{k}, \xi_{k}\right)\right]=f\left(x_{k}\right) \tag{5.1}
\end{equation*}
$$

We are going to apply the randomized smoothing techniques (see e.g., $[9,11,27,29]$ ) to explore the zeroth-order information of $f$. Hence, throughout this section, we also assume $F\left(\cdot, \xi_{k}\right) \in \mathcal{C}_{L}^{1,1}\left(\mathbb{R}^{n}\right)$ almost surely with respect to $\xi_{k} \in \Xi$, which together with Assumption A3 imply $f \in \mathcal{C}_{L}^{1,1}\left(\mathbb{R}^{n}\right)$. Also, throughout this section, we assume that $\|\cdot\|$ is the standard Euclidean norm.

Suppose $v$ is a random vector in $\mathbb{R}^{n}$ with density function $\rho$, a smooth approximation of $f$ is defined as

$$
\begin{equation*}
f_{\mu}(x)=\int f(x+\mu v) \rho(v) d v \tag{5.2}
\end{equation*}
$$

where $\mu>0$ is the smoothing parameter. For different choices of smoothing distribution, the smoothed function $f_{\mu}$ would have different properties. In this section, we only consider the Gaussian smoothing distribution. That is we assume that $v$ is a $n$-dimensional standard Gaussian random vector and

$$
\begin{equation*}
f_{\mu}(x)=\frac{1}{(2 \pi)^{\frac{n}{2}}} \int f(x+\mu v) e^{-\frac{1}{2}\|v\|^{2}} d v=\mathbb{E}_{v}[f(x+\mu v)] . \tag{5.3}
\end{equation*}
$$

Nesterov [29] showed that the Gaussian smoothing approximation and $f_{\mu}$ have the following nice properties.

Lemma 5 If $f \in \mathcal{C}_{L}^{1,1}\left(\mathbb{R}^{n}\right)$, then
(a) $f_{\mu}$ is also Lipschitz continuously differentiable with gradient Lipschitz constant $L_{\mu} \leq L$ and

$$
\begin{equation*}
\nabla f_{\mu}(x)=\frac{1}{(2 \pi)^{\frac{n}{2}}} \int \frac{f(x+\mu v)-f(x)}{\mu} v e^{-\frac{1}{2}\|v\|^{2}} d v \tag{5.4}
\end{equation*}
$$

(b) for any $x \in \mathbb{R}^{n}$, we have

$$
\begin{align*}
& \left|f_{\mu}(x)-f(x)\right| \leq \frac{\mu^{2}}{2} L n  \tag{5.5}\\
& \left\|\nabla f_{\mu}(x)-\nabla f(x)\right\| \leq \frac{\mu}{2} L(n+3)^{\frac{3}{2}}  \tag{5.6}\\
& \mathbb{E}_{v}\left[\left\|\frac{f(x+\mu v)-f(x)}{\mu} v\right\|^{2}\right] \leq 2(n+4)\|\nabla f(x)\|^{2}+\frac{\mu^{2}}{2} L^{2}(n+6)^{3} . \tag{5.7}
\end{align*}
$$

(c) $f_{\mu}$ is also convex provided $f$ is convex.

In the following, let us define the approximated stochastic gradient of $f$ at $x_{k}$ as

$$
\begin{equation*}
G_{\mu}\left(x_{k}, \xi_{k}, v\right)=\frac{F\left(x_{k}+\mu v, \xi_{k}\right)-F\left(x_{k}, \xi_{k}\right)}{\mu} v, \tag{5.8}
\end{equation*}
$$

and define $G\left(x_{k}, \xi_{k}\right)=\nabla_{x} F\left(x_{k}, \xi_{k}\right)$. We assume the Assumption 1 holds for $G\left(x_{k}, \xi_{k}\right)$. Then, by the Assumption A3 and Lemma 5(a), we directly get

$$
\begin{equation*}
\mathbb{E}_{v, \xi_{k}}\left[G_{\mu}\left(x_{k}, \xi_{k}, v\right)\right]=\nabla f_{\mu}\left(x_{k}\right), \tag{5.9}
\end{equation*}
$$

where the expectation is taken with respect to $v$ and $\xi_{k}$.
Now based on the RSPG algorithm, we state an algorithm which only uses zerothorder information to solve problem (1.1).
A randomized stochastic projected gradient free (RSPGF) algorithm

Input: Given initial point $x_{1} \in X$, iteration limit $N$, the stepsizes $\left\{\gamma_{k}\right\}$ with $\gamma_{k}>$ $0, k \geq 1$, the batch sizes $\left\{m_{k}\right\}$ with $m_{k}>0, k \geq 1$, and the probability mass function $P_{R}$ supported on $\{1, \ldots, N\}$.
Step 0 . Let $R$ be a random variable with probability mass function $P_{R}$.
Step $k=1, \ldots, R-1$. Call the $\mathcal{S Z O} m_{k}$ times to obtain $G_{\mu}\left(x_{k}, \xi_{k, i}, v_{k, i}\right), i=$ $1, \ldots, m_{k}$, set

$$
\begin{equation*}
G_{\mu, k}=\frac{1}{m_{k}} \sum_{i=1}^{m_{k}} G_{\mu}\left(x_{k}, \xi_{k, i}, v_{k, i}\right) \tag{5.10}
\end{equation*}
$$

and compute

$$
\begin{equation*}
x_{k+1}=\arg \min _{u \in X}\left\{\left\langle G_{\mu, k}, u\right\rangle+\frac{1}{\gamma_{k}} V\left(u, x_{k}\right)+h(u)\right\} . \tag{5.11}
\end{equation*}
$$

Output: $x_{R}$.
Compared with RSPG algorithm, we can see at the $k$-th iteration, the RSPGF algorithm simply replaces the stochastic gradient $G_{k}$ by the approximated stochastic gradient $G_{\mu, k}$. By (5.9), $G_{\mu, k}$ can be simply viewed as an unbiased stochastic gradient of the smoothed function $f_{\mu}$. However, to apply the results developed in the previous section, we still need an estimation of the bound on the variations of the stochastic gradient $G_{\mu, k}$. In addition, the role that the smoothing parameter $\mu$ plays and the proper selection of $\mu$ in the RSPGF algorithm are still not clear now. We answer these questions in the following series of theorems and their corollaries.

