Noname manuscript No.
(will be inserted by the editor)

Mini-batch Stochastic Approximation Methods for Nonconvex Stochastic Composite Optimization

Saeed Ghadimi · Guanghui Lan · Hongchao Zhang

the date of receipt and acceptance should be inserted later

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.

keywords constrained stochastic programming, mini-batch of samples, stochastic approximation, nonconvex optimization, stochastic programming, first-order method, zeroth-order method

---

August, 2013. This research was partially supported by NSF grants CMMI-1000347, CMMI-1254446, DMS-1319050, DMS-1016204 and ONR grant N00014-13-1-0036.

sghadimi@ufl.edu, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611.

glan@ise.ufl.edu, http://www.ise.ufl.edu/glan, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611.

hozhang@math.lsu.edu , https://www.math.lsu.edu/∼hozhang, Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803.

Address(es) of author(s) should be given

1 Introduction

In this paper, we consider the following problem

$$\Psi^* := \min_{x \in X} \{\Psi(x) := f(x) + h(x)\}, \quad (1)$$

where X is a closed convex set in Euclidean space $\mathbb{R}^n$, $f : X \to \mathbb{R}$ 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.,

$$\|\nabla f(y) - \nabla f(x)\| \le L\|y - x\|, \quad \text{for any } x, y \in X, \quad (2)$$

and Ψ is bounded below over X, i.e. Ψ^* is finite. Although f is Lipschitz continuously differentiable, we assume that only the noisy gradient of f is available via subsequent calls to a *stochastic first-order oracle* ($\mathcal{SFO}$). Specifically, at the k-th call, $k \ge 1$, for the input $x_k \in X$, $\mathcal{SFO}$ would output a *stochastic gradient* $G(x_k, \xi_k)$, where ξ_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 $G(x_k, \xi_k)$.

A1: For any $k \ge 1$, we have

$$\text{a)} \quad \mathbb{E}[G(x_k, \xi_k)] = \nabla f(x_k), \quad (3)$$

$$\text{b)} \quad \mathbb{E}\left[\|G(x_k, \xi_k) - \nabla f(x_k)\|^2\right] \le \sigma^2, \quad (4)$$

where $\sigma > 0$ is a constant. For some examples which fit our setting, one may refer the problems in references [1,11,12,15,16,20–22,31].

Stochastic programming (SP) problems have been the subject of intense studies for more than 50 years. In the seminal 1951 paper, Robbins and Monro [29] 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., [32, Section 4.5.3]). Later, Polyak [27] and Polyak and Juditsky [28] 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., Ψ in (1) is a convex function). Motivated by the complexity theory in convex optimization [24], these studies focus on the convergence properties of SA-type algorithms in a finite number of iterations. For example, Nemirovski et al. [23] presented a mirror descent SA approach for solving general nonsmooth convex stochastic programming problems. They showed that the mirror descent SA exhibits an optimal $\mathcal{O}(1/\epsilon^2)$ iteration complexity for solving these problems with an essentially unimprovable constant factor. Also, Lan [19] presented a unified optimal method for smooth, nonsmooth and stochastic optimization. This unified optimal method also leads to optimal methods for strongly convex problems [13,14]. 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) 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 (4), the complexity of the gradient descent method for solving problem (1) has been studied in [6,25]. Very recently, Ghadimi and Lan [15] 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) with $h \equiv 0$ and $X = \mathbb{R}^n$. In their algorithm, a trajectory $\{x_1, \ldots, x_N\}$ 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 $\mathcal{SFO}$ required by this algorithm to find an ϵ-solution, i.e., a point $\bar{x}$ such that $\mathbb{E}[\|\nabla f(\bar{x})\|_2^2] \le \epsilon$, is bounded by $\mathcal{O}(\sigma^2/\epsilon^2)$. 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 (ϵ, Λ)-*solution*, i.e., a point $\bar{x}$ satisfying $\text{Prob}\{\|\nabla f(\bar{x})\|_2^2 \le \epsilon\} \ge 1 - \Lambda$, 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 $\mathcal{O}(n\sigma^2/\epsilon^2)$.

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) 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 $\mathcal{SFO}$ to find a solution $\bar{x} \in X$ such that $\mathbb{E}[\|g_X(\bar{x})\|^2] \le \epsilon$, is still $\mathcal{O}(\sigma^2/\epsilon^2)$, where $g_X(\bar{x})$ is a generalized projected gradient of Ψ 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 [15]. 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 $\mathcal{SFO}$. 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 have been 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 Section 2. In section 3, a deterministic first-order method for problem (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) in Section 4. In section 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 Section 6. Finally, in Section 7, we give some concluding remarks.

Notation. We use $\|\cdot\|$ to denote a general norm with associated inner product $\langle \cdot, \cdot \rangle$. For any $p \geq 1$, $\|\cdot\|_p$ denote the standard p-norm in $\mathbb{R}^n$, i.e.

$$\|x\|_p^p = \sum_{i=1}^n |x_i|^p, \qquad \text{for any } x \in \mathbb{R}^n.$$

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.,

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|, \qquad \text{for any } x, y \in \Omega, \tag{5}$$

which clearly implies

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|^2, \qquad \text{for any } x, y \in \Omega. \tag{6}$$

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 Subsection 2.1. Then, in Subsection 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 [2,3,5,19,23,33]. Hence, in this paper we would like to set up the projection based on the so-called prox-function.

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

$$\langle x - z, \nabla \omega(x) - \nabla \omega(z) \rangle \geq \alpha \|x - z\|^2, \ \ \forall x, z \in X. \tag{7}$$

Then, the *prox-function* associated with ω is defined as

$$V(x, z) = \omega(x) - [\omega(z) + \langle \nabla \omega(z), x - z \rangle]. \tag{8}$$

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

$$x^+ = \arg\min_{u \in X} \left\{ \langle g, u \rangle + \frac{1}{\gamma} V(u, x) + h(u) \right\} \tag{9}$$

is easily solvable for any $\gamma > 0$, $g \in \mathbb{R}^n$ and $x \in X$. Apparently, different choices of ω 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, x^+ is just the usual Euclidean projection. Some less trivial examples can be found, e.g., in [2,4,8,18,24].

2.2 Properties of Projection

In this subsection, we discuss some important properties of the generalized projection defined in (9). Let us first define

$$P_X(x, g, \gamma) = \frac{1}{\gamma}(x - x^+), \tag{10}$$

where x^+ is given in (9). We can see that $P_X(x, \nabla f(x), \gamma)$ can be viewed as a generalized projected gradient of Ψ 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 (9). Then, for any $x \in X$, $g \in \mathbb{R}^n$ and $\gamma > 0$, we have*

$$\langle g, P_X(x, g, \gamma)\rangle \geq \alpha \|P_X(x, g, \gamma)\|^2 + \frac{1}{\gamma}\left[h(x^+) - h(x)\right]. \tag{11}$$

Proof By the optimality condition of (9) and the definition of prox-function in (8), there exists a $p \in \partial h(x^+)$ such that

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

Letting $u = x$ in the above inequality, by the convexity of h and (7), we obtain

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

which in the view of (10) and $\gamma > 0$ clearly imply (11).

It is well-known [30] 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 (9) with g replaced by g_1 and g_2 respectively. Then,*

$$\|x_2^+ - x_1^+\| \leq \frac{\gamma}{\alpha}\|g_2 - g_1\|, \tag{12}$$

where $\alpha > 0$ is the modulus of strong convexity of ω defined in (7).

Proof By the optimality condition of (9), for any $u \in X$, there exist $p_1 \in \partial h(x_1^+)$ and $p_2 \in \partial h(x_2^+)$ such that

$$\langle g_1 + \frac{1}{\gamma}\left[\nabla\omega(x_1^+) - \nabla\omega(x)\right] + p_1, u - x_1^+\rangle \geq 0, \tag{13}$$

and

$$\langle g_2 + \frac{1}{\gamma}\left[\nabla\omega(x_2^+) - \nabla\omega(x)\right] + p_2, u - x_2^+\rangle \geq 0. \tag{14}$$

Letting $u = x_2^+$ in (13), by the convexity of h, we have

$$\begin{aligned}\langle g_1, x_2^+ - x_1^+\rangle &\geq \frac{1}{\gamma}\langle\nabla\omega(x) - \nabla\omega(x_1^+), x_2^+ - x_1^+\rangle + \langle p_1, x_1^+ - x_2^+\rangle \\ &\geq \frac{1}{\gamma}\langle\nabla\omega(x_2^+) - \nabla\omega(x_1^+), x_2^+ - x_1^+\rangle + \frac{1}{\gamma}\langle\nabla\omega(x) - \nabla\omega(x_2^+), x_2^+ - x_1^+\rangle \\ &\quad + h(x_1^+) - h(x_2^+).\end{aligned} \tag{15}$$

Similarly, letting $u = x_1^+$ in (14), we have

$$\begin{aligned}\langle g_2, x_1^+ - x_2^+\rangle &\geq \frac{1}{\gamma}\langle\nabla\omega(x) - \nabla\omega(x_2^+), x_1^+ - x_2^+\rangle + \langle p_2, x_2^+ - x_1^+\rangle \\ &\geq \frac{1}{\gamma}\langle\nabla\omega(x) - \nabla\omega(x_2^+), x_1^+ - x_2^+\rangle + h(x_2^+) - h(x_1^+).\end{aligned} \tag{16}$$

Summing up (15) and (16), by the strong convexity (7) of ω, we obtain

$$\|g_1 - g_2\|\|x_2^+ - x_1^+\| \geq \langle g_1 - g_2, x_2^+ - x_1^+\rangle \geq \frac{\alpha}{\gamma}\|x_2^+ - x_1^+\|^2,$$

which gives (12).

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 (10). Then, for any g_1 and g_2 in $\mathbb{R}^n$, we have*

$$\|P_X(x, g_1, \gamma) - P_X(x, g_2, \gamma)\| \leq \frac{1}{\alpha}\|g_1 - g_2\|, \tag{17}$$

where α is the modulus of strong convexity of ω defined in (7).

Proof Noticing (10), (13) and (14), we have

$$\|P_X(x, g_1, \gamma) - P_X(x, g_2, \gamma)\| = \|\frac{1}{\gamma}(x - x_1^+) - \frac{1}{\gamma}(x - x_2^+)\| = \frac{1}{\gamma}\|x_2^+ - x_1^+\| \leq \frac{1}{\alpha}\|g_1 - g_2\|,$$

where the last inequality follows from (12).

The following lemma (see, e.g., Lemma 1 of [19] and Lemma 2 of [13]) characterizes the solution of the generalized projection.

Lemma 3 *Let x^+ be given in (9). Then, for any $u \in X$, we have*

$$\langle g, x^+\rangle + h(x^+) + \frac{1}{\gamma}V(x^+, x) \leq \langle g, u\rangle + h(u) + \frac{1}{\gamma}[V(u, x) - V(u, x^+)]. \tag{18}$$

3 Deterministic first-order methods

In this section, we consider the problem (1) with $f \in \mathcal{C}_L^{1,1}(X)$, and for each input $x_k \in X$, we assume that the exact gradient $\nabla f(x_k)$ is available. Using the exact gradient information, we give a deterministic projected gradient (PG) algorithm for solving (1), which mainly provides a basis for us to develop the stochastic first-order algorithms in the next section.

