STOCHASTIC FIRST- AND ZEROTH-ORDER METHODS FOR NONCONVEX STOCHASTIC PROGRAMMING *

SAEED GHADIMI † AND GUANGHUI LAN ‡

Abstract. In this paper, we introduce a new stochastic approximation (SA) type algorithm, namely the randomized stochastic gradient (RSG) method, for solving an important class of nonlinear (possibly nonconvex) stochastic programming (SP) problems. We establish the complexity of this method for computing an approximate stationary point of a nonlinear programming problem. We also show that this method possesses a nearly optimal rate of convergence if the problem is convex. We discuss a variant of the algorithm which consists of applying a post-optimization phase to evaluate a short list of solutions generated by several independent runs of the RSG method, and show that such modification allows to improve significantly the large-deviation properties of the algorithm. These methods are then specialized for solving a class of simulation-based optimization problems in which only stochastic zeroth-order information is available.

Keywords: stochastic approximation, nonconvex optimization, stochastic programming, simulation-based optimization

1. Introduction. In 1951, Robbins and Monro in their seminal work [34] proposed a classical stochastic approximation (SA) algorithm for solving stochastic programming (SP) problems. This approach mimics the simplest gradient descent method by using noisy gradient information in place of the exact gradients, and possesses the "asymptotically optimal" rate of convergence for solving a class of strongly convex SP problems [4, 38]. However, it is usually difficult to implement the "asymptotically optimal" stepsize policy, especially in the beginning, so that the algorithms often perform poorly in practice (e.g., [40, Section 4.5.3]). An important improvement of the classical SA was developed by Polyak [32] and Polyak and Juditsky [33], where longer stepsizes were suggested together with the averaging of the obtained iterates. Their methods were shown to be more robust with respect to the selection of stepsizes than the classical SA and also exhibit the "asymptotically optimal" rate of convergence for solving strongly convex SP problems. We refer to [24] for an account of the earlier history of SA methods.

The last few years have seen some significant progress for the development of SA methods for SP. On one hand, new SA type methods are being introduced to solve SP problems which are not necessarily strongly convex. On the other hand, these developments, motivated by complexity theory in convex optimization [25], concerned the convergence properties of SA methods during a finite number of iterations. For example, Nemirovski et al. [24] presented a properly modified SA approach, namely, mirror descent SA for solving general non-smooth convex SP problems. They demonstrated that the mirror descent SA exhibits an optimal $\mathcal{O}(1/\epsilon^2)$ iteration complexity for solving these problems. This method has been shown in [19, 24] to be competitive to the widely-accepted sample average approximation approach (see, e.g., [17, 39]) and even significantly outperform it for solving a class of convex SP problems. Similar techniques, based on subgradient averaging, have been proposed in [14, 16, 28]. While

*The first author was partially supported by NSF Grant CMMI-1000347 and the second author was partially supported by NSF grant CMMI-1000347, ONR grant N00014-13-1-0036 and NSF CAREER Award CMMI-1254446.

† Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, (email: sghadimi@ufl.edu).

‡Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, (email: glan@ise.ufl.edu).

these techniques dealt with non-smooth convex programming problems, Lan [18] presented a unified optimal method for smooth, non-smooth and stochastic optimization, which explicitly takes into account the smoothness of the objective function (see also [11, 10] for discussions about strong convexity). However, note that convexity has played an important role in establishing the convergence of all these SA algorithms. To the best of our knowledge, none of existing SA algorithms can handle more general SP problems whose objective function is possibly nonconvex.

This paper focuses on the theoretical development of SA type methods for solving an important class of nonconvex SP problems. More specifically, we study the classical unconstrained nonlinear programming (NLP) problem given in the form of (e.g., [27, 31])

$$f^* := \inf_{x\in\mathbb{R}^n} f(x), \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable (not necessarily convex), bounded from below, and its gradient $\nabla f(\cdot)$ satisfies

$$\|\nabla f(y) - \nabla f(x)\| \le L\|y - x\|, \quad \forall x, y \in \mathbb{R}^n.$$

However, different from the standard NLP, we assume throughout the paper that we only have access to noisy function values or gradients about the objective function f in (1.1). In particular, in the basic setting, we assume that problem (1.1) is to be solved by iterative algorithms which acquire the gradients of f via subsequent calls to a stochastic first-order oracle ($\mathcal{SFO}$). At iteration k of the algorithm, x_k being the input, the $\mathcal{SFO}$ outputs a *stochastic gradient* $G(x_k, \xi_k)$, where ξ_k, $k \ge 1$, are random variables whose distributions P_k are supported on $\Xi_k \subseteq \mathbb{R}^d$. The following assumptions are made for the Borel functions $G(x_k, \xi_k)$.