Theorem 4 Suppose that the stepsizes $\left\{\gamma_{k}\right\}$ in the RSPGF algorithm are chosen such that $0<\gamma_{k} \leq \alpha / L$ with $\gamma_{k}<\alpha / L$ for at least one $k$, and the probability mass function $P_{R}$ are chosen as (4.3). If $\|\nabla f(x)\| \leq M$ for all $x \in X$, then under Assumptions AI and A3,
(a) for any $N \geq 1$, we have

$$
\begin{equation*}
\mathbb{E}\left[\left\|\bar{g}_{\mu, X, R}\right\|^{2}\right] \leq \frac{L D_{\psi}^{2}+\mu^{2} L n+\left(\tilde{\sigma}^{2} / \alpha\right) \sum_{k=1}^{N}\left(\gamma_{k} / m_{k}\right)}{\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right)}, \tag{5.12}
\end{equation*}
$$

where the expectation is taken with respect to $R, \xi_{[N]}$ and $v_{[N]}:=\left(v_{1}, \ldots, v_{N}\right)$, $D_{\Psi}$ is defined in (3.5),

$$
\begin{equation*}
\tilde{\sigma}^{2}=2(n+4)\left[M^{2}+\sigma^{2}+\mu^{2} L^{2}(n+4)^{2}\right], \tag{5.13}
\end{equation*}
$$

and

$$
\begin{equation*}
\bar{g}_{\mu, X, k}=P_{X}\left(x_{k}, G_{\mu, k}, \gamma_{k}\right), \tag{5.14}
\end{equation*}
$$

with $P_{X}$ defined in(2.4);
(b) if, in addition, $f$ in problem (1.1) is convex with an optimal solution $x^{*}$, and the stepsizes $\left\{\gamma_{k}\right\}$ are non-decreasing as (4.6), we have

$$
\begin{equation*}
\mathbb{E}\left[\Psi\left(x_{R}\right)-\Psi\left(x^{*}\right)\right] \leq \frac{\left(\alpha-L \gamma_{1}\right) V\left(x^{*}, x_{1}\right)+\left(\tilde{\sigma}^{2} / 2\right) \sum_{k=1}^{N}\left(\gamma_{k}^{2} / m_{k}\right)}{\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right)}+\mu^{2} L n \tag{5.15}
\end{equation*}
$$

where the expectation is taken with respect to $R, \xi_{[N]}$ and $v_{[N]}$.
Proof By our assumption that $F\left(\cdot, \xi_{k}\right) \in \mathcal{C}_{L}^{1,1}\left(\mathbb{R}^{n}\right)$ almost surely and (5.7) (applying $f=F\left(\cdot, \xi_{k}\right)$ ), we have

$$
\begin{aligned}
\mathbb{E}_{v_{k}, \xi_{k}}\left[\left\|G_{\mu}\left(x_{k}, \xi_{k}, v_{k}\right)\right\|^{2}\right] & =\mathbb{E}_{\xi_{k}}\left[\mathbb{E}_{v_{k}}\left[\left\|G_{\mu}\left(x_{k}, \xi_{k}, v_{k}\right)\right\|^{2}\right]\right] \\
& \leq 2(n+4)\left[\mathbb{E}_{\xi_{k}}\left[\left\|G\left(x_{k}, \xi\right)\right\|^{2}\right]+\frac{\mu^{2}}{2} L^{2}(n+6)^{3}\right. \\
& \leq 2(n+4)\left[\mathbb{E}_{\xi_{k}}\left[\left\|\nabla f\left(x_{k}\right)\right\|^{2}\right]+\sigma^{2}\right]+2 \mu^{2} L^{2}(n+4)^{3}
\end{aligned}
$$

where the last inequality follows from Assumption 1 with $G\left(x_{k}, \xi_{k}\right)=\nabla_{x} F\left(x_{k}, \xi_{k}\right)$. Then, from (5.9), the above inequality, and $\left\|\nabla f\left(x_{k}\right)\right\| \leq M$, we have

$$
\begin{align*}
& \mathbb{E}_{v_{k}, \xi_{k}}\left[\left\|G_{\mu}\left(x_{k}, \xi_{k}, v_{k}\right)-\nabla f_{\mu}\left(x_{k}\right)\right\|^{2}\right] \\
& =\mathbb{E}_{v_{k}, \xi_{k}}\left[\left\|G_{\mu}\left(x_{k}, \xi_{k}, v_{k}\right)\right\|^{2}+\left\|\nabla f_{\mu}\left(x_{k}\right)\right\|^{2}-2\left\langle G_{\mu}\left(x_{k}, \xi_{k}, v_{k}\right), \nabla f_{\mu}\left(x_{k}\right)\right\rangle\right] \\
& =\mathbb{E}_{v_{k}, \xi_{k}}\left[\left\|G_{\mu}\left(x_{k}, \xi_{k}, v_{k}\right)\right\|^{2}\right]+\left\|\nabla f_{\mu}\left(x_{k}\right)\right\|^{2}-2\left\langle\mathbb{E}_{v_{k}, \xi_{k}}\left[G_{\mu}\left(x_{k}, \xi_{k}, v_{k}\right)\right], \nabla f_{\mu}\left(x_{k}\right)\right\rangle \\
& =\mathbb{E}_{v_{k}, \xi_{k}}\left[\left\|G_{\mu}\left(x_{k}, \xi_{k}, v_{k}\right)\right\|^{2}\right]+\left\|\nabla f_{\mu}\left(x_{k}\right)\right\|^{2}-2\left\|\nabla f_{\mu}\left(x_{k}\right)\right\|^{2} \\
& \leq \mathbb{E}_{v_{k}, \xi_{k}}\left[\left\|G_{\mu}\left(x_{k}, \xi_{k}, v_{k}\right)\right\|^{2}\right] \leq 2(n+4)\left[M^{2}+\sigma^{2}+\mu^{2} L^{2}(n+4)^{2}\right]=\tilde{\sigma}^{2} . \tag{5.16}
\end{align*}
$$

Now let $\Psi_{\mu}(x)=f_{\mu}(x)+h(x)$ and $\Psi_{\mu}^{*}=\min _{x \in X} \Psi_{\mu}(x)$. We have from (5.5) that

$$
\begin{equation*}
\left|\left(\Psi_{\mu}(x)-\Psi_{\mu}^{*}\right)-\left(\Psi(x)-\Psi^{*}\right)\right| \leq \mu^{2} \operatorname{Ln} . \tag{5.17}
\end{equation*}
$$

By Lemma 5(a), we have $L_{\mu} \leq L$ and therefore $f_{\mu} \in \mathcal{C}_{L}^{1,1}\left(\mathbb{R}^{n}\right)$. With this observation, noticing (5.9) and (5.16), viewing $G_{\mu}\left(x_{k}, \xi_{k}, v_{k}\right)$ as a stochastic gradient of $f_{\mu}$, then by part (a) of Theorem 2 we can directly get

$$
\mathbb{E}\left[\left\|\bar{g}_{\mu, X, R}\right\|^{2}\right] \leq \frac{L D_{\Psi_{\mu}}^{2}+\left(\tilde{\sigma}^{2} / \alpha\right) \sum_{k=1}^{N}\left(\gamma_{k} / m_{k}\right)}{\sum_{k=1}^{N}\left(\alpha \gamma_{k}-L \gamma_{k}^{2}\right)},
$$

where $D_{\Psi_{\mu}}=\left[\left(\Psi_{\mu}\left(x_{1}\right)-\Psi_{\mu}^{*}\right) / L\right]^{1 / 2}$ and the expectation is taken with respect to $R, \xi_{[N]}$ and $v_{[N]}$. Then, the conclusion (5.12) follows the above inequality and (5.17).