A projected gradient (PG) algorithm

Input: Given initial point $x_1 \in X$, total number of iterations N, and the stepsizes $\{\gamma_k\}$ with $\gamma_k > 0$, $k \geq 1$.
Step $k = 1, \ldots, N$. Compute

$$x_{k+1} = \arg\min_{u \in X} \left\{ \langle \nabla f(x_k), u \rangle + \frac{1}{\gamma_k} V(u, x_k) + h(u) \right\}. \tag{19}$$

Output: $x_R \in \{x_k, \ldots, x_N\}$ such that

$$R = \arg\min_{k \in \{1,\ldots,N\}} \|g_{X,k}\|, \tag{20}$$

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

$$g_{X,k} = P_X(x_k, \nabla f(x_k), \gamma_k). \tag{21}$$

We can see that the above algorithm outputs the iterate with the minimum norm of the generalized projected gradient. In the above algorithm, we have not specified the selection of the stepsizes $\{\gamma_k\}$. We will return to this issue after establishing the following convergence results.

Theorem 1 *Suppose that the stepsizes $\{\gamma_k\}$ 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*

$$\|g_{X,R}\|^2 \leq \frac{L D_\Psi^2}{\sum_{k=1}^N (\alpha\gamma_k - L\gamma_k^2/2)}, \tag{22}$$

where

$$g_{X,R} = P_X(x_R, \nabla f(x_R), \gamma_R) \quad \textit{and} \quad D_\Psi := \left[\frac{(\Psi(x_1) - \Psi^*)}{L} \right]^{\frac{1}{2}}. \tag{23}$$

Proof Since $f \in \mathcal{C}_L^{1,1}(X)$, it follows from (6), (10), (19) and (21) that for any $k = 1, \ldots, N$, we have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= f(x_k) - \gamma_k \langle \nabla f(x_k), g_{X,k} \rangle + \frac{L}{2}\gamma_k^2 \|g_{X,k}\|^2. \end{aligned} \tag{24}$$

Then, by Lemma 1 with $x = x_k$, $\gamma = \gamma_k$ and $g = \nabla f(x_k)$, we obtain

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

which implies

$$\Psi(x_{k+1}) \leq \Psi(x_k) - \left(\alpha\gamma_k - \frac{L}{2}\gamma_k^2 \right) \|g_{X,k}\|^2. \tag{25}$$

Summing up the above inequalities for $k = 1, \ldots, N$, by (20) and $\gamma_k \le 2\alpha/L$, we have

$$\|g_{X,R}\|^2 \sum_{k=1}^{N} \left(\alpha\gamma_k - \frac{L}{2}\gamma_k^2\right) \le \sum_{k=1}^{N} \left(\alpha\gamma_k - \frac{L}{2}\gamma_k^2\right) \|g_{X,k}\|^2$$
$$\le \Psi(x_1) - \Psi(x_{k+1}) \le \Psi(x_1) - \Psi^*. \tag{26}$$

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 (22).

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*

$$\|g_{X,R}\|^2 \le \frac{2L^2 D_\Psi^2}{\alpha^2 N}. \tag{27}$$

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

$$\frac{L D_\Psi^2}{\sum_{k=1}^{N}(\alpha\gamma_k - L\gamma_k^2/2)} = \frac{2L^2 D_\Psi^2}{N\alpha^2}, \tag{28}$$

which together with (22), clearly imply (27).

4 Stochastic first-order methods

In this section, we consider problem (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 $\mathcal{SFO}$. In particular, given the k-th iteration $x_k \in X$ of our algorithm, the $\mathcal{SFO}$ will output the stochastic gradient $G(x_k, \xi_k)$, where ξ_k is a random vector whose distribution is supported on $\Xi_k \subseteq \mathbb{R}^d$. We assume the stochastic gradient $G(x_k, \xi_k)$ satisfies Assumption A1.

This section also consists of two subsections. In Subsection 4.1, we present a stochastic variant of the PG algorithm in Section 3 incorporated with a randomized stopping criterion, called the RSPG algorithm. Then, in Subsection 4.2, we describe a two phase RSPG algorithm, called the 2-RSPG algorithm, which can significantly reduce the large-deviations resulted from the RSPG algorithm.

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 [13,14,**?**,**?**,19]. 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 $\{\xi_k\}, k \ge 1$ could depend on the iterates $\{x_k\}$.

A randomized stochastic projected gradient (RSPG) algorithm

Input: Given initial point $x_1 \in X$, iteration limit N, the stepsizes $\{\gamma_k\}$ with $\gamma_k > 0$, $k \geq 1$, the batch sizes $\{m_k\}$ 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 $\mathcal{SFO}$ m_k times to obtain $G(x_k, \xi_{k,i})$, $i = 1, \ldots, m_k$, set

$$G_k = \frac{1}{m_k} \sum_{i=1}^{m_k} G(x_k, \xi_{k,i}), \tag{29}$$

and compute

$$x_{k+1} = \arg\min_{u \in X} \left\{ \langle G_k, u \rangle + \frac{1}{\gamma_k} V(u, x_k) + h(u) \right\}. \tag{30}$$

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 $\{\gamma_k\}$, the batch sizes $\{m_k\}$ 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 $\{\gamma_k\}$ 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 $k = 1, \ldots, N$,*

$$P_R(k) := \text{Prob}\{R = k\} = \frac{\alpha\gamma_k - L\gamma_k^2}{\sum_{k=1}^N (\alpha\gamma_k - L\gamma_k^2)}. \tag{31}$$

Then, under Assumption A1,

(a) for any $N \geq 1$, we have

$$\mathbb{E}[\|\tilde{g}_{X,R}\|^2] \leq \frac{LD_\Psi^2 + (\sigma^2/\alpha)\sum_{k=1}^N (\gamma_k/m_k)}{\sum_{k=1}^N (\alpha\gamma_k - L\gamma_k^2)}, \tag{32}$$

where the expectation is taken with respect to R and $\xi_{[N]} := (\xi_1, \ldots, \xi_N)$, D_Ψ is defined in (23), and the stochastic projected gradient

$$\tilde{g}_{X,k} := P_X(x_k, G_k, \gamma_k), \tag{33}$$

with P_X defined in(10);

(b) if, in addition, f in problem (1) is convex with an optimal solution x^*, and the stepsizes $\{\gamma_k\}$ are non-decreasing, i.e.,

$$0 \leq \gamma_1 \leq \gamma_2 \leq ... \leq \gamma_N \leq \frac{\alpha}{L}, \tag{34}$$

we have

$$\mathbb{E}\left[\Psi(x_R) - \Psi(x^*)\right] \leq \frac{(\alpha - L\gamma_1)V(x^*, x_1) + (\sigma^2/2)\sum_{k=1}^N (\gamma_k^2/m_k)}{\sum_{k=1}^N (\alpha\gamma_k - L\gamma_k^2)}, \tag{35}$$

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

$$\frac{\alpha}{L} \geq \gamma_1 \geq \gamma_2 \geq ... \geq \gamma_N \geq 0, \tag{36}$$

we have

$$\mathbb{E}\left[\Psi(x_R) - \Psi(x^*)\right] \leq \frac{(\alpha - L\gamma_N)\bar{V}(x^*) + (\sigma^2/2)\sum_{k=1}^N(\gamma_k^2/m_k)}{\sum_{k=1}^N(\alpha\gamma_k - L\gamma_k^2)}, \tag{37}$$

where $\bar{V}(x^*) := \max_{u \in X} V(x^*, u)$.

Proof Let $\delta_k \equiv G_k - \nabla f(x_k)$, $k \geq 1$. Since $f \in \mathcal{C}_L^{1,1}(X)$, it follows from (6), (10), (30) and (33) that, for any $k = 1, \ldots, N$, we have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= f(x_k) - \gamma_k \langle \nabla f(x_k), \tilde{g}_{X,k} \rangle + \frac{L}{2}\gamma_k^2 \|\tilde{g}_{X,k}\|^2 \\ &= f(x_k) - \gamma_k \langle G_k, \tilde{g}_{X,k} \rangle + \frac{L}{2}\gamma_k^2 \|\tilde{g}_{X,k}\|^2 + \gamma_k \langle \delta_k, \tilde{g}_{X,k} \rangle. \end{aligned} \tag{38}$$

So, by Lemma 1 with $x = x_k$, $\gamma = \gamma_k$ and $g = G_k$, we obtain

$$\begin{aligned} f(x_{k+1}) \leq f(x_k) - \Big[\alpha\gamma_k \|\tilde{g}_{X,k}\|^2 + h(x_{k+1}) - h(x_k)\Big] + \frac{L}{2}\gamma_k^2 \|\tilde{g}_{X,k}\|^2 \\ + \gamma_k \langle \delta_k, g_{X,k} \rangle + \gamma_k \langle \delta_k, \tilde{g}_{X,k} - g_{X,k} \rangle, \end{aligned}$$

where the projected gradient $g_{X,k}$ is defined in (21). Then, from the above inequality, (21) and (33), we obtain

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

where the last inequality follows from Proposition 1 with $x = x_k, \gamma = \gamma_k, g_1 = G_k$ and $g_2 = \nabla f(x_k)$. Summing up the above inequalities for $k = 1, \ldots, N$ and noticing that $\gamma_k \leq \alpha/L$, we obtain

$$\begin{aligned} \sum_{k=1}^N \left(\alpha\gamma_k - L\gamma_k^2\right)\|\tilde{g}_{X,k}\|^2 &\leq \sum_{k=1}^N \left(\alpha\gamma_k - \frac{L}{2}\gamma_k^2\right)\|\tilde{g}_{X,k}\|^2 \\ &\leq \Psi(x_1) - \Psi(x_{k+1}) + \sum_{k=1}^N \left\{\gamma_k \langle \delta_k, g_{X,k} \rangle + \frac{\gamma_k}{\alpha}\|\delta_k\|^2\right\} \\ &\leq \Psi(x_1) - \Psi^* + \sum_{k=1}^N \left\{\gamma_k \langle \delta_k, g_{X,k} \rangle + \frac{\gamma_k}{\alpha}\|\delta_k\|^2\right\}. \end{aligned} \tag{39}$$

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 $\mathbb{E}[\langle \delta_k, g_{X,k} \rangle | \xi_{[k-1]}] =$

0. In addition, denoting $S_j = \sum_{i=1}^{j} \delta_{k,i}$, and noting that $\mathbb{E}[\langle S_{i-1}, \delta_{k,i}\rangle | S_{i-1}] = 0$ for all $i = 1, \ldots, m_k$, we have

$$\mathbb{E}[\|S_{m_k}\|^2] = \mathbb{E}\left[\|S_{m_k-1}\|^2 + 2\langle S_{m_k-1}, \delta_{k,m_k}\rangle + \|\delta_{k,m_k}\|^2\right]$$
$$= \mathbb{E}[\|S_{m_k-1}\|^2] + \mathbb{E}[\|\delta_{k,m_k}\|^2] = \ldots = \sum_{i=1}^{m_k} \|\delta_{k,i}\|^2,$$

which, in view of (29) and Assumption A1.b), then implies that

$$\mathbb{E}[\|\delta_k\|^2] = \frac{1}{m_k^2}\mathbb{E}[\|S_{m_k}\|^2] = \frac{1}{m_k^2}\sum_{i=1}^{m_k}\mathbb{E}[\|\delta_{k,i}\|^2] \le \frac{\sigma^2}{m_k}. \tag{40}$$