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

$$\text{a) } \mathbb{E}[G(x_k, \xi_k)] = \nabla f(x_k), \tag{1.2}$$
$$\text{b) } \mathbb{E}\left[\|G(x_k, \xi_k) - \nabla f(x_k)\|^2\right] \le \sigma^2, \tag{1.3}$$

for some parameter $\sigma \ge 0$. Observe that, by (1.2), $G(x_k, \xi_k)$ is an unbiased estimator of $\nabla f(x_k)$ and, by (1.3), the variance of the random variable $\|G(x_k, \xi_k) - \nabla f(x_k)\|$ is bounded. It is worth noting that in the standard setting for SP, the random vectors ξ_k, $k = 1, 2, \ldots$, are independent of each other (and also of x_k) (see, e.g., [25, 24]). Our assumption here is slightly weaker since we do not need to assume ξ_k, $k = 1, 2, \ldots$, to be independent.

Our study on the aforementioned SP problems has been motivated by a few interesting applications which are briefly outlined as follows.

- In many *machine learning* problems, we intend to minimize a regularized loss function $f(\cdot)$ given by

$$f(x) = \int_\Xi L(x, \xi) dP(\xi) + r(x), \tag{1.4}$$

  where either the loss function $L(x, \xi)$ or the regularization $r(x)$ is nonconvex (see, e.g., [22, 23]).
- Another important class of problems originate from the so-called *endogenous uncertainty* in SP. More specifically, the objective functions for these

SP problems are given in the form of

$$f(x) = \int_{\Xi(x)} F(x, \xi) dP_x(\xi), \tag{1.5}$$

where the support $\Xi(x)$ and the distribution function P_x of the random vector ξ depend on x. The function f in (1.5) is usually nonconvex even if $F(x, \xi)$ is convex with respect to x. For example, if the support Ξ does not depend on x, it is often possible to represent $dP_x = H(x)dP$ for some fixed distribution P. Typically this transformation results in a nonconvex integrand function. Other techniques have also been developed to compute unbiased estimators for the gradient of $f(\cdot)$ in (1.5) (see, e.g., [8, 13, 20, 36]).

- Finally, in *simulation-based optimization*, the objective function is given by $f(x) = \mathbb{E}_\xi[F(x, \xi)]$, where $F(\cdot, \xi)$ is not given explicitly, but through a black-box simulation procedure (e.g., [1, 7]). Therefore, we do not know if the function f is convex or not. Moreover, in these cases, we usually only have access to stochastic zeroth-order information about the function values of $f(\cdot)$ rather than its gradients.

The complexity of the gradient descent method for solving problem (1.1) has been well-understood under the deterministic setting (i.e., $\sigma = 0$ in (1.3)). In particular, Nesterov [27] shows that after running the method for at most $N = \mathcal{O}(1/\epsilon)$ steps, we have $\min_{k=1,\ldots,N} \|\nabla f(x_k)\|^2 \leq \epsilon$ (see Gratton et al. [37] for a similar bound for the trust-region methods). Cartis et al. [2] show that this bound is actually tight for the gradient descent method. Note, however, that the analysis in [27] is not applicable to the stochastic setting (i.e., $\sigma > 0$ in (1.3)). Moreover, even if we have $\min_{k=1,\ldots,N} \|\nabla f(x_k)\|^2 \leq \epsilon$, to find the best solution from $\{x_1, \ldots, x_N\}$ is still difficult since $\|\nabla f(x_k)\|$ is not known exactly. Our major contributions in this paper are summarized as follows. Firstly, to solve the aforementioned nonconvex SP problem, we present a randomized stochastic gradient (RSG) method by introducing the following modifications to the classical SA. Instead of taking average of the iterates as in the mirror descent SA for convex SP, we randomly select a solution $\bar{x}$ from $\{x_1, \ldots, x_N\}$ according to a certain probability distribution as the output. We show that such a solution satisfies $\mathbb{E}[\|\nabla f(\bar{x})\|^2] \leq \epsilon$ after running the method for at most $N = \mathcal{O}(1/\epsilon^2)$ iterations [1]. Moreover, if $f(\cdot)$ is convex, we show that the relation $\mathbb{E}[f(\bar{x}) - f^*] \leq \epsilon$ always holds. We demonstrate that such a complexity result is nearly optimal for solving convex SP problems (see the discussions after Corollary 2.2).