We now show part (b). Since $f$ is convex, by Lemma 5(c), $f_{\mu}$ is also convex. Again by (5.17), we have

$$
\mathbb{E}\left[\Psi\left(x_{R}\right)-\Psi\left(x^{*}\right)\right] \leq \mathbb{E}\left[\Psi_{\mu}\left(x_{R}\right)-\Psi_{\mu}\left(x^{*}\right)\right]+\mu^{2} L n .
$$

Then, by this inequality and the convexity of $f_{\mu}$, it follows from part (b) of Theorem 2 and similar arguments in showing the part (a) of this theorem, the conclusion (5.15) holds.

Using the previous Theorem 4, similar to the Corollary 3, we can give the following corollary on the RSPGF algorithm with a certain constant stepsize and batch size at each iteration.

Corollary 7 Suppose that in the RSPGF algorithm the stepsizes $\gamma_{k}=\alpha /(2 L)$ for all $k=1, \ldots, N$, the batch sizes $m_{k}=m$ for all $k=1, \ldots, N$, and the probability mass function $P_{R}$ is set to (4.3).

Then under Assumptions A1 and A3, we have

$$
\begin{equation*}
\mathbb{E}\left[\left\|\bar{g}_{\mu, X, R}\right\|^{2}\right] \leq \frac{4 L^{2} D_{\Psi}^{2}+4 \mu^{2} L^{2} n}{\alpha^{2} N}+\frac{2 \tilde{\sigma}^{2}}{\alpha^{2} m} \tag{5.18}
\end{equation*}
$$

and

$$
\begin{equation*}
\mathbb{E}\left[\left\|g_{X, R}\right\|^{2}\right] \leq \frac{\mu^{2} L^{2}(n+3)^{2}}{2 \alpha^{2}}+\frac{16 L^{2} D_{\Psi}^{2}+16 \mu^{2} L^{2} n}{\alpha^{2} N}+\frac{12 \tilde{\sigma}^{2}}{\alpha^{2} m} \tag{5.19}
\end{equation*}
$$

where the expectation is taken with respect to $R, \xi_{[N]}$ and $v_{[N]}$, and $\tilde{\sigma}, \bar{g}_{\mu, X, R}$ and $g_{X, R}$ are defined in (5.13), (5.14) and (3.3), respectively.

If, in addition, $f$ in the problem (1.1) is convex with an optimal solution $x^{*}$, then

$$
\begin{equation*}
\mathbb{E}\left[\Psi\left(x_{R}\right)-\Psi\left(x^{*}\right)\right] \leq \frac{2 L V\left(x^{*}, x_{1}\right)}{N \alpha}+\frac{\tilde{\sigma}^{2}}{2 L m}+\mu^{2} L n . \tag{5.20}
\end{equation*}
$$

Proof (5.18) immediately follows from (5.12) with $\gamma_{k}=\alpha /(2 L)$ and $m_{k}=m$ for all $k=1, \ldots, N$. Now let $g_{\mu, X, R}=P_{X}\left(x_{R}, \nabla f_{\mu}\left(x_{R}\right), \gamma_{R}\right)$, we have from (5.6) and Proposition 1 with $x=x_{R}, \gamma=\gamma_{R}, g_{1}=\nabla f\left(x_{R}\right)$ and $g_{2}=\nabla f_{\mu}\left(x_{R}\right)$ that

$$
\begin{equation*}
\mathbb{E}\left[\left\|g_{X, R}-g_{\mu, X, R}\right\|^{2}\right] \leq \frac{\mu^{2} L^{2}(n+3)^{2}}{4 \alpha^{2}} \tag{5.21}
\end{equation*}
$$

Similarly, by Proposition 1 with $x=x_{R}, \gamma=\gamma_{R}, g_{1}=\bar{G}_{\mu, k}$ and $g_{2}=\nabla f_{\mu}\left(x_{R}\right)$, we have

$$
\begin{equation*}
\mathbb{E}\left[\left\|\bar{g}_{\mu, X, R}-g_{\mu, X, R}\right\|^{2}\right] \leq \frac{\tilde{\sigma}^{2}}{\alpha^{2} m} \tag{5.22}
\end{equation*}
$$

Then, it follows from (5.21), (5.22) and (5.18) that

$$
\begin{aligned}
\mathbb{E}\left[\left\|g_{X, R}\right\|^{2}\right] & \leq 2 \mathbb{E}\left[\left\|g_{X, R}-g_{\mu, X, R}\right\|^{2}\right]+2 \mathbb{E}\left[\left\|g_{\mu, X, R}\right\|^{2}\right] \\
& \leq \frac{\mu^{2} L^{2}(n+3)^{2}}{2 \alpha^{2}}+4 \mathbb{E}\left[\left\|g_{\mu, X, R}-\bar{g}_{\mu, X, R}\right\|^{2}\right]+4 \mathbb{E}\left[\left\|\bar{g}_{\mu, X, R}\right\|^{2}\right] \\
& \leq \frac{\mu^{2} L^{2}(n+3)^{2}}{2 \alpha^{2}}+\frac{12 \tilde{\sigma}^{2}}{\alpha^{2} m}+\frac{16 L^{2} D_{\Psi}^{2}+16 \mu^{2} L^{2} n}{\alpha^{2} N}
\end{aligned}
$$

Moreover, if $f$ is convex, then (5.20) immediately follows from (5.15), and the constant stepsizes $\gamma_{k}=\alpha /(2 L)$ for all $k=1, \ldots, N$.

Similar to the Corollary 3 for the RSPG algorithm, the above results also depend on the number of samples $m$ at each iteration. In addition, the above results depend on the smoothing parameter $\mu$ as well. The following corollary, analogous to the Corollary 4, shows how to choose $m$ and $\mu$ appropriately.