With these observations, now taking expectations with respect to $\xi_{[N]}$ on both sides of (39), we get

$$\sum_{k=1}^{N}\left(\alpha\gamma_k - L\gamma_k^2\right)\mathbb{E}_{\xi_{[N]}}\|\tilde{g}_{X,k}\|^2 \le \Psi(x_1) - \Psi^* + (\sigma^2/\alpha)\sum_{k=1}^{N}(\gamma_k/m_k).$$

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

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

we have (32) 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

$$\langle G_k, x_{k+1}\rangle + h(x_{k+1}) + \frac{1}{\gamma_k}V(x_{k+1}, x_k) \le \langle G_k, x^*\rangle + h(x^*) + \frac{1}{\gamma_k}[V(x^*, x_k) - V(x^*, x_{k+1})],$$

which together with (6) and definition of δ_k give

$$f(x_{k+1}) + \langle \nabla f(x_k) + \delta_k, x_{k+1}\rangle + h(x_{k+1}) + \frac{1}{\gamma_k}V(x_{k+1}, x_k)$$
$$\le f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 + \langle \nabla f(x_k) + \delta_k, x^*\rangle + h(x^*)$$
$$+ \frac{1}{\gamma_k}[V(x^*, x_k) - V(x^*, x_{k+1})].$$

Simplifying the above inequality, we have

$$\Psi(x_{k+1}) \le f(x_k) + \langle \nabla f(x_k), x^* - x_k\rangle + h(x^*) + \langle \delta_k, x^* - x_{k+1}\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$
$$- \frac{1}{\gamma_k}V(x_{k+1}, x_k) + \frac{1}{\gamma_k}[V(x^*, x_k) - V(x^*, x_{k+1})].$$

Then, it follows from the convexity of f, (7) and (8) that

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

where the last inequality follows from Young's inequality. Noticing $\gamma_k \le \alpha/L$, multiplying both sides of the above inequality by $(\alpha\gamma_k - L\gamma_k^2)$ and summing them up for $k = 1, \ldots, N$, we obtain

$$\begin{aligned}
\sum_{k=1}^{N}(\alpha\gamma_k - L\gamma_k^2)\left[\Psi(x_{k+1}) - \Psi(x^*)\right] &\le \sum_{k=1}^{N}(\alpha\gamma_k - L\gamma_k^2)\langle \delta_k, x^* - x_k\rangle + \sum_{k=1}^{N}\frac{\gamma_k^2}{2}\|\delta_k\|^2 \\
&\quad + \sum_{k=1}^{N}(\alpha - L\gamma_k)\left[V(x^*, x_k) - V(x^*, x_{k+1})\right]. \quad (41)
\end{aligned}$$

Now, if the increasing stepsize condition (34) is satisfied, we have from $V(x^*, x_{N+1}) \ge 0$ that

$$\begin{aligned}
&\sum_{k=1}^{N}(\alpha - L\gamma_k)\left[V(x^*, x_k) - V(x^*, x_{k+1})\right] \\
&= (\alpha - L\gamma_1)V(x^*, x_1) + \sum_{k=2}^{N}(\alpha - L\gamma_k)V(x^*, x_k) - \sum_{k=1}^{N}(\alpha - L\gamma_k)V(x^*, x_{k+1}) \\
&\le (\alpha - L\gamma_1)V(x^*, x_1) + \sum_{k=2}^{N}(\alpha - L\gamma_{k-1})V(x^*, x_k) - \sum_{k=1}^{N}(\alpha - L\gamma_k)V(x^*, x_{k+1}) \\
&= (\alpha - L\gamma_1)V(x^*, x_1) - (\alpha - L\gamma_N)V(x^*, x_{N+1}) \\
&\le (\alpha - L\gamma_1)V(x^*, x_1).
\end{aligned}$$

Taking expectation on both sides of (41) with respect to $\xi_{[N]}$, again using the observations that $\mathbb{E}[\|\delta_k^2\|] \le \sigma^2$ and $\mathbb{E}[\langle \delta_k, g_{X,k}\rangle | \xi_{[k-1]}] = 0$, then it follows from the above inequality that

$$\sum_{k=1}^{N}(\alpha\gamma_k - L\gamma_k^2)\mathbb{E}_{\xi_{[N]}}\left[\Psi(x_{k+1}) - \Psi(x^*)\right] \le (\alpha - L\gamma_1)V(x^*, x_1) + \frac{\sigma^2}{2}\sum_{k=1}^{N}(\gamma_k^2/m_k).$$

Finally, (35) follows from the above inequality and the arguments similar to the proof in part (a). Now, if the decreasing stepsize condition (36) is satisfied, we have from the definition $\bar{V}(x^*) := \max_{u \in X} V(x^*, u) \geq 0$ and $V(x^*, x_{N+1}) \geq 0$ that

$$
\begin{aligned}
&\sum_{k=1}^{N}(\alpha - L\gamma_k)\left[V(x^*, x_k) - V(x^*, x_{k+1})\right] \\
&= (\alpha - L\gamma_1)V(x^*, x_1) + L\sum_{k=1}^{N-1}(\gamma_k - \gamma_{k+1})V(x^*, x_{k+1}) - (\alpha - L\gamma_N)V(x^*, x_{N+1}) \\
&\leq (\alpha - L\gamma_1)\bar{V}(x^*) + L\sum_{k=1}^{N-1}(\gamma_k - \gamma_{k+1})\bar{V}(x^*) - (\alpha - L\gamma_N)V(x^*, x_{N+1}) \\
&\leq (\alpha - L\gamma_N)\bar{V}(x^*),
\end{aligned}
$$

which together with (41) and similar arguments used above would give (37).

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 $\{\gamma_k\}$ (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 (35) or (37), and [23,19]). However, if f is possibly nonconvex and $m_k = 1$, then the RHS of (32) is bounded from below by

$$\frac{LD_\Psi^2 + (\sigma^2/\alpha)\sum_{k=1}^{N}\gamma_k}{\sum_{k=1}^{N}(\alpha\gamma_k - L\gamma_k^2)} \geq \frac{\sigma^2}{\alpha^2},$$

which does not necessarily guarantee the converge of the RSPG algorithm, no matter how the stepsizes $\{\gamma_k\}$ are specified. This is exactly the reason why we consider taking multiple samples $G(x_k, \xi_{k,i})$, $i = 1, \ldots, m_k$, for some $m_k > 1$ at each iteration of the RSPG method.

Secondly, from (39) in the proof of Theorem 2, we see that the stepsize policies can be further relaxed to get a similar result as (32). More specifically, we can have the following corollary.

Corollary 2 *Suppose that the stepsizes $\{\gamma_k\}$ 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 $k = 1, \ldots, N$*

$$P_R(k) := \text{Prob}\{R = k\} = \frac{\alpha\gamma_k - L\gamma_k^2/2}{\sum_{k=1}^{N}(\alpha\gamma_k - L\gamma_k^2/2)}. \tag{42}$$

Then, under Assumption A1, we have

$$\mathbb{E}[\|\tilde{g}_{X,R}\|^2] \leq \frac{LD_\Psi^2 + (\sigma^2/\alpha)\sum_{k=1}^{N}(\gamma_k/m_k)}{\sum_{k=1}^{N}(\alpha\gamma_k - L\gamma_k^2/2)}, \tag{43}$$

where the expectation is taken with respect to R and $\xi_{[N]} := (\xi_1, ..., \xi_N)$.

Based on the Theorem 2, we can establish the following complexity results of the RSPG algorithm with proper selection of stepsizes $\{\gamma_k\}$ and batch sizes $\{m_k\}$ at each iteration.

Corollary 3 *Suppose that in the RSPG algorithm the stepsizes $\gamma_k = \alpha/(2L)$ for all $k = 1, \ldots, N$, and the probability mass function P_R are chosen as (31). Also assume that the batch sizes $m_k = m$, $k = 1, \ldots, N$, for some $m \geq 1$. Then under Assumption A1, we have*

$$\mathbb{E}[\|g_{X,R}\|^2] \leq \frac{8L^2 D_\Psi^2}{\alpha^2 N} + \frac{6\sigma^2}{\alpha^2 m} \quad \textit{and} \quad \mathbb{E}[\|\tilde{g}_{X,R}\|^2] \leq \frac{4L^2 D_\Psi^2}{\alpha^2 N} + \frac{2\sigma^2}{\alpha^2 m}, \tag{44}$$

where $g_{X,R}$ and $\tilde{g}_{X,R}$ are defined in (21) and (33), respectively. If, in addition, f in the problem (1) is convex with an optimal solution x^*, then

$$\mathbb{E}\left[\Psi(x_R) - \Psi(x^*)\right] \leq \frac{2LV(x^*, x_1)}{N\alpha} + \frac{\sigma^2}{2Lm}. \tag{45}$$

Proof By (32), we have

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

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

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

Then, by Proposition 1 with $x = x_R, \gamma = \gamma_R, g_1 = \nabla f(x_R), g_2 = G_k$, we have from the above inequality and (40) that

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

Moreover, since $\gamma_k = \alpha/(2L)$ for all $k = 1, \ldots, N$, the stepsize conditions (34) are satisfied. Hence, if the problem is convex, (45) can be derived in a similar way as (35).

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 $\mathcal{SFO}$.

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

$$m = \left\lceil \min\left\{\max\left\{1, \frac{\sigma\sqrt{6\bar{N}}}{4L\tilde{D}}\right\}, \bar{N}\right\}\right\rceil, \tag{46}$$

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

$$\mathcal{B}_{\bar{N}} := \frac{16LD_\Psi^2}{\bar{N}} + \frac{4\sqrt{6}\sigma}{\sqrt{\bar{N}}}\left(\frac{D_\Psi^2}{\tilde{D}} + \tilde{D}\max\left\{1, \frac{\sqrt{6}\sigma}{4L\tilde{D}\sqrt{\bar{N}}}\right\}\right). \tag{47}$$

If, in addition, f *in problem (1) is convex, then* $\mathbb{E}[\Psi(x_R) - \Psi(x^*)] \le \mathcal{C}_{\bar{N}}$*, where* x^* *is an optimal solution and*

$$\mathcal{C}_{\bar{N}} := \frac{4LV(x^*, x_1)}{\alpha\bar{N}} + \frac{\sqrt{6}\sigma}{\alpha\sqrt{\bar{N}}}\left(\frac{V(x^*, x_1)}{\tilde{D}} + \frac{\alpha\tilde{D}}{3}\max\left\{1, \frac{\sqrt{6}\sigma}{4L\tilde{D}\sqrt{\bar{N}}}\right\}\right). \tag{48}$$