Secondly, in order to improve the large deviation properties and hence the reliability of the RSG method, we present a two-phase randomized stochastic gradient (2-RSG) method by introducing a post-optimization phase to evaluate a short list of solutions generated by several independent runs of the RSG method. We show that the complexity of the 2-RSG method for computing an (ϵ, Λ)-*solution* of problem (1.1), i.e., a point $\bar{x}$ such that $\text{Prob}\{\|\nabla f(\bar{x})\|^2 \leq \epsilon\} \geq 1 - \Lambda$ for some $\epsilon > 0$ and $\Lambda \in (0, 1)$, can be bounded by

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

[1] It should not be too surprising to see that the complexity for the stochastic case is much worse than that for the deterministic case. For example, in the convex case, it is known [27, 18] that the complexity for finding an solution $\bar{x}$ satisfying $f(\bar{x}) - f^* \leq \epsilon$ will be substantially increased from $\mathcal{O}(1/\sqrt{\epsilon})$ to $\mathcal{O}(1/\epsilon^2)$ as one moves from the deterministic to stochastic setting.

We further show that, under certain light-tail assumption about the $\mathcal{SFO}$, the above complexity bound can be reduced to

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

Thirdly, we specialize the RSG method for the case where only stochastic zeroth-order information is available. There exists a somewhat long history for the development of zeroth-order (or derivative-free) methods in nonlinear programming (see the monograph by Conn et al. [5] and references therein). However, only few complexity results are available for these types of methods, mostly for convex programming (e.g., [25, 29]) and deterministic nonconvex programming problems (e.g., [3, 9, 29, 41]). The stochastic zeroth-order methods studied in this paper are directly motivated by a recent important work due to Nesterov [29]. More specifically, Nesterov proved in [29] some tight bounds for approximating first-order information by zeroth-order information using the Gaussian smoothing technique (see Theorem 3.1). Based on this technique, he presented a series of new complexity results for zeroth-order methods. For example, he established the $\mathcal{O}(n/\epsilon)$ complexity, in terms of $\mathbb{E}[f(\bar{x})-f^*]\leq\epsilon$, for a zeroth-order method applied to smooth convex programming problems (see in p.19 of [29]) along with some possible acceleration schemes. Here the expectation is taken with respect to the Gaussian random variables used in the algorithms. He had also proved the $\mathcal{O}(n/\epsilon)$ complexity, in terms of $\mathbb{E}[\|\nabla f(\bar{x})\|^2]\leq\epsilon$, for solving smooth nonconvex problems (see p.24 of [29]). While these bounds were obtained for solving deterministic optimization problems, Nesterov established the $\mathcal{O}(n^2/\epsilon^2)$ complexity, in terms of $\mathbb{E}[f(\bar{x})-f^*]\leq\epsilon$, for solving general nonsmooth convex SP problems (see p.17 of [29]).

By incorporating the Gaussian smoothing technique [29] into the RSG method, we present a randomized stochastic gradient free (RSGF) method for solving a class of simulation-based optimization problems and demonstrate that its iteration complexity for finding the aforementioned ϵ-solution (i.e., $\mathbb{E}[\|\nabla f(\bar{x})\|^2]\leq\epsilon$) can be bounded by $\mathcal{O}(n/\epsilon^2)$. To the best of our knowledge, this appears to be the first complexity result for nonconvex stochastic zeroth-order methods in the literature. Moreover, the same RSGF algorithm possesses an $\mathcal{O}(n/\epsilon^2)$ complexity bound, in terms of $\mathbb{E}[f(\bar{x})-f^*]\leq\epsilon$, for solving smooth convex SP problems. It is interesting to observe that this bound has a much weaker dependence on n than the one previously established by Nesterov for solving general nonsmooth convex SP problems (see p.17 of [29]). Such an improvement is obtained by explicitly making use of the smoothness properties of the objective function and carefully choosing the stepsizes and smoothing parameter used in the RSGF method.