Corollary 8 Suppose that all the conditions in Corollary 7 are satisfied. Given a fixed total number of calls to the $\mathcal{S Z O} \bar{N}$, if the smoothing parameter satisfies

$$
\begin{equation*}
\mu \leq \frac{D_{\Psi}}{\sqrt{(n+4) \bar{N}}}, \tag{5.23}
\end{equation*}
$$

and the number of calls to the $\mathcal{S Z O}$ at each iteration of the RSPGF method is

$$
\begin{equation*}
m=\left\lceil\min \left\{\max \left\{\frac{\sqrt{(n+4)\left(M^{2}+\sigma^{2}\right) \bar{N}}}{L \tilde{D}}, n+4\right\}, \bar{N}\right\}\right\rceil, \tag{5.24}
\end{equation*}
$$

for some $\tilde{D}>0$, then we have $\left(\alpha^{2} / L\right) \mathbb{E}\left[\left\|g_{X, R}\right\|^{2}\right] \leq \overline{\mathcal{B}}_{\bar{N}}$, where

$$
\begin{equation*}
\overline{\mathcal{B}}_{\bar{N}}:=\frac{\left(24 \theta_{2}+41\right) L D_{\Psi}^{2}(n+4)}{\bar{N}}+\frac{32 \sqrt{(n+4)\left(M^{2}+\sigma^{2}\right)}}{\sqrt{\bar{N}}}\left(\frac{D_{\Psi}^{2}}{\tilde{D}}+\tilde{D} \theta_{1}\right), \tag{5.25}
\end{equation*}
$$

and

$$
\begin{equation*}
\theta_{1}=\max \left\{1, \frac{\sqrt{(n+4)\left(M^{2}+\sigma^{2}\right)}}{L \tilde{D} \sqrt{\bar{N}}}\right\} \text { and } \theta_{2}=\max \left\{1, \frac{n+4}{\bar{N}}\right\} \tag{5.26}
\end{equation*}
$$

If, in addition, $f$ in the problem (1.1) is convex and the smoothing parameter satisfies

$$
\begin{equation*}
\mu \leq \sqrt{\frac{V\left(x^{*}, x_{1}\right)}{\alpha(n+4) \bar{N}}}, \tag{5.27}
\end{equation*}
$$

then $\mathbb{E}\left[\Psi\left(x_{R}\right)-\Psi\left(x^{*}\right)\right] \leq \overline{\mathcal{C}}_{\bar{N}}$, where $x^{*}$ is an optimal solution and

$$
\begin{equation*}
\overline{\mathcal{C}}_{\bar{N}}:=\frac{\left(5+\theta_{2}\right) L V\left(x^{*}, x_{1}\right)(n+4)}{\alpha \bar{N}}+\frac{\sqrt{(n+4)\left(M^{2}+\sigma^{2}\right)}}{\alpha \sqrt{\bar{N}}}\left(\frac{4 V\left(x^{*}, x_{1}\right)}{\tilde{D}}+\alpha \tilde{D} \theta_{1}\right) . \tag{5.28}
\end{equation*}
$$

Proof By the definitions of $\theta 1$ and $\theta_{2}$ in (5.26) and $m$ in (5.24), we have

$$
\begin{equation*}
m=\left\lceil\max \left\{\frac{\sqrt{(n+4)\left(M^{2}+\sigma^{2}\right) \bar{N}}}{L \tilde{D} \theta_{1}}, \frac{n+4}{\theta_{2}}\right\}\right\rceil \tag{5.29}
\end{equation*}
$$

Given the total number of calls to the $\mathcal{S Z O} \bar{N}$ and the number $m$ of calls to the $\mathcal{S Z O}$ at each iteration, the RSPGF algorithm can perform at most $N=\lfloor\bar{N} / \mathrm{m}\rfloor$ iterations. Obviously, $N \geq \bar{N} /(2 m)$. With this observation $\bar{N} \geq m, \theta_{1} \geq 1$ and $\theta_{2} \geq 1$, by (5.19), (5.23) and (5.29), we have

$$
\begin{aligned}
& \mathbb{E}\left[\left\|g_{X, R}\right\|^{2}\right] \\
& \leq \frac{L^{2} D_{\Psi}^{2}(n+3)}{2 \alpha^{2} \bar{N}}+\frac{24(n+4)\left(M^{2}+\sigma^{2}\right)}{\alpha^{2} m}+\frac{24 L^{2} D_{\Psi}^{2}(n+4)^{2}}{\alpha^{2} m \bar{N}}+\frac{32 L^{2} D_{\Psi}^{2} m}{\alpha^{2} \bar{N}} \\
& \leq \frac{L^{2} D_{\Psi}^{2}\left(n+\frac{1}{\bar{N}}\right)}{2 \alpha^{2} \bar{N}}+\frac{24 \theta_{1} L \tilde{D} \sqrt{(n+4)\left(M^{2}+\sigma^{2}\right)}}{\alpha^{2} \sqrt{\bar{N}}}+\frac{24 \theta_{2} L^{2} D_{\Psi}^{2}(n+4)}{\alpha^{2} \bar{N}} \\
&+\frac{32 L^{2} D_{\Psi}^{2}}{\alpha^{2} \bar{N}}\left(\frac{\sqrt{(n+4)\left(M^{2}+\sigma^{2}\right) \bar{N}}}{L \tilde{D} \theta_{1}}+\frac{n+4}{\theta_{2}}\right)+\frac{32 L^{2} D_{\Psi}^{2}}{\alpha^{2} \bar{N}} \\
& \leq \frac{L^{2} D_{\Psi}^{2}(n+4)}{2 \alpha^{2} \bar{N}}+\frac{24 \theta_{1} L \tilde{D} \sqrt{(n+4)\left(M^{2}+\sigma^{2}\right)}}{\alpha^{2} \sqrt{\bar{N}}}+\frac{24 \theta_{2} L^{2} D_{\Psi}^{2}(n+4)}{\alpha^{2} \bar{N}} \\
&+\frac{32 L D_{\Psi}^{2} \sqrt{(n+4)\left(M^{2}+\sigma^{2}\right)}}{\alpha^{2} \tilde{D} \sqrt{\bar{N}}}+\frac{32 L^{2} D_{\Psi}^{2}(n+4)}{\alpha^{2} \bar{N}}+\frac{32 L^{2} D_{\Psi}^{2}}{\alpha^{2} \bar{N}},
\end{aligned}
$$

which after integrating the terms give (5.25). The conclusion (5.28) follows similarly by (5.27) and (5.20).