Proof Given the total number of calls to the stochastic first-order oracle $\bar{N}$ and the number m of calls to the $\mathcal{SFO}$ at each iteration, the RSPG algorithm can perform at most $N = \lfloor \bar{N}/m \rfloor$ iterations. Obviously, $N \ge \bar{N}/(2m)$. With this observation and (44), we have

$$\begin{aligned}
\mathbb{E}[\|g_{X,R}\|^2] &\le \frac{16mL^2D_\Psi^2}{\alpha^2\bar{N}} + \frac{6\sigma^2}{\alpha^2 m} \\
&\le \frac{16L^2D_\Psi^2}{\alpha^2\bar{N}}\left(1 + \frac{\sigma\sqrt{6\bar{N}}}{4L\tilde{D}}\right) + \max\left\{\frac{4\sqrt{6}L\tilde{D}\sigma}{\alpha^2\sqrt{\bar{N}}}, \frac{6\sigma^2}{\alpha^2\bar{N}}\right\} \\
&= \frac{16L^2D_\Psi^2}{\alpha^2\bar{N}} + \frac{4\sqrt{6}L\sigma}{\alpha^2\sqrt{\bar{N}}}\left(\frac{D_\Psi^2}{\tilde{D}} + \tilde{D}\max\left\{1, \frac{\sqrt{6}\sigma}{4L\tilde{D}\sqrt{\bar{N}}}\right\}\right),
\end{aligned} \tag{49}$$

which gives (47). The bound (48) 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 $\sigma_k^2 := \mathbb{E}\left[\|G(x_k, \xi_k) - \nabla f(x_k)\|^2\right]$. Secondly, we need to specify the parameter $\tilde{D}$ in (46). It can be seen from (47) and (48) that when $\bar{N}$ is relatively large such that

$$\max\left\{1, \sqrt{6}\sigma/(4L\tilde{D}\sqrt{\bar{N}})\right\} = 1, \quad \text{i.e.,} \quad \bar{N} \ge 3\sigma^2/(8L^2\tilde{D}^2), \tag{50}$$

an optimal choice of $\tilde{D}$ would be D_Ψ and $\sqrt{3V(x^*, x_1)/\alpha}$ for solving nonconvex and convex SP problems, respectively. With this selection of $\tilde{D}$, the bounds in (47) and (48), respectively, reduce to

$$\frac{\alpha^2}{L}\mathbb{E}[\|g_{X,R}\|^2] \le \frac{16LD_\Psi^2}{\bar{N}} + \frac{8\sqrt{6}D_\Psi\sigma}{\sqrt{\bar{N}}} \tag{51}$$

and

$$\mathbb{E}[\Psi(x^*) - \Psi(x_1)] \le \frac{4LV(x^*, x_1)}{\alpha\bar{N}} + \frac{2\sqrt{2V(x^*, x_1)}\sigma}{\sqrt{\alpha\bar{N}}}. \tag{52}$$

Thirdly, the stepsize policy in Corollary 3 and the probability mass function (31) together with the number of samples (46) 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 (52) is unimprovable (see e.g., [24]), while the first term in (52) can be considerably improved [19].

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 (ϵ, Λ)-*solution* of the problem (1), i.e., a point $x \in X$ satisfying $\text{Prob}\{\|g_X(x)\|^2 \le \epsilon\} \ge 1 - \Lambda$, for some $\epsilon > 0$ and $\Lambda \in (0, 1)$. Noticing that by the Markov's inequality and (47), we can directly have

$$\text{Prob}\left\{\|g_{X,R}\|^2 \ge \frac{\lambda L \mathcal{B}_{\bar{N}}}{\alpha^2}\right\} \le \frac{1}{\lambda}, \qquad \text{for any } \lambda > 0. \tag{53}$$

This implies that the total number of calls to the $\mathcal{SFO}$ performed by the RSPG algorithm for finding an (ϵ, Λ)-*solution*, after disregarding a few constant factors, can be bounded by

$$\mathcal{O}\left\{\frac{1}{\Lambda\epsilon} + \frac{\sigma^2}{\Lambda^2\epsilon^2}\right\}. \tag{54}$$

In this subsection, we present a approach to improve the dependence of the above bound on Λ. 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 $\mathcal{SFO}$ 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 (46), stepsizes $\gamma_k = \alpha/(2L)$ for $k = 1, \ldots, N$, batch sizes $m_k = m$, and probability mass function P_R in (31).

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 $\{\bar{x}_1, \ldots, \bar{x}_S\}$ such that

$$\|\bar{g}_X(\bar{x}^*)\| = \min_{s=1,\ldots,S} \|\bar{g}_X(\bar{x}_s)\|, \quad \bar{g}_X(\bar{x}_s) := P_X(\bar{x}_s, \bar{G}_T(\bar{x}_s), \gamma_{R_s}), \tag{55}$$

where $\bar{G}_T(x) = \frac{1}{T}\sum_{k=1}^T G(x, \xi_k)$ and $P_X(x, g, \gamma)$ is defined in (10).

Output: $\bar{x}^*$.

In the 2-RSPG algorithm, the total number of calls of $\mathcal{SFO}$ 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, $\bar{N}$ and T for finding an (ϵ, Λ)-*solution* of problem (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 [17] for a general result using possibly non-Euclidean norm).

Lemma 4 *Assume that we are given a polish space with Borel probability measure μ and a sequence of $\mathcal{F}_0 = \{\emptyset, \Omega\} \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots$ of σ-sub-algebras of Borel σ-algebra of Ω. Let $\zeta_i \in \mathbb{R}^n$, $i = 1, \dots, \infty$, be a martingale-difference sequence of Borel functions on Ω such that ζ_i is $\mathcal{F}_i$ measurable and $\mathbb{E}[\zeta_i|i-1] = 0$, where $\mathbb{E}[\cdot|i]$, $i = 1, 2, \dots$, denotes the conditional expectation w.r.t. $\mathcal{F}_i$ and $\mathbb{E} \equiv \mathbb{E}[\cdot|0]$ is the expectation w.r.t. μ.*

a) If $\mathbb{E}[\|\zeta_i\|^2] \le \sigma_i^2$ for any $i \ge 1$, then $\mathbb{E}[\|\sum_{i=1}^N \zeta_i\|^2] \le \sum_{i=1}^N \sigma_i^2$. As a consequence, we have

$$\forall N \ge 1, \lambda \ge 0 : \text{Prob}\left\{ \| \sum_{i=1}^N \zeta_i \|^2 \ge \lambda \sum_{i=1}^N \sigma_i^2 \right\} \le \frac{1}{\lambda};$$

b) If $\mathbb{E}\left[\exp\left(\|\zeta_i\|^2/\sigma_i^2\right)|i-1\right] \le \exp(1)$ almost surely for any $i \ge 1$, then

$$\forall N \ge 1, \lambda \ge 0 : \text{Prob}\left\{ \| \sum_{i=1}^N \zeta_i \| \ge \sqrt{2}(1+\lambda)\sqrt{\sum_{i=1}^N \sigma_i^2} \right\} \le \exp(-\lambda^2/3).$$

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).*

(a) Let $\mathcal{B}_{\bar{N}}$ be defined in (47). Then, for all $\lambda > 0$

$$\text{Prob}\left\{ \|g_X(\bar{x}^*)\|^2 \ge \frac{2}{\alpha^2}\left(4L\mathcal{B}_{\bar{N}} + \frac{3\lambda\sigma^2}{T}\right) \right\} \le \frac{S}{\lambda} + 2^{-S}; \tag{56}$$

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

$$S(\Lambda) := \lceil \log_2(2/\Lambda) \rceil, \tag{57}$$

$$\bar{N}(\epsilon) := \left\lceil \max\left\{ \frac{512L^2 D_\Psi^2}{\alpha^2 \epsilon}, \left[\left(\tilde{D} + \frac{D_\Psi^2}{\tilde{D}}\right)\frac{128\sqrt{6}L\sigma}{\alpha^2\epsilon}\right]^2, \frac{3\sigma^2}{8L^2\tilde{D}^2} \right\} \right\rceil, \tag{58}$$

$$T(\epsilon, \Lambda) := \left\lceil \frac{24 S(\Lambda)\sigma^2}{\alpha^2 \Lambda \epsilon} \right\rceil, \tag{59}$$

then the 2-RSPG algorithm computes an (ϵ, Λ)-solution of the problem (1) after taking at most

$$S(\Lambda)\ \left[\bar{N}(\epsilon) + T(\epsilon, \Lambda)\right] \tag{60}$$

calls of the stochastic first order oracle.

Proof We first show part (a). Let $g_X(\bar{x}_s) = P_X(\bar{x}_s, \nabla f(\bar{x}_s), \gamma_{R_s})$. Then, it follows from the definition of $\bar{x}^*$ in (55) that

$$\begin{aligned}
\|\bar{g}_X(\bar{x}^*)\|^2 &= \min_{s=1,\dots,S} \|\bar{g}_X(\bar{x}_s)\|^2 = \min_{s=1,\dots,S} \|g_X(\bar{x}_s) + \bar{g}_X(\bar{x}_s) - g_X(\bar{x}_s)\|^2 \\
&\le \min_{s=1,\dots,S} \left\{ 2\|g_X(\bar{x}_s)\|^2 + 2\|\bar{g}_X(\bar{x}_s) - g_X(\bar{x}_s)\|^2 \right\} \\
&\le 2 \min_{s=1,\dots,S} \|g_X(\bar{x}_s)\|^2 + 2 \max_{s=1,\dots,S} \|\bar{g}_X(\bar{x}_s) - g_X(\bar{x}_s)\|^2,
\end{aligned}$$

which implies that

$$\begin{aligned}\|g_X(\bar{x}^*)\|^2 &\le 2\|\bar{g}_X(\bar{x}^*)\|^2 + 2\|g_X(\bar{x}^*) - \bar{g}_X(\bar{x}^*)\|^2 \\ &\le 4 \min_{s=1,\ldots,S} \|g_X(\bar{x}_s)\|^2 + 4 \max_{s=1,\ldots,S} \|\bar{g}_X(\bar{x}_s) - g_X(\bar{x}_s)\|^2 + 2\|g_X(\bar{x}^*) - \bar{g}_X(\bar{x}^*)\|^2 \\ &\le 4 \min_{s=1,\ldots,S} \|g_X(\bar{x}_s)\|^2 + 6 \max_{s=1,\ldots,S} \|\bar{g}_X(\bar{x}_s) - g_X(\bar{x}_s)\|^2. \end{aligned} \tag{61}$$

We now provide certain probabilistic bounds to the two terms in the right hand side of the above inequality. Firstly, from the fact that $\bar{x}_s$, $1 \le s \le S$, are independent and (53) (with $\lambda = 2$), we have

$$\mathrm{Prob}\left\{\min_{s\in[1,S]} \|g_X(\bar{x}_s)\|^2 \ge \frac{2L\mathcal{B}_{\bar{N}}}{\alpha^2}\right\} = \prod_{s=1}^{S} \mathrm{Prob}\left\{\|g_X(\bar{x}_s)\|^2 \ge \frac{2L\mathcal{B}_{\bar{N}}}{\alpha^2}\right\} \le 2^{-S}. \tag{62}$$

Moreover, denoting $\delta_{s,k} = G(\bar{x}_s, \xi_k) - \nabla f(\bar{x}_s)$, $k = 1, \ldots, T$, by Proposition 1 with $x = \bar{x}_s, \gamma = \gamma_{R_s}, g_1 = \bar{G}_T(\bar{x}_s), g_2 = \nabla f(\bar{x}_s)$, we have

$$\|\bar{g}_X(\bar{x}_s) - g_X(\bar{x}_s)\| \le \frac{1}{\alpha}\|\sum_{k=1}^{T} \delta_{s,k}/T\|. \tag{63}$$

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

$$\mathrm{Prob}\left\{\|\bar{g}_X(\bar{x}_s) - g_X(\bar{x}_s)\|^2 \ge \frac{\lambda\sigma^2}{\alpha^2 T}\right\} \le \mathrm{Prob}\left\{\|\sum_{k=1}^{T}\delta_{s,k}\|^2 \ge \lambda T\sigma^2\right\} \le \frac{1}{\lambda},$$

which implies

$$\mathrm{Prob}\left\{\max_{s=1,\ldots,S} \|\bar{g}_X(\bar{x}_s) - g_X(\bar{x}_s)\|^2 \ge \frac{\lambda\sigma^2}{\alpha^2 T}\right\} \le \frac{S}{\lambda}. \tag{64}$$

Then, the conclusion (56) follows from (61), (62) and (64).