We now would like to add a few remarks about the above the results in Corollary 8. Firstly, the above complexity bounds are similar to those of the first-order RSPG method in Corollary 4 in terms of their dependence on the total number of stochastic oracle $\bar{N}$ called by the algorithm. However, for the zeroth-order case, the complexity in Corollary 8 also depends on the size of the gradient $M$ and the problem dimension $n$. Secondly, the value of $\tilde{D}$ has not been specified. It can be easily seen from (5.25) and (5.28) that when $\bar{N}$ is relatively large such that $\theta_{1}=1$ and $\theta_{2}=1$, i.e.,

$$
\begin{equation*}
\bar{N} \geq \max \left\{\frac{(n+4)^{2}\left(M^{2}+\sigma^{2}\right)}{L^{2} \tilde{D}^{2}}, n+4\right\} \tag{5.30}
\end{equation*}
$$

the optimal choice of $\tilde{D}$ would be $D_{\Psi}$ and $2 \sqrt{V\left(x^{*}, x_{1}\right) / \alpha}$ for solving nonconvex and convex SP problems, respectively. With this selection of $\tilde{D}$, the bounds in (5.25) and (5.28), respectively, reduce to

$$
\begin{equation*}
\frac{\alpha^{2}}{L} \mathbb{E}\left[\left\|g_{X, R}\right\|^{2}\right] \leq \frac{65 L D_{\Psi}^{2}(n+4)}{\bar{N}}+\frac{64 \sqrt{(n+4)\left(M^{2}+\sigma^{2}\right)}}{\sqrt{\bar{N}}} \tag{5.31}
\end{equation*}
$$

and
$\mathbb{E}\left[\Psi\left(x_{R}\right)-\Psi\left(x^{*}\right)\right] \leq \frac{6 L V\left(x^{*}, x_{1}\right)(n+4)}{\alpha \bar{N}}+\frac{4 \sqrt{V\left(x^{*}, x_{1}\right)(n+4)\left(M^{2}+\sigma^{2}\right)}}{\sqrt{\alpha \bar{N}}}$.

Thirdly, the complexity result in (5.28) implies that when $\Psi$ is convex, if $\epsilon$ sufficiently small, then the number of calls to the $\mathcal{S Z O}$ to find a solution $\bar{x}$ such that $\mathbb{E}[\Psi(\bar{x})-$ $\left.\Psi^{*}\right] \leq \epsilon$ can be bounded by $\mathcal{O}\left(n / \epsilon^{2}\right)$, which is better than the complexity of $\mathcal{O}\left(n^{2} / \epsilon^{2}\right)$ established by Nesterov [29] to find such a solution for general convex SP problems.

## 6 Numerical results

In this section, we present the numerical results of our computational experiments for solving two SP problems: a stochastic nonconvex semi-supervised support vector machine problem and a simulation-based inventory optimization problem.

Algorithmic schemes We implement the RSPG algorithm and its two-phase variants 2-RSPG and 2-RSPG-V algorithms described in Section 4, where the prox-function $V(x, z)=\|x-z\|^{2} / 2$, the stepsizes $\gamma_{k}=\alpha /(2 L)$ with $\alpha=1$ for all $k \geq 1$, and the probability mass function $P_{R}$ is set to (4.3). Also, in the optimization phase of the 2-RSPG (2-RSPG-V) algorithm, we take $S=5$ independent (consecutive) runs of the RSPG algorithm to compute 5 candidate solutions. Then, we use an i.i.d. sample of size $T=\bar{N} / 2$ in the post-optimization phase to estimate the projected gradients at these candidate solutions and then choose the best one, $\bar{x}^{*}$, according to (4.27). Finally, the solution quality at $\bar{x}^{*}$ is evaluated by using another i.i.d. sample of size $K \gg \bar{N}$.
Estimation of parameters We use an initial i.i.d. sample of size $N_{0}=200$ to estimate the problem parameters, namely, $L$ and $\sigma$. In particular, for the first problem in our numerical experiments, we know the structure of the objective functions. Thus, we compute $l_{2}$-norm of the Hessian of the deterministic approximation of the objective functions obtained by the SAA approach with 200 samples, as an estimation of $L$. Using these sample, we also compute the stochastic gradients of the objective function 20 times at 10 randomly selected points and then take the average of the variances of the stochastic gradients for each point as an estimation of $\sigma^{2}$.

For the inventory problem, since we have no information about the objective function, we randomly generate 10 points and for each point, we call the stochastic oracle 20 times. Then, we estimate the stochastic gradients by (5.8) and take the average of them for each point, say $\bar{G}_{\mu}\left(x_{i}\right), \quad i=1, \ldots 10$, as an approximation of the true gradient. Finally, we consider the average of $\left\|\bar{G}_{\mu}\left(x_{i}\right)-\bar{G}_{\mu}\left(x_{j}\right)\right\| /\left\|x_{i}-x_{j}\right\|$ for all pairs of $i$ and $j$ as an estimation of $L$ as well as the average of variances of the stochastic gradients used for computing each $\bar{G}_{\mu}\left(x_{i}\right), \quad i=1, \ldots 10$, as an estimation of $\sigma^{2}$. We also estimate the parameter $\tilde{D}=D_{\Psi}$ by (3.5). More specifically, since the problems considered in this section have nonnegative optimal values, i.e., $\Psi^{*} \geq 0$, we have $D_{\Psi} \leq\left(\Psi\left(x_{1}\right) / L\right)^{\frac{1}{2}}$, where $x_{1}$ denotes the starting point of the algorithms.

Since previously there do not exist SA type methods with guaranteed convergence for solving nonconvex composite SP problems discussed in this paper, in practice one might simply assume that these problems are convex and then apply some existing convex SA methods to solve them. Hence, in our experiments, we also report the solutions obtained by taking the average of the trajectory of running the RSPG method for $N$ iterations. This approach is essentially the mirror descent SA (MD-SA) method in [21,22,26].

## Notation in the tables

- $\bar{N} S$ denotes the maximum number of calls to the stochastic oracle performed in the optimization phase of the above algorithms. For example, $\bar{N} S=1,000$ has the following implications.
- For the RSPG algorithm, the number of samples per iteration $m$ is computed according to (4.18) with $\bar{N}=1,000$ and the iteration limit $N$ is set to $\lfloor 1,000 / \mathrm{m}\rfloor$;
- For the 2-RSPG and 2-RSPG-V algorithms, since $S=5$, we set $\bar{N}=200$. The $m$ and $N$ are computed as mentioned above. In this case, total number of calls to the stochastic oracle will be at most 1,000 (this does not include the samples used in the post optimization phase);
- For the MD-SA method, after computing $m$ according to (4.18) with $\bar{N}=$ 1,000 , we run the RSPG method for $\lfloor 1,000 / \mathrm{m}\rfloor$ iterations and take the average of the iterates as the output.
$-\bar{x}^{*}$ is the output solution of the above algorithms.
- Mean and Var. represent, respectively, the average and variance of the results obtained over different runs of each algorithm.


### 6.1 Semi-supervised support vector machine problem

In the first experiment, we consider a binary classification problem. The training set is divided to two types of data, which consists of labeled and unlabeled examples, respectively. The linear semi-supervised support vector machine problem can be formulated as follows [7]:

$$
\begin{aligned}
\min _{b \in \mathbb{R}, x \in \mathbb{R}^{n}} \Psi(x, b):= & \lambda_{1} \mathbb{E}_{u_{1}, v}\left[\max \left\{0,1-v\left(\left\langle x, u_{1}\right\rangle+b\right)\right\}^{2}\right] \\
& +\lambda_{2} \mathbb{E}_{u_{2}}\left[\max \left\{0,1-\left|\left\langle x, u_{2}\right\rangle+b\right|\right\}^{2}\right]+\frac{\lambda_{3}}{2}\|x\|_{2}^{2},
\end{aligned}
$$

where $\left(u_{1}, v\right)$ and $u_{2}$ are labeled and unlabeled examples, respectively. Clearly, the above problem is nonsmooth, nonconvex, and does not fit the setting of the problem (1.1). Using a smooth approximation of the above problem [7], we can reformulate it as

$$
\begin{align*}
\min _{(x, b) \in \mathbb{R}^{n+1}} \Psi(x, b):= & \mathbb{E}_{u_{1}, u_{2}, v}\left[\lambda_{1} \max \left\{0,1-v\left(\left\langle x, u_{1}\right\rangle+b\right)\right\}^{2}+\lambda_{2} e^{-5\left\{\left\langle x, u_{2}\right\rangle+b\right\}^{2}}\right] \\
& +\frac{\lambda_{3}}{2}\|x\|_{2}^{2} . \tag{6.1}
\end{align*}
$$

Here, we assume that the feature vectors $u_{1}$ and $u_{2}$ are drawn from standard normal distribution with approximately $5 \%$ nonzero elements. Moreover, we assume that label $v \in\{-1,1\}$ with $v=\operatorname{sgn}\left(\left\langle\bar{x}, u_{1}\right\rangle+b\right)$ for some $\bar{x} \in \mathbb{R}^{n}$. The parameters are set to $\lambda_{1}=0.5, \quad \lambda_{2}=0.5$ and $\lambda_{3}=1$ and also three different problem sizes with $n=100,500$ and 1,000 are considered in this experiment.

We also want to determine the labels of unlabeled examples such that the ratio of new positive labels is close to that of the already labeled examples. It is shown in [7] that if the examples come from a distribution with zero mean, then, to have balanced new labels, we can consider the following constraint

$$
\begin{equation*}
|b-2 r+1| \leq \delta \tag{6.2}
\end{equation*}
$$

where $r$ is the ratio of positive labels in the already labeled examples and $\delta$ is a tolerance setting to 0.1 in our experiment. We also consider the $l_{2}$ regularization term as a simple convex term in the objective function i.e., $h(x)=\lambda_{3}\|x\|_{2}^{2} / 2$. Therefore, (6.1) together with the constraint (6.2) is a constrained nonconvex composite problem, which fits the setting of problem (1.1). Table 1 shows the mean and variance of the 2-norm of the projected gradient at the solutions obtained by 20 runs of the RSPG algorithms, and Fig. 1 gives the corresponding average objective values.

The following conclusions can be made from the numerical results. First, over 20 runs of the algorithm, the solutions of the RSPG algorithm have relatively large variance. Second, 2-RSPG, 2-RSPG-V can significantly reduce the variance of the RSPG algorithm for many instances. Third, for a given fixed $\bar{N} S$, the solution quality of the 2-RSPG-V algorithm is significantly better than that of the 2-RSPG algorithm when the problem size increases and $\bar{N} S$ is small. The possible reason is that, unlike the 2-RSPG algorithm, a candidate solution obtained in each run of the RSPG in the optimization phase of the 2-RSPG-V algorithm is used to generate the next candidate solution and hence the possibility of having better solution is increased. Finally, the solution quality of the 2-RSPG algorithm is much better than that of the MD-SA algorithm in almost all cases (see Table 1).

Table 1 Estimated $\left\|g_{X}\left(\bar{x}^{*}\right)\right\|^{2}$ for the semi-supervised support vector machine problem $(K=75,000)$

| $\bar{N} S$ |  | RSPG | 2 -RSPG | $2-R S P G-V$ | MD-SA |
| :--- | :--- | :--- | :--- | :--- | :--- |
| $n=100$ |  |  |  |  |  |
| 1,000 | Mean | 0.6364 | 0.0919 | 0.0824 | 0.2066 |
|  | Var. | $2.66 \mathrm{e}+000$ | $1.48 \mathrm{e}-004$ | $8.11 \mathrm{e}-005$ | $1.77 \mathrm{e}-004$ |
| 5,000 | Mean | 0.0490 | 0.0376 | 0.0368 | 0.0366 |
|  | Var. | $1.68 \mathrm{e}-002$ | $1.60 \mathrm{e}-005$ | $1.39 \mathrm{e}-005$ | $4.84 \mathrm{e}-006$ |
| 25,000 | Mean | 0.3665 | 0.0181 | 0.0173 | 0076 |
|  | Var. | $2.54 \mathrm{e}+000$ | $5.69 \mathrm{e}-006$ | $3.00 \mathrm{e}-006$ | $1.72 \mathrm{e}-007$ |
| $n=500$ |  |  |  |  |  |
| 1,000 | Mean | 10.0422 | 1.007 | 0.2001 | 1.9494 |
|  | Var. | $5.17 \mathrm{e}+002$ | $1.83 \mathrm{e}+000$ | $2.75 \mathrm{e}-004$ | $1.41 \mathrm{e}-003$ |
| 5,000 | Mean | 7.5432 | 0.0827 | 0.0777 | 0.3261 |
|  | Var. | $1.01 \mathrm{e}+003$ | $1.31 \mathrm{e}-004$ | $1.76 \mathrm{e}-005$ | $3.17 \mathrm{e}-005$ |
| 25,000 | Mean | 3.6799 | 0.0344 | 0.0339 | 0.0603 |
|  | Var. | $2.51 \mathrm{e}+002$ | $5.10 \mathrm{e}-006$ | $4.22 \mathrm{e}-006$ | $6.04 \mathrm{e}-007$ |
| $n=1,000$ |  |  |  |  |  |
| 1,000 | Mean | 55.7736 | 10.0097 | 0.4060 | 9.9998 |
|  | Var. | $6.33 \mathrm{e}+003$ | $8.41 \mathrm{e}+001$ | $7.60 \mathrm{e}-002$ | $5.23 \mathrm{e}-003$ |
| 5,000 | Mean | 7.1839 | 0.2753 | 0.1489 | 1.7826 |
|  | Var. | $9.60 \mathrm{e}+002$ | $4.32 \mathrm{e}-002$ | $2.98 \mathrm{e}-003$ | $6.71 \mathrm{e}-005$ |
| 25,000 | Mean | 1.0753 | 0.0633 | 0.0621 | 0.3286 |
|  | Var. | $2.13 \mathrm{e}+001$ | $4.98 \mathrm{e}-006$ | $4.33 \mathrm{e}-006$ | $6.77 \mathrm{e}-006$ |

### 6.2 Simulation-based inventory optimization problem

In the second experiment, we consider the classical $(s, S)$ inventory problem. More specifically, we consider the following simple case study in [20]. A Widgets company carries inventory of one product. Customers arrive according to Poisson distribution with mean 10 persons per day and they demand $1,2,3$, and 4 items of this product with probabilities $0.167,0.333,0.333$ and 0.167 , respectively, with back order permitted.