We now show part (b). With the settings in part (b), it is easy to count the total number of calls of the $\mathcal{SFO}$ in the 2-RSPG algorithm is bounded up by (60). Hence, we only need to show that the $\bar{x}^*$ returned by the 2-RSPG algorithm is indeed an (ϵ, Λ)-solution of the problem (1). With the choice of $\bar{N}(\epsilon)$ in (58), we can see that (50) holds. So, we have from (47) and (58) that

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

By the above inequality and (59), setting $\lambda = 2S/\Lambda$ in (56), we have

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

which together with (56), (57) and $\lambda = 2S/\Lambda$ imply

$$\mathrm{Prob}\left\{\|g_X(\bar{x}^*)\|^2 \ge \epsilon\right\} \le \frac{\Lambda}{2} + 2^{-S} \le \Lambda.$$

Hence,$\bar{x}^*$ is an (ϵ, Λ)-solution of the problem (1).

Now, it is interesting to compare the complexity bound in (60) with the one in (54). In view of (57), (58) and (59), the complexity bound in (60) for finding an (ϵ, Λ)-solution, after discarding a few constant factors, is equivalent to

$$\mathcal{O}\left\{\frac{1}{\epsilon}\log_2\frac{1}{\Lambda}+\frac{\sigma^2}{\epsilon^2}\log_2\frac{1}{\Lambda}+\frac{\sigma^2}{\Lambda\epsilon}\log_2^2\frac{1}{\Lambda}\right\}. \tag{65}$$

When the second terms are the dominating terms in both bounds, the above bound (65) can be considerably smaller than the one in (54) 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

$$\mathbb{E}[\exp\{\|G(x_k,\xi_k)-\nabla f(x)\|^2/\sigma^2\}] \le \exp\{1\}, \tag{66}$$

the bound (60) 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).*

(a) Let $\mathcal{B}_{\bar N}$ *is defined in (47). Then, for all* $\lambda > 0$

$$\text{Prob}\left\{\|g_X(\bar x^*)\|^2 \ge \left[\frac{8L\mathcal{B}_{\bar N}}{\alpha^2}+\frac{12(1+\lambda)^2\sigma^2}{T\alpha^2}\right]\right\} \le S\exp(-\frac{\lambda^2}{3})+2^{-S}; \tag{67}$$

(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 (57) and (58), respectively, and the sample size* T *is set to*

$$T'(\epsilon,\Lambda) := \frac{24\sigma^2}{\alpha^2\epsilon}\left[1+\left(3\log_2\frac{2S(\Lambda)}{\Lambda}\right)^{\frac{1}{2}}\right]^2, \tag{68}$$

then the 2*-RSPG algorithm can compute an* (ϵ,Λ)*-solution of the problem (1) after taking at most*

$$S(\Lambda)\left[\bar N(\epsilon)+T'(\epsilon,\Lambda)\right] \tag{69}$$

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 $\delta_{s,k} = G(\bar x_s, \xi_k) - \nabla f(\bar x_s)$, $k = 1,\ldots,T$, again by Proposition 1, we have (63) holds. Then, by Assumption A2 and Lemma 4.b), for any $\lambda > 0$ and any $s = 1,\ldots,S$, we have

$$\begin{aligned}&\text{Prob}\left\{\|\bar g_X(\bar x_s)-g_X(\bar x_s)\|^2 \ge (1+\lambda)^2\frac{2\sigma^2}{\alpha^2 T}\right\}\\ &\le \text{Prob}\left\{\|\sum_{k=1}^{T}\delta_{s,k}\| \ge \sqrt{2T}(1+\lambda)\sigma\right\} \le \exp(-\frac{\lambda^2}{3}),\end{aligned}$$

which implies that for any $\lambda > 0$

$$\text{Prob}\left\{\max_{s=1,\ldots,S}\|\bar g_X(\bar x_s)-g_X(\bar x_s)\|^2 \ge (1+\lambda)^2\frac{2\sigma^2}{\alpha^2 T}\right\} \le S\exp(-\frac{\lambda^2}{3}), \tag{70}$$

Then, the conclusion (67) follows from (61), (62) and (70).

In view of (57), (58) and (68), the bound in (69), after discarding a few constant factors, is equivalent to

$$\mathcal{O}\left\{\frac{1}{\epsilon}\log_2\frac{1}{\Lambda}+\frac{\sigma^2}{\epsilon^2}\log_2\frac{1}{\Lambda}+\frac{\sigma^2}{\epsilon}\log_2^2\frac{1}{\Lambda}\right\}. \tag{71}$$

Clearly, the third term of the above bound is smaller than the third term in (65) by a factor of $1/\Lambda$.

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) are available. More specifically, we assume that we can only access the noisy zeroth-order information of f by a *stochastic zeroth-order oracle* ($\mathcal{SZO}$). For any input x_k and ξ_k, the $\mathcal{SZO}$ would output a quantity $F(x_k,\xi_k)$, where x_k is the k-th iterate of our algorithm and ξ_k is a random variable whose distribution is supported on $\Xi \in \mathbb{R}^d$ (noting that Ξ does not depend on x_k). Throughout this section, we assume $F(x_k,\xi_k)$ is an unbiased estimator of $f(x_k)$, that is

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

$$\mathbb{E}[F(x_k,\xi_k)] = f(x_k). \tag{72}$$

We are going to apply the randomized smoothing techniques developed by [9,26] to explore the zeroth-order information of f. Hence, throughout this section, we also assume $F(\cdot,\xi_k)\in\mathcal{C}_L^{1,1}(\mathbb{R}^n)$ almost surely with respect to $\xi_k\in\Xi$, which together with Assumption A3 imply $f\in\mathcal{C}_L^{1,1}(\mathbb{R}^n)$. 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 ρ, a smooth approximation of f is defined as

$$f_\mu(x) = \int f(x+\mu v)\rho(v)dv, \tag{73}$$

where $\mu > 0$ is the smoothing parameter. For different choices of smoothing distribution, the smoothed function f_μ 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

$$f_\mu(x) = \frac{1}{(2\pi)^{\frac{n}{2}}}\int f(x+\mu v)e^{-\frac{1}{2}\|v\|^2}\,dv = \mathbb{E}_v[f(x+\mu v)]. \tag{74}$$

Nesterov [26] showed that the Gaussian smoothing approximation and f_μ have the following nice properties.

Lemma 5 *If $f\in\mathcal{C}_L^{1,1}(\mathbb{R}^n)$, then*

a) f_μ is also Lipschitz continuously differentiable with gradient Lipschitz constant $L_\mu \leq L$ and

$$\nabla f_\mu(x) = \frac{1}{(2\pi)^{\frac{n}{2}}}\int \frac{f(x+\mu v)-f(x)}{\mu}ve^{-\frac{1}{2}\|v\|^2}\,dv. \tag{75}$$

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

$$|f_\mu(x) - f(x)| \le \frac{\mu^2}{2} Ln, \tag{76}$$

$$\|\nabla f_\mu(x) - \nabla f(x)\| \le \frac{\mu}{2} L(n+3)^{\frac{3}{2}}, \tag{77}$$

$$\mathbb{E}_v\left[\left\|\frac{f(x+\mu v) - f(x)}{\mu} v\right\|^2\right] \le 2(n+4)\|\nabla f(x)\|^2 + \frac{\mu^2}{2} L^2 (n+6)^3. \tag{78}$$

c) f_μ is also convex provided f is convex.

In the following, let us define the approximated stochastic gradient of f at x_k as

$$G_\mu(x_k, \xi_k, v) = \frac{F(x_k + \mu v, \xi_k) - F(x_k, \xi_k)}{\mu} v, \tag{79}$$

and define $G(x_k, \xi_k) = \nabla_x F(x_k, \xi_k)$. We assume the Assumption 1 holds for $G(x_k, \xi_k)$. Then, by the Assumption A3 and Lemma 5.a), we directly get

$$\mathbb{E}_{v,\xi_k}[G_\mu(x_k, \xi_k, v)] = \nabla f_\mu(x_k), \tag{80}$$

where the expectation is taken with respect to v and ξ_k.

Now based on the RSPG algorithm, we state an algorithm which only uses zeroth-order information to solve problem (1).

A randomized stochastic projected gradient free (RSPGF) algorithm

Input: Given initial point $x_1 \in X$, iteration limit N, the stepsizes $\{\gamma_k\}$ with $\gamma_k > 0$, $k \ge 1$, the batch sizes $\{m_k\}$ with $m_k > 0$, $k \ge 1$, 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 $\mathcal{SZO}$ m_k times to obtain $G_\mu(x_k, \xi_{k,i}, v_{k,i})$, $i = 1, \ldots, m_k$, set

$$G_{\mu,k} = \frac{1}{m_k} \sum_{i=1}^{m_k} G_\mu(x_k, \xi_{k,i}, v_{k,i}) \tag{81}$$

and compute

$$x_{k+1} = \arg\min_{u \in X} \left\{ \langle G_{\mu,k}, u \rangle + \frac{1}{\gamma_k} V(u, x_k) + h(u) \right\}. \tag{82}$$

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 (80), $G_{\mu,k}$ can be simply viewed as an unbiased stochastic gradient of the smoothed function f_μ. 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 μ plays and the proper selection of μ 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 $\{\gamma_k\}$ in the RSPGF algorithm are chosen such that $0 < \gamma_k \le \alpha/L$ with $\gamma_k < \alpha/L$ for at least one k, and the probability mass function P_R are chosen as (31). If $\|\nabla f(x)\| \le M$ for all $x \in X$, then under Assumptions A1 and A3,*

(a) for any $N \ge 1$, we have

$$\mathbb{E}[\|\bar{g}_{\mu,X,R}\|^2] \le \frac{LD_\Psi^2 + \mu^2 Ln + (\tilde{\sigma}^2/\alpha)\sum_{k=1}^N (\gamma_k/m_k)}{\sum_{k=1}^N (\alpha\gamma_k - L\gamma_k^2)}, \tag{83}$$

where the expectation is taken with respect to R, $\xi_{[N]}$ and $v_{[N]} := (v_1, \ldots, v_N)$, D_Ψ is defined in (23),

$$\tilde{\sigma}^2 = 2(n+4)[M^2 + \sigma^2 + \mu^2 L^2 (n+4)^2], \tag{84}$$

and

$$\bar{g}_{\mu,X,k} = P_X(x_k, G_{\mu,k}, \gamma_k), \tag{85}$$

with P_X defined in(10);

(b) if, in addition, f in problem (1) is convex with an optimal solution x^*, and the stepsizes $\{\gamma_k\}$ are non-decreasing as (34), we have

$$\mathbb{E}\left[\Psi(x_R) - \Psi(x^*)\right] \le \frac{(\alpha - L\gamma_1)V(x^*, x_1) + (\tilde{\sigma}^2/2)\sum_{k=1}^N (\gamma_k^2/m_k)}{\sum_{k=1}^N (\alpha\gamma_k - L\gamma_k^2)} + \mu^2 Ln, \tag{86}$$

where the expectation is taken with respect to R, $\xi_{[N]}$ and $v_{[N]}$.