At the beginning of each day, the company checks the inventory level. If it is less than $s$, an order is placed to replenish the inventory up to $S$. Also, the lead time (the time between an order is placed and the ordered products arrive at the company) is distributed uniformly between 0.5 and 1 day. There is a fixed order cost of $\$ 32$ plus $\$ 3$ per item ordered. Also, a holding cost of $\$ 1$ per item per day and a shortage cost of $\$ 5$ per item per day are incurred. The company needs to choose $s$ and $S$ appropriately to minimize the total daily inventory cost. Since the inventory cost can only be evaluated by using simulation, we consider the following simulation-based optimization problem of

$$
\begin{equation*}
\left.\min _{100 \geq s \geq s \geq 0} \mathbb{E} \text { [daily inventory cost }\right] . \tag{6.3}
\end{equation*}
$$



Fig. 1 Average objective values at $\bar{x}^{*}$, obtained in Table 1. a $n=100, \mathbf{b} n=500, \mathbf{c} n=1,000$

We implement the RSPGF method as described in Sect. 5.
Moreover, we compute the value of the objective function in (6.3) by simulating the inventory system over 100 days. The other zeroth-order methods are implemented similarly to their corresponding first-order methods as described in the beginning of this section. Also, the smoothing parameter $\mu$ satisfying (5.23) is set to 0.05 for all these zeroth-order methods.

Table 2 Estimated gradients $\left\|\nabla f_{\mu}\left(\bar{x}^{*}\right)\right\|^{2}$ for the inventory problem ( $K=10,000$ )

| $x_{1}=\left(s_{1}, S_{1}\right)$ | $N S$ |  | RSPGF | 2-RSPGF | 2-RSPGF-V | MD-SA-GF |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| $(10,100)$ | 1,000 | Mean | 1.8179 | 0.0302 | 0.0306 | 2.2789 |
|  |  | Var. | $1.19 \mathrm{e}+001$ | $1.13 \mathrm{e}-003$ | $6.22 \mathrm{e}-004$ | $1.41 \mathrm{e}+001$ |
|  | 5,000 | Mean | 0.3167 | 0.0191 | 0.0187 | 0.0276 |
|  |  | Var. | $8.51 \mathrm{e}-001$ | $1.20 \mathrm{e}-004$ | $2.24 \mathrm{e}-004$ | $4.16 \mathrm{e}-004$ |
|  | 1,000 | mean | 4.7312 | 1.2958 | 0.7973 | 3.6128 |
|  |  | Var. | $5.26 \mathrm{e}+001$ | $1.64 \mathrm{e}+000$ | $2.54 \mathrm{e}-002$ | $2.99 \mathrm{e}+001$ |
|  | 5,000 | Mean | 3.9971 | 0.7562 | 0.7607 | 0.9055 |
|  |  | Var. | $4.68 \mathrm{e}+001$ | $1.68 \mathrm{e}-002$ | $1.99 \mathrm{e}-002$ | $1.75 \mathrm{e}-003$ |
|  | $10,50)$ | Mean | 2.5126 | 2.1447 | 1.1228 | 1.6830 |
|  |  | Var. | $1.21 \mathrm{e}+001$ | $8.66 \mathrm{e}+000$ | $1.86 \mathrm{e}-001$ | $1.45 \mathrm{e}+000$ |
|  | 5,000 | Mean | 2.2591 | 1.2505 | 0.7375 | 1.9244 |
|  |  | Var. | $1.09 \mathrm{e}+001$ | $3.93 \mathrm{e}-001$ | $2.96 \mathrm{e}-001$ | $4.28 \mathrm{e}+000$ |

Table 3 Average daily inventory costs

| $x_{1}=\left(s_{1}, S_{1}\right)$ | $N S$ | RSPGF | 2-RSPGF | 2-RSPGF-V | MD-SA-GF |
| :--- | :--- | :--- | :--- | :--- | :--- |
| $(10,100)$ | 1,000 | 129.71 | 129.31 | 129.38 | 129.3315 |
|  | 5,000 | 129.16 | 129.32 | 129.09 | 129.4647 |
| $(50,100)$ | 1,000 | 137.04 | 137.66 | 137.40 | 137.0039 |
|  | 5,000 | 136.56 | 137.10 | 135.11 | 136.7674 |
| $(10,50)$ | 1,000 | 126.78 | 126.51 | 125.51 | 125.8781 |
|  | 5,000 | 124.28 | 125.46 | 123.43 | 125.5707 |

Table 2 reports the mean and variance of $\left\|g_{\mu}\left(\bar{x}^{*}\right)\right\|^{2}$ over 10 runs of these methods with different initial solutions $x_{1}$. Similar to the results for the nonconvex semi-supervised support vector machine problem, the solution quality (in term of $\left\|g_{\mu}\left(\bar{x}^{*}\right)\right\|^{2}$ ) of the RSPGF method is not as good as the 2-RSPGF method and, for a given $\bar{N} S$, the 2-RSPGF-V method outperforms the 2-RSPGF method in many cases. Moreover, in most cases the 2-RPSGF method has much better performance than the MD-SA-GF method (i.e., the gradient free version of the MD-SA method described in beginning of this section). Table 3 shows the corresponding average daily inventory costs for the solutions, $\bar{x}^{*}$, computed by these algorithms. The best solution given by $(s, S)=(20.05,53.83)$ with an estimated average daily inventory cost $\$ 119.77$ has been obtained by running the 2-RSPG-V method starting from the initial solution $(10,50)$.

## 7 Conclusion

This paper proposes a new stochastic approximation algorithm with its variants for solving a class of nonconvex stochastic composite optimization problems. This new
randomized stochastic projected gradient (RSPG) algorithm uses mini-batch of samples at each iteration to handle the constraints. The proposed algorithm is set up in a way that a more general gradient projection according to the geometry of the constraint set could be used. The complexity bound of our algorithm is established in a unified way, including both convex and nonconvex objective functions. Our results show that the RSPG algorithm would automatically maintain a nearly optimal rate of convergence for solving stochastic convex programming problems. To reduce the variance of the RSPG algorithm, a two-phase RSPG algorithm is also proposed. It is shown that with a special post-optimization phase, the variance of the solutions returned by the RSPG algorithm could be significantly reduced, especially when a light tail condition holds. Based on this RSPG algorithm, a stochastic projected gradient free algorithm, which only uses the stochastic zeroth-order information, has been also proposed and analyzed. Our preliminary numerical results show that our two-phase RSPG algorithms, the 2-RSPG and its variant 2-RSPG-V algorithms, could be very effective and stable for solving the aforementioned nonconvex stochastic composite optimization problems.

It should be noted that in this paper we focus on the case when the regularization term $h$ in problem (1.1) is convex. In the future, it will be interesting to consider nonconvex and nonsmooth regularization terms in the objective function, especially due to the importance of these types of problems in a few application areas, such as signal processing.

## References

1. Andradóttir, S.: A review of simulation optimization techniques. In: Proceedings of the Winter Simulation Conference, pp. 151-158 (1998)
2. Auslender, A., Teboulle, M.: Interior gradient and proximal methods for convex and conic optimization. SIAM J. Optim. 16, 697-725 (2006)
3. Bauschke, H., Borwein, J., Combettes, P.: Bregman monotone optimization algorithms. SIAM J. Control Optim. 42, 596-636 (2003)
4. Ben-Tal, A., Margalit, T., Nemirovski, A.S.: The ordered subsets mirror descent optimization method with applications to tomography. SIAM J. Optim. 12, 79-108 (2001)
5. Bregman, L.: The relaxation method of finding the common point convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Phys. 7, 200-217 (1967)
6. Cartis, C., Gould, N.I.M., Toint, P.L.: On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization. SIAM J. Optim. 20(6), 2833-2852 (2010)
7. Chapelle, O., Sindhwani, V., Keerthi, S.S.: Optimization techniques for semi-supervised support vector machines. J. Mach. Learn. Res. 9, 203-233 (2008)
8. Dang, C.D., Lan, G.: On the Convergence Properties of Non-Euclidean Extragradient Methods for Variational Inequalities with Generalized Monotone Operators, manuscript, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, April 2012. Available on http://www.optimization-online.org/
9. Duchi, J.C., Bartlett, P.L., Wainwright, M.J.: Randomized smoothing for stochastic optimization. SIAM J. Optim. 22, 674-701 (2012)
10. Duchi, J.C., Shalev-shwartz, S., Singer, Y., Tewari, A.: Composite objective mirror descent. In: Proceedings of the Twenty Third Annual Conference on Computational Learning Theory (2010)
11. Flaxman, A.D., Kalai, A.T., McMahan, H.B.: Online convex optimization in the bandit setting: gradient descent without a gradient. J. Am. Stat. Assoc., 385-394 (2005). Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms
12. Fu, M.: Gradient estimation. In: Henderson, S.G., Nelson, B.L. (eds). Handbooks in Operations Research and Management Science: Simulation. Elsevier, Amsterdam, pp. 575-616 (2008)
13. Fu, M.: Optimization for simulation: theory vs. practice. INFORMS J. Comput. 14, 192-215 (2002)
14. 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)
15. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. SIAM J. Optim. 23, 20612089 (2013)
16. Ghadimi, S., Lan, G.: Accelerated Gradient Methods for Nonconvex Nonlinear and Stochastic Optimization. Technical Report. Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA (2013)
17. Glasserman, P.: Gradient Estimation Via Perturbation Analysis. Kluwer, Boston, MA (2003)
18. Juditsky, A., Nemirovski, A.S.: Large Deviations of Vector-Valued Martingales in 2-Smooth Normed Spaces, manuscript, Georgia Institute of Technology, Atlanta, GA, E-print: www2.isye.gatech.edu/ $\sim$ nemirovs/LargeDevSubmitted.pdf (2008)
19. Juditsky, A., Nemirovski, A.S.: First-order methods for nonsmooth convex large-scale optimization. I: general purpose methods. In: Sra, S., Nowozin, S., Wright, S.J. (eds.) Optimization for Machine Learning. MIT Press, Cambridge, MA (2011)
20. Kelton, R.P.S.W.D., Sturrock, D.T.: Simulation with Arena, 4th edn. McGraw-Hill, New York (2007)
21. Lan, G.: An optimal method for stochastic composite optimization. Math. Program. 133(1), 365-397 (2012)
22. Lan, G., Nemirovski, A.S., Shapiro, A.: Validation analysis of mirror descent stochastic approximation method. Math. Program. 134, 425-458 (2012)
23. LÉcuyer, P.: A unified view of the IPA, SF, and LR gradient estimation techniques. Manag. Sci. 36(11), 1364-1383 (1990)
24. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online dictionary learning for sparse coding. In: ICML, pp. 689-696 (2009)
25. Mason, L., Baxter, J., Bartlett, P., Frean, M.: Boosting algorithms as gradient descent in function space. Proc. NIPS 12, 512-518 (1999)
26. Nemirovski, A.S., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. 19, 1574-1609 (2009)
27. Nemirovski, A.S., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. WileyInterscience Series in Discrete Mathematics. John Wiley, Chichester, New York (1983)
28. Nesterov, Y.E.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer, Boston, MA (2004)
29. Nesterov, Y.E.: Random Gradient-Free Minimization of Convex Functions, Technical Report. Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (2010)
30. Polyak, B.: New stochastic approximation type procedures. Automat. i Telemekh. 7, 98-107 (1990)
31. Polyak, B., Juditsky, A.: Acceleration of stochastic approximation by averaging. SIAM J. Control Optim. 30, 838-855 (1992)
32. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. 22, 400-407 (1951)
33. Rockafellar, R.T., Wets, R.J.-B.: Variational analysis, ser. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer, Berlin (1998)
34. Rubinstein, R., Shapiro, A.: Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method. Wiley, New York (1993)
35. Schmidt, M., Roux, N.L., Bach, F.: Minimizing Finite Sums with the Stochastic Average Gradient, Technical Report (2013)
36. Spall, J.: Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control. Wiley, Hoboken, NJ (2003)
37. Sra, S.: Scalable nonconvex inexact proximal splitting. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 530-538. Curran Associates, Inc. (2012)
38. Teboulle, M.: Convergence of proximal-like algorithms. SIAM J. Optim. 7, 1069-1083 (1997)

[^0]:    This research was partially supported by NSF grants CMMI-1000347, CMMI-1254446, DMS-1319050, DMS-1016204 and ONR grant N00014-13-1-0036.
    S. Ghadimi • G. Lan ( $\triangle$ )

    Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA
    e-mail: glan@ise.ufl.edu
    URL: http://www.ise.ufl.edu/glan
    S. Ghadimi
    e-mail: sghadimi@ufl.edu
    H. Zhang

    Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803, USA
    e-mail: hozhang @math.lsu.edu
    URL: https://www.math.lsu.edu/~hozhang