Proof By our assumption that $F(\cdot, \xi_k) \in \mathcal{C}_L^{1,1}(\mathbb{R}^n)$ almost surely, (78) (applying $f = F(\cdot, \xi_k)$) and Assumption 1 with $G(x_k, \xi_k) = \nabla_x F(x_k, \xi_k)$, we have

$$\begin{aligned} \mathbb{E}_{v_k,\xi_k}[\|G_\mu(x_k, \xi_k, v_k)\|^2] &= \mathbb{E}_{\xi_k}[\mathbb{E}_{v_k}[\|G_\mu(x_k, \xi_k, v_k)\|^2]] \\ &\le 2(n+4)[\mathbb{E}_{\xi_k}[\|G(x_k, \xi)\|^2] + \frac{\mu^2}{2} L^2 (n+6)^3 \\ &\le 2(n+4)[\mathbb{E}_{\xi_k}[\|\nabla f(x_k)\|^2] + \sigma^2] + 2\mu^2 L^2 (n+4)^3. \end{aligned}$$

Then, from the above inequality, (80) and $\|\nabla f(x_k)\| \le M$, we have

$$\begin{aligned} &\mathbb{E}_{v_k,\xi_k}[\|G_\mu(x_k, \xi_k, v_k) - \nabla f_\mu(x_k)\|^2] = \mathbb{E}_{v_k,\xi_k}[\|G_\mu(x_k, \xi_k, v_k)\|^2] \\ &\le 2(n+4)[M^2 + \sigma^2 + \mu^2 L^2 (n+4)^2] = \tilde{\sigma}^2. \end{aligned} \tag{87}$$

Now let $\Psi_\mu(x) = f_\mu(x) + h(x)$ and $\Psi_\mu^* = \min_{x \in X} \Psi_\mu(x)$. We have from (76) that

$$|(\Psi_\mu(x) - \Psi_\mu^*) - (\Psi(x) - \Psi^*)| \le \mu^2 Ln. \tag{88}$$

By Lemma (5).a), we have $L_\mu \le L$ and therefore $f_\mu \in \mathcal{C}_L^{1,1}(\mathbb{R}^n)$. With this observation, noticing (80) and (87), viewing $G_\mu(x_k, \xi_k, v_k)$ as a stochastic gradient of f_μ, then by part (a) of Theorem 2 we can directly get

$$\mathbb{E}[\|\bar{g}_{\mu,X,R}\|^2] \le \frac{LD_{\Psi_\mu}^2 + (\tilde{\sigma}^2/\alpha)\sum_{k=1}^N (\gamma_k/m_k)}{\sum_{k=1}^N (\alpha\gamma_k - L\gamma_k^2)},$$

where $D_{\Psi_\mu} = [(\Psi_\mu(x_1) - \Psi_\mu^*)/L]^{1/2}$ and the expectation is taken with respect to R, $\xi_{[N]}$ and $v_{[N]}$. Then, the conclusion (83) follows the above inequality and (88).

We now show part (b). Since f is convex, by Lemma (5).c), f_μ is also convex. Again by (88), we have

$$\mathbb{E}\left[\Psi(x_R) - \Psi(x^*)\right] \le \mathbb{E}\left[\Psi_\mu(x_R) - \Psi_\mu(x^*)\right] + \mu^2 Ln.$$

Then, by this inequality and the convexity of f_μ, it follows from part (b) of Theorem 2 and similar arguments in showing the part (a) of this theorem, the conclusion (86) 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 6 *Suppose that in the RSPGF algorithm the stepsizes $\gamma_k = \alpha/(2L)$ 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 (31). Then under Assumptions A1 and A3, we have*

$$\mathbb{E}[\|\bar{g}_{\mu,X,R}\|^2] \le \frac{4L^2 D_\Psi^2 + 4\mu^2 L^2 n}{\alpha^2 N} + \frac{2\tilde{\sigma}^2}{\alpha^2 m} \tag{89}$$

and

$$\mathbb{E}[\|g_{X,R}\|^2] \le \frac{\mu^2 L^2 (n+3)^2}{2\alpha^2} + \frac{16L^2 D_\Psi^2 + 16\mu^2 L^2 n}{\alpha^2 N} + \frac{12\tilde{\sigma}^2}{\alpha^2 m}, \tag{90}$$

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 (84), (85) and (21), respectively.

If, in addition, f in the problem (1) is convex with an optimal solution x^*, then

$$\mathbb{E}\left[\Psi(x_R) - \Psi(x^*)\right] \le \frac{2LV(x^*, x_1)}{N\alpha} + \frac{\tilde{\sigma}^2}{2Lm} + \mu^2 Ln. \tag{91}$$

Proof (89) immediately follows from (83) with $\gamma_k = \alpha/(2L)$ and $m_k = m$ for all $k = 1, \ldots, N$. Now let $g_{\mu,X,R} = P_X(x_R, \nabla f_\mu(x_R), \gamma_R)$, we have from (77) and Proposition 1 with $x = x_R$, $\gamma = \gamma_R$, $g_1 = \nabla f(x_R)$ and $g_2 = \nabla f_\mu(x_R)$ that

$$\mathbb{E}[\|g_{X,R} - g_{\mu,X,R}\|^2] \le \frac{\mu^2 L^2 (n+3)^2}{4\alpha^2}. \tag{92}$$

Similarly, by Proposition 1 with $x = x_R$, $\gamma = \gamma_R$, $g_1 = \bar{G}_{\mu,k}$ and $g_2 = \nabla f_\mu(x_R)$, we have

$$\mathbb{E}[\|\bar{g}_{\mu,X,R} - g_{\mu,X,R}\|^2] \le \frac{\tilde{\sigma}^2}{\alpha^2 m}. \tag{93}$$

Then, it follows from (92), (93) and (89) that

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

Moreover, if f is convex, then (91) immediately follows from (86), and the constant stepsizes $\gamma_k = \alpha/(2L)$ 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 μ as well. The following corollary, analogous to the Corollary 4, shows how to choose m and μ appropriately.

Corollary 7 *Suppose that all the conditions in Corollary 6 are satisfied. Given a fixed total number of calls to the $\mathcal{SZO}$ $\bar{N}$, if the smoothing parameter satisfies*

$$\mu \leq \frac{D_\Psi}{\sqrt{(n+4)\bar{N}}}, \tag{94}$$

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

$$m = \left\lceil \min \left\{ \max \left\{ \frac{\sqrt{(n+4)(M^2+\sigma^2)\bar{N}}}{L\tilde{D}}, n+4 \right\}, \bar{N} \right\} \right\rceil, \tag{95}$$

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

$$\bar{\mathcal{B}}_{\bar{N}} := \frac{(24\theta_2+41)LD_\Psi^2(n+4)}{\bar{N}} + \frac{32\sqrt{(n+4)(M^2+\sigma^2)}}{\sqrt{\bar{N}}} \left(\frac{D_\Psi^2}{\tilde{D}} + \tilde{D}\theta_1 \right), \tag{96}$$

and

$$\theta_1 = \max\left\{1, \frac{\sqrt{(n+4)(M^2+\sigma^2)}}{L\tilde{D}\sqrt{\bar{N}}}\right\} \quad \textit{and} \quad \theta_2 = \max\left\{1, \frac{n+4}{\bar{N}}\right\}. \tag{97}$$

If, in addition, f in the problem (1) is convex and the smoothing parameter satisfies

$$\mu \leq \sqrt{\frac{V(x^*, x_1)}{\alpha(n+4)\bar{N}}}, \tag{98}$$

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

$$\bar{\mathcal{C}}_{\bar{N}} := \frac{(5+\theta_2)LV(x^*, x_1)(n+4)}{\alpha\bar{N}} + \frac{\sqrt{(n+4)(M^2+\sigma^2)}}{\alpha\sqrt{\bar{N}}} \left(\frac{4V(x^*, x_1)}{\tilde{D}} + \alpha\tilde{D}\theta_1 \right). \tag{99}$$

Proof By the definitions of $\theta 1$ and θ_2 in (97) and m in (95), we have

$$m = \left\lceil \max\left\{ \frac{\sqrt{(n+4)(M^2+\sigma^2)\bar{N}}}{L\tilde{D}\theta_1}, \frac{n+4}{\theta_2} \right\} \right\rceil. \tag{100}$$

Given the total number of calls to the $\mathcal{SZO}$ $\bar{N}$ and the the number m of calls to the $\mathcal{SZO}$ at each iteration, the RSPGF algorithm can perform at most $N = \lfloor \bar{N}/m \rfloor$

iterations. Obviously, $N \geq \bar{N}/(2m)$. With this observation $\bar{N} \geq m$, $\theta_1 \geq 1$ and $\theta_2 \geq 1$, by (90), (94) and (100), we have

$$
\begin{aligned}
&\mathbb{E}[\|g_{X,R}\|^2] \\
\leq\ & \frac{L^2 D_\Psi^2 (n+3)}{2\alpha^2 \bar{N}} + \frac{24(n+4)(M^2+\sigma^2)}{\alpha^2 m} + \frac{24 L^2 D_\Psi^2 (n+4)^2}{\alpha^2 m \bar{N}} + \frac{32 L^2 D_\Psi^2 m}{\alpha^2 \bar{N}} \left(1 + \frac{1}{\bar{N}}\right) \\
\leq\ & \frac{L^2 D_\Psi^2 (n+4)}{2\alpha^2 \bar{N}} + \frac{24\theta_1 L \tilde{D} \sqrt{(n+4)(M^2+\sigma^2)}}{\alpha^2 \sqrt{\bar{N}}} + \frac{24\theta_2 L^2 D_\Psi^2 (n+4)}{\alpha^2 \bar{N}} \\
& + \frac{32 L^2 D_\Psi^2}{\alpha^2 \bar{N}} \left(\frac{\sqrt{(n+4)(M^2+\sigma^2)\bar{N}}}{L \tilde{D} \theta_1} + \frac{n+4}{\theta_2} \right) + \frac{32 L^2 D_\Psi^2}{\alpha^2 \bar{N}} \\
\leq\ & \frac{L^2 D_\Psi^2 (n+4)}{2\alpha^2 \bar{N}} + \frac{24\theta_1 L \tilde{D} \sqrt{(n+4)(M^2+\sigma^2)}}{\alpha^2 \sqrt{\bar{N}}} + \frac{24\theta_2 L^2 D_\Psi^2 (n+4)}{\alpha^2 \bar{N}} \\
& + \frac{32 L D_\Psi^2 \sqrt{(n+4)(M^2+\sigma^2)}}{\alpha^2 \tilde{D} \sqrt{\bar{N}}} + \frac{32 L^2 D_\Psi^2 (n+4)}{\alpha^2 \bar{N}} + \frac{32 L^2 D_\Psi^2}{\alpha^2 \bar{N}},
\end{aligned}
$$

which after integrating the terms give (96). The conclusion (99) follows similarly by (98) and (91).

We now would like to add a few remarks about the above the results in Corollary 7. 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 7 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 (96) and (99) that when $\bar{N}$ is relatively large such that $\theta_1 = 1$ and $\theta_2 = 1$, i.e.,

$$\bar{N} \geq \max\left\{ \frac{(n+4)^2(M^2+\sigma^2)}{L^2 \tilde{D}^2}, n+4 \right\}, \tag{101}$$

the optimal choice of $\tilde{D}$ would be D_Ψ and $2\sqrt{V(x^*, x_1)/\alpha}$ for solving nonconvex and convex SP problems, respectively. With this selection of $\tilde{D}$, the bounds in (96) and (99), respectively, reduce to

$$\frac{\alpha^2}{L}\mathbb{E}[\|g_{X,R}\|^2] \leq \frac{65 L D_\Psi^2 (n+4)}{\bar{N}} + \frac{64\sqrt{(n+4)(M^2+\sigma^2)}}{\sqrt{\bar{N}}} \tag{102}$$

and

$$\mathbb{E}[\Psi(x_R) - \Psi(x^*)] \leq \frac{6 L V(x^*, x_1)(n+4)}{\alpha \bar{N}} + \frac{4\sqrt{V(x^*, x_1)(n+4)(M^2+\sigma^2)}}{\sqrt{\alpha \bar{N}}}. \tag{103}$$

Thirdly, the complexity result in (99) implies that when f is convex, if ϵ sufficiently small, then the number of calls to the $\mathcal{SZO}$ to find a solution $\bar{x}$ such that $\mathbb{E}[f(\bar{x}) - f^*] \leq \epsilon$ can be bounded by $\mathcal{O}(n/\epsilon^2)$, which is better than the complexity of $\mathcal{O}(n^2/\epsilon^2)$ established by Nesterov [26] 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 least square problem with a nonconvex regularization term and a stochastic nonconvex semi-supervised support vector machine problem.

Algorithmic schemes. We implement the RSPG algorithm and its two-phase variant 2-RSPG algorithm described in Section 4, where the prox-function $V(x,z) = \|x - z\|^2/2$, the stepsizes $\gamma_k = \alpha/(2L)$ with $\alpha = 1$ for all $k \geq 1$, and the probability mass function P_R is set to (31). Also, in the optimization phase of the 2-RSPG algorithm, we take $S = 5$ independent runs of the RSPG algorithm to compute 5 candidate solutions. Then, we use an i.i.d. sample of size $T = N/2$ in the post-optimization phase to estimate the projected gradients at these candidate solutions and then choose the best one, $\bar{x}^*$, according to (55). Finally, the solution quality at $\bar{x}^*$ is evaluated by using another i.i.d. sample of size $K >> N$. In addition to the above algorithms, we also implement another variant of the 2-RSPG algorithm, namely, 2-RSPG-V algorithm. This algorithm also consists of two phases similar to the 2-RSPG algorithm. In the optimization phase, instead of terminating the RSPG algorithm by using a random count R, we terminate the algorithm by using a fixed number of iterations, say NS. We then randomly pick up $S = 5$ solutions from the generated trajectory according to P_R defined in (31). The post-optimization phase of the 2-RSPG-V algorithm is the same as that of the 2-RSPG algorithm. Note that, in the 2-RSPG-V algorithm, unlike the 2-RSPG algorithm, the S candidate solutions are not independent and hence, we cannot provide the large-deviation results similar to the 2-RSPG algorithm. We also implement the RSG, 2-RSG and 2-RSG-V algorithms developed in [15] to compare with our results.

Estimation of parameters. We use an initial i.i.d. sample of size $N_0 = 200$ to estimate the problem parameters, namely, L and σ. We also estimate the parameter $\tilde{D} = D_\Psi$ by (23). More specifically, since the problems considered in this section have nonnegative optimal values, i.e., $\Psi^* \geq 0$, we have $D_\Psi \leq (2\Psi(x_1)/L)^{\frac{1}{2}}$, where x_1 denotes the starting point of the algorithms.

Notation in the tables.

- NS denotes the maximum number of calls to the stochastic oracle performed in the optimization phase of the above algorithms. For example, $NS = 1,000$ has the following implications.
  - For the RSPG algorithm, the number of samples per iteration m is computed according to (46) with $\bar{N} = 1000$ and the iteration limit N is set to $\lfloor 1000/m \rfloor$;
  - For the 2-RSPG algorithm, 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 2-RSPG-V algorithm, m is computed as mentioned above and we run the RSPG method for $\lfloor 1000/m \rfloor$ iterations and randomly select $S = 5$ solutions from the trajectory according to P_R defined in (31).
- $\bar{x}^*$ is the output solution of the above algorithms.
- *Mean* and *Var.* represent, respectively, the average and variance of the results obtained over 20 runs of each algorithm.

Table 1 Estimated $\|\nabla f(\bar{x}^*)\|^2$ for the least square problem ($K = 75,000$)

NS		RSG	2-RSG	2-RSG-V	RSPG	2-RSPG	2-RSPG-V
		$n = 100, \tilde{\sigma} = 0.1$					
1000	mean	0.2509	0.3184	0.0794	0.1564	0.3176	0.0422
	var.	4.31e-2	1.68e-2	1.23e-3	4.58e-2	2.54e-2	8.99e-3
5000	mean	0.0828	0.0841	0.0042	0.0113	0.0164	0.0009
	var.	6.75e-3	1.03e-3	1.35e-5	4.22e-4	3.37e-4	4.36e-8
25000	mean	0.0056	0.0070	0.0002	0.0006	0.0010	0.0004
	var.	1.69e-4	1.08e-4	3.41e-8	2.05e-7	1.43e-7	7.83e-9
		$n = 100, \tilde{\sigma} = 1$					
1000	mean	0.3731	0.3761	0.1230	0.2379	0.3567	0.0364
	var.	3.38e-2	1.40e-2	3.28e-3	4.01e-2	1.41e-2	1.24e-3
5000	mean	0.1095	0.1314	0.0135	0.0436	0.0323	0.0075
	var.	2.22e-2	3.96e-3	4.67e-5	1.44e-2	8.69e-4	7.97e-5
25000	mean	0.0374	0.0172	0.0078	0.0138	0.0048	0.0046
	var.	8.46e-3	1.83e-4	4.54e-4	1.95e-3	8.48e-7	5.60e-5
		$n = 500, \tilde{\sigma} = 0.1$					
1000	mean	0.5479	0.6865	0.4121	0.4212	0.8977	0.2579
	var.	3.47e-2	6.17e-3	1.09e-2	5.13e-2	2.64e-3	1.34e-2
5000	mean	0.2481	0.3560	0.0873	0.1030	0.1997	0.0154
	var.	4.38e-2	3.45e-3	1.28e-3	2.57e-2	2.21e-3	1.83e-4
25000	mean	0.2153	0.0876	0.0084	0.1093	0.0136	0.0011
	var.	6.77e-2	1.13e-3	3.97e-5	4.07e-2	3.24e-5	3.04e-8
		$n = 500, \tilde{\sigma} = 1$					
1000	mean	0.5869	0.7444	0.4828	0.4371	0.7771	0.4190
	var.	2.14e-2	4.18e-3	9.40e-3	3.40e-2	5.15e-3	4.13e-2
5000	mean	0.3603	0.4732	0.1699	0.1745	0.2987	0.0411
	var.	3.77e-2	8.13e-3	1.22e-3	3.51e-2	1.87e-2	6.21e-4
25000	mean	0.2467	0.1584	0.0342	0.1271	0.0351	0.0189
	var.	6.49e-2	1.87e-3	3.72e-4	4.30e-2	2.83e-4	3.89e-5
		$n = 1000, \tilde{\sigma} = 0.1$					
1000	mean	1.853	2.417	1.549	1.855	3.092	1.937
	var.	1.73e-1	1.31e-2	1.62e-2	1.88e-1	1.29e-1	2.64e-1
5000	mean	0.9555	1.501	0.5422	0.4944	1.832	0.1368
	var.	3.62e-1	6.39e-2	3.73e-2	4.82e-1	2.36e-1	8.78e-3
25000	mean	0.6305	0.4725	0.0839	0.3402	0.1100	0.0071
	var.	6.38e-1	2.08e-2	1.19e-2	4.40e-1	4.54e-3	1.97e-4
		$n = 1000, \tilde{\sigma} = 1$					
1000	mean	1.868	2.407	1.560	1.701	3.208	1.662
	var.	1.44e-1	1.22e-2	4.37e-2	1.84e-1	1.54e-1	2.75e-1
5000	mean	1.297	1.596	0.6438	0.8032	1.403	0.2408
	var.	5.25e-1	5.26e-2	3.04e-2	6.38e-1	1.10e-1	3.26e-2
25000	mean	0.575	0.6309	0.0793	0.2079	0.1806	0.0336
	var.	3.43e-1	4.65e-2	1.38e-3	1.17e-1	1.43e-2	3.67e-6

6.1 Nonconvex least square problem

In our first experiment, we consider the following least square problem with a smoothly clipped absolute deviation penalty term given in [10]:

$$\min_{x \in \mathbb{R}^n} f(x) := \mathbb{E}_{u,v}[(\langle x, u \rangle - v)^2] + \sum_{j=1}^{d} p_\lambda(|x_j|).$$

Table 2 Average ratio of true recovered zeros for the penalized least square problem

$\bar{N}S$	RSG	2-RSG	2-RSG-V	RSPG	2-RSPG	2-RSPG-V
			$n = 100, \tilde{\sigma} = 0.1$			
1000	0.18	0.14	0.11	0.17	0.13	0.19
5000	0.26	0.11	0.60	0.56	0.19	0.98
25000	0.82	0.56	0.98	0.97	0.95	1.00
			$n = 100, \tilde{\sigma} = 1$			
1000	0.16	0.14	0.09	0.12	0.11	0.09
5000	0.16	0.09	0.21	0.14	0.08	0.15
25000	0.41	0.20	0.51	0.26	0.16	0.24
			$n = 500, \tilde{\sigma} = 0.1$			
1000	0.14	0.17	0.09	0.08	0.27	0.06
5000	0.18	0.08	0.22	0.31	0.06	0.56
25000	0.47	0.21	0.84	0.63	0.55	0.99
			$n = 500, \tilde{\sigma} = 1$			
1000	0.16	0.23	0.10	0.09	0.17	0.12
5000	0.16	0.10	0.17	0.12	0.06	0.16
25000	0.39	0.17	0.60	0.33	0.17	0.45
			$n = 1000, \tilde{\sigma} = 0.1$			
1000	0.1	0.17	0.07	0.05	0.25	0.10
5000	0.10	0.05	0.09	0.10	0.04	0.11
25000	0.31	0.09	0.51	0.55	0.12	0.91
			$n = 1000, \tilde{\sigma} = 1$			
1000	0.12	0.14	0.05	0.09	0.20	0.08
5000	0.11	0.06	0.09	0.08	0.03	0.08
25000	0.20	0.09	0.40	0.28	0.09	0.45

Here, the penalty term $p_\lambda : \mathbb{R}_+ \to \mathbb{R}$ satisfies $p_\lambda(0) = 0$ and has derivatives as

$$p'_\lambda(\beta) = \lambda \left\{ I(\beta \le \lambda) + \frac{\max(0, a\lambda - \beta)}{(a-1)\lambda} I(\beta > \lambda) \right\},$$

where $a > 2$ and $\lambda > 0$ are constant parameter, and I is the indicator function. As it can be seen, $p_\lambda(|\cdot|)$ is nonconvex and non-differentiable at 0. Therefore, we replace p_λ by its smooth nonconvex approximation $q_\lambda : \mathbb{R}_+ \to \mathbb{R}$, satisfying $q_\lambda(0) = 0$ with derivative defined as

$$q'_\lambda(\beta) = \left\{ \beta I(\beta \le \lambda) + \frac{\max(0, a\lambda - \beta)}{(a-1)} I(\beta > \lambda) \right\}.$$

In this experiment, we assume that u is a sparse vector, whose components are standard normal, and v is obtained by $v = \langle \bar{x}, u \rangle + \xi$, where $\xi \sim N(0, \bar{\sigma}^2)$ is the random noise independent of u and the coefficient $\bar{x}$ defines the true linear relationship between u and v. Also, we set the parameters to $a = 3.7$ and $\lambda = 0.01$ in the numerical experiments.

We consider three different problem sizes with $n = 100, 500$ and $1,000$, and two different noise levels with $\bar{\sigma} = 0.1$ and 1. Also, we set the data sparsity to 5%, which means that approximately five percent of u is nonzero for each data point (u, v). We also use a sparse multivariate standard normal $\bar{x}$ for generating data points. Also, for all problem sizes, the initial point is set to $x_1 = 5 * \bar{x}_0 \in \mathbb{R}^n$, where $\bar{x}_0$ is a multivariate standard normal vector with approximately 10% nonzero elements. Since this problem is unconstrained, we also implement variants of the RSG algorithm developed in [15]. Table 1 shows the mean and variance of the 2-norm of the gradient at the solutions returned by 20 runs of the comparing algorithms. Moreover, we are also interested in recovering sparse solutions. Hence, we set the component of output solutions to be zero if its absolute value is less than a threshold 0.02. We call such solutions as the recovered truncated zeros and compute their ratio with respect to the number of true zeros. Table 2 shows the average of this ratio over 20 runs of the algorithms. The following observations can be made from the numerical results.

- **Different variants of the RSPG algorithm:** Firstly, over 20 runs of the algorithm, the solutions of the RSPG algorithm has relatively large variance. Secondly, both 2-RSPG and 2-RSPG-V can significantly reduce the variance of the RSPG algorithm. Moreover, for a given fixed NS, the solution quality of the 2-RSPG-V algorithm is significantly better than that of the 2-RSPG algorithm. The reason might be that, for fixed NS, $S = 5$ times more iterations are being used in the 2-RSPG-V algorithm to generate new solutions in the trajectory.
- **Different variants of the RSG algorithm:** The differences among different variants of the RSG algorithm are similar to those of the corresponding variants of the RSPG algorithm.
- **RSPG algorithm vs. RSG algorithm:** In terms of the mean value, the solutions give by RSG and RSPG algorithms are comparable. However, the solution of the RSPG algorithm usually have less variance than that of the corresponding RSG algorithm. The possible reason is that we use a better approximation for stochastic gradient by incorporating mini-batch of samples during the execution of the RSPG method.

6.2 Semi-supervised support vector machine problem

In this second 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]:

$$\min_{b\in\mathbb{R},\ x\in\mathbb{R}^n} f(x,b) := \lambda_1 \mathbb{E}_{u_1,v}\left[\max\{0, 1 - v(\langle x, u_1\rangle + b)\}^2\right] \\ +\lambda_2 \mathbb{E}_{u_2}\left[\max\{0, 1 - |\langle x, u_2\rangle + b|\}^2\right] + \lambda_3 \|x\|_2^2,$$

where (u_1, v) 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). Using a smooth approximation of the above problem [7], we can reformulate it as

$$\min_{(x,b)\in\mathbb{R}^{n+1}} f(x,b) := \mathbb{E}_{u_1,u_2,v}\left[\lambda_1 \max\{0, 1 - v(\langle x, u_1\rangle + b)\}^2 + \lambda_2 e^{-5\{\langle x,u_2\rangle + b\}^2}\right] \\ +\lambda_3 \|x\|_2^2. \tag{104}$$

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 $v \in \{0, 1\}$ with $v = \text{sgn}(\langle \bar{x}, u' \rangle + b)$ for some $\bar{x} \in \mathbb{R}^n$. The parameters are also set to $\lambda_1 = 1$, $\lambda_2 = 0.5$ and $\lambda_3 = 0.5$. The choices of problem size and the noise variance are same as those of the nonconvex penalized least square problem in the previous subsection. We also want to determine the labels of unlabeled examples such that the

Table 3 Estimated $\|g_X(\bar{x}^*)\|^2$ for the semi-supervised support vector machine problem ($K = 75,000$)

NS		RSPG	2-RSPG	2-RSPG-V
$n = 100$				
1000	mean	1.355	0.2107	0.1277
	var.	1.21e+1	9.50e-3	5.45e-3
5000	mean	0.1032	0.1174	0.0899
	var.	4.96e-2	4.42e-3	6.28e-3
25000	mean	0.0352	0.0699	0.0239
	var.	1.13e-3	3.42e-3	1.73e-5
$n = 500$				
1000	mean	5.976	0.7955	0.1621
	var.	1.93e+2	6.07e-1	1.15e-3
5000	mean	0.2237	0.1703	0.0928
	var.	2.77e-1	4.39e-3	1.29e-3
25000	mean	0.2174	0.0832	0.0339
	var.	2.35e-1	2.41e-4	8.04e-6
$n = 1000$				
1000	mean	27.06	2.417	0.3167
	var.	6.00e+3	1.73e+1	1.19e-2
5000	mean	16.24	0.4726	0.1463
	var.	2.20e+3	2.85e+1	1.46e-3
25000	mean	0.1007	0.1378	0.0672
	var.	2.46e-2	5.63e-5	5.10e-5

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

$$|b - 2r + 1| \leq \delta, \tag{105}$$

where r is the ratio of positive labels in the already labeled examples and δ is a tolerance setting to 0.1 in our experiment. Therefore, (104) together with the constraint (105) is a constrained nonconvex smooth problem, which fits the setting of problem (1). Table 3 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 Table 4 gives the corresponding average objective values at the solutions given in Table 3. Similar to the conclusions in the previous subsection, we again can see 2-RSPG-V algorithm has the best performance among the variants of the RSPG algorithms and 2-RSPG algorithms is more stable than the RSPG algorithm.

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

Table 4 Average objective values at $\bar{x}^*$, obtained in Table 3

$\bar{N}S$	RSPG	2-RSPG	2-RSPG-V
$n = 100$			
1000	1.497	0.9331	0.9094
5000	0.9131	0.9078	0.8873
25000	0.8736	0.8862	0.8728
$n = 500$			
1000	3.813	1.364	1.038
5000	1.062	1.043	0.9998
25000	1.062	0.9982	0.9719
$n = 1000$			
1000	14.05	2.100	1.055
5000	8.77	1.128	0.9767
25000	0.9513	0.9719	0.9351

RSPG algorithm would automatically maintains 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 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.

References

1. S. Andradóttir, *A review of simulation optimization techniques*, Proceedings of the 1998 Winter Simulation Conference, pp. 151–158.
2. A. Auslender and M. Teboulle, *Interior gradient and proximal methods for convex and conic optimization*, SIAM Journal on Optimization, 16 (2006), pp. 697–725.
3. H.H. Bauschke, J.M. Borwein, and P.L. Combettes, *Bregman monotone optimization algorithms*, SIAM Journal on Controal and Optimization, 42 (2003), pp. 596–636.
4. A. Ben-Tal, T. Margalit, and A. S. Nemirovski, *The ordered subsets mirror descent optimization method with applications to tomography*, SIAM Journal on Optimization, 12 (2001), pp. 79–108.
5. L.M. Bregman, *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 (1967), pp. 200–217.
6. C. Cartis, N. I. M. Gould, and Ph. L. Toint, *On the complexity of steepest descent, newton's and regularized newton's methods for nonconvex unconstrained optimization*, SIAM Journal on Optimization, 20(6) (2010), pp. 2833–2852.
7. O. Chapelle, V. Sindhwani, and S. S. Keerthi, *Optimization techniques for semi-supervised support vector machines*, Journal of Machine Learning Research, 9 (2008), pp. 203–233.
8. C. D. Dang and G. Lan, *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. J. C. Duchi, P. L. Bartlett, and M. J. Wainwright, *Randomized smoothing for stochastic optimization*, SIAM Journal on Optimization, 22 (2012), pp. 674–701.
10. J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American Statistical Association, 96 (2001), p. 13481360.

11. M.C. Fu, *Gradient estimation*, in Handbooks in Operations Research and Management Science: Simulation, S. G. Henderson and B. L. Nelson, eds., Elsevier, pp. 575–616.
12. M. Fu, *Optimization for simulation: Theory vs. practice*, INFORMS Journal on Computing, 14 (2002), pp. 192–215.
13. S. Ghadimi and G. Lan, *Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, I: a generic algorithmic framework*, SIAM Journal on Optimization, 22 (2012), pp. 1469–1492.
14. ———, *Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms*, SIAM Journal on Optimization, (2013). to appear.
15. ———, *Stochastic first- and zeroth-order methods for nonconvex stochastic programming*, technical report, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, June 2012. *SIAM Journal on Optimization* (under second-round review).
16. P. Glasserman, *Gradient Estimation via Perturbation Analysis*, Kluwer Academic Publishers, Boston,Massachusetts, 2003.
17. A. Juditsky and A. S. Nemirovski, *Large deviations of vector-valued martingales in* 2-*smooth normed spaces*, manuscript, Georgia Institute of Technology, Atlanta, GA, 2008. E-print: www2.isye.gatech.edu/∼ nemirovs/LargeDevSubmitted.pdf.
18. ———, *First-Order Methods for Nonsmooth Convex Large-Scale Optimization, I: General Purpose Methods* , in Optimization for Machine Learning, Eds: S. Sra, S. Nowozin and S.J. Wright, MIT press, 2011.
19. G. Lan, *An optimal method for stochastic composite optimization*, Mathematical Programming, 133(1) (2012), pp. 365–397.
20. P. LÉcuyer, *A unified view of the IPA, SF, and LR gradient estimation techniques*, Management Science, 36(11) (1990), pp. 1364–1383.
21. J. Mairal, F. Bach, J. Ponce, and G. Sapiro, *Online dictionary learning for sparse coding*, in In ICML, 2009, pp. 689–696.
22. L. Mason, J. Baxter, P. Bartlett, and M. Frean, *Boosting algorithms as gradient descent in function space*, Proc. NIPS, 12 (1999), pp. 512–518.
23. A. S. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal on Optimization, 19 (2009), pp. 1574–1609.
24. A. S. Nemirovski and D. Yudin, *Problem complexity and method efficiency in optimization*, Wiley-Interscience Series in Discrete Mathematics, John Wiley, XV, 1983.
25. Y. E. Nesterov, *Introductory Lectures on Convex Optimization: a basic course*, Kluwer Academic Publishers, Massachusetts, 2004.
26. ———, *Random gradient-free minimization of convex functions*, tech. report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, January 2010.
27. B.T. Polyak, *New stochastic approximation type procedures*, Automat. i Telemekh., 7 (1990), pp. 98–107.
28. B.T. Polyak and A.B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control and Optimization, 30 (1992), pp. 838–855.
29. H. Robbins and S. Monro, *A stochastic approximation method*, Annals of Mathematical Statistics, 22 (1951), pp. 400–407.
30. R. T. Rockafellar and R. J.-B. Wets, *Variational analysis, ser. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*, Springer-Verlag, Berlin, 1998.
31. R.Y. Rubinstein and A. Shapiro, *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*, John Wiley & Sons, 1993.
32. J.C. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, John Wiley, Hoboken, NJ, 2003.
33. M. Teboulle, *Convergence of proximal-like algorithms*, SIAM Journal on Optimization, 7 (1997), pp. 1069–1083.