This paper is organized as follows. We introduce two stochastic first-order methods, i.e., the RSG and 2-RSG methods, for nonconvex SP, and establish their convergence properties in Section 2. We then specialize these methods for solving a class of simulation-based optimization problems in Section 3. Some brief concluding remarks are also presented in Section 4.

1.1. Notation and terminology. As stated in [27], we say that $f\in\mathcal{C}_L^{1,1}(\mathbb{R}^n)$ if it is differentiable and

$$\|\nabla f(y)-\nabla f(x)\|\leq L\|y-x\|,\quad\forall x,y\in\mathbb{R}^n.$$

Clearly, we have

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^n. \tag{1.6}$$

If, in addition, $f(\cdot)$ is convex, then

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \geq \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2, \tag{1.7}$$

and

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2, \quad \forall x, y \in \mathbb{R}^n. \tag{1.8}$$

2. Stochastic first-order methods. Our goal in this section is to present and analyze a new class of SA algorithms for solving general smooth nonlinear (possibly nonconvex) SP problems. More specifically, we present the RSG method and establish its convergence properties in Subsection 2.1, and then introduce the 2-RSG method which can significantly improve the large-deviation properties of the RSG method in Subsection 2.2.

We assume throughout this section that Assumption A1 holds. In some cases, Assumption A1 is augmented by the following "light-tail" assumption.

A2: For any $x \in \mathbb{R}^n$ and $k \geq 1$, we have

$$\mathbb{E}\left[\exp\{\|G(x, \xi_k) - g(x)\|^2/\sigma^2\}\right] \leq \exp\{1\}. \tag{2.1}$$

It can be easily seen that Assumption A2 implies Assumption A1.b) by Jensen's inequality.

2.1. The randomized stochastic gradient method. The convergence of existing SA methods requires $f(\cdot)$ to be convex [24, 19, 18, 11, 10]. Moreover, in order to guarantee the convexity of $f(\cdot)$, one often need to assume that the random variables ξ_k, $k \geq 1$, to be independent of the search sequence $\{x_k\}$. Below we present a new SA-type algorithm that can deal with both convex and nonconvex SP problems, and allow random noises to be dependent on the search sequence. This algorithm is obtained by incorporating a certain randomization scheme into the classical SA method.

A randomized stochastic gradient (RSG) method

Input: Initial point x_1, iteration limit N, stepsizes $\{\gamma_k\}_{k\geq 1}$ and probability mass function $P_R(\cdot)$ supported on $\{1, \ldots, N\}$.

Step 0. Let R be a random variable with probability mass function P_R.

Step $k = 1, \ldots, R$. Call the stochastic first-order oracle for computing $G(x_k, \xi_k)$ and set

$$x_{k+1} = x_k - \gamma_k G(x_k, \xi_k). \tag{2.2}$$

Output x_R.

A few remarks about the above RSG method are in order. Firstly, in comparison with the classical SA, we have used a random iteration count, R, to terminate the execution of the RSG algorithm. Equivalently, one can view such a randomization

scheme from a slightly different perspective described as follows. Instead of terminating the algorithm at the R-th step, one can also run the RSG algorithm for N iterations but randomly choose a search point x_R (according to P_R) from its trajectory as the output of the algorithm. Clearly, using the latter scheme, we just need to run the algorithm for the first R iterations and the remaining $N - R$ iterations are surpluses. Note however, that the primary goal to introduce the random iteration count R is to derive new complexity results for nonconvex SP, rather than save the computational efforts in the last $N - R$ iterations of the algorithm. Indeed, if R is uniformly distributed, the computational gain from such a randomization scheme is simply a factor of 2. Secondly, the RSG algorithm described above is conceptual only because we have not specified the selection of the stepsizes $\{\gamma_k\}$ and the probability mass function P_R yet. We will address this issue after establishing some basic convergence properties of the RSG method.

The following result describes some convergence properties of the RSG method.

THEOREM 2.1. *Suppose that the stepsizes $\{\gamma_k\}$ and the probability mass function $P_R(\cdot)$ in the RSG method are chosen such that $\gamma_k < 2/L$ and*

$$P_R(k) := \text{Prob}\{R = k\} = \frac{2\gamma_k - L\gamma_k^2}{\sum_{k=1}^N (2\gamma_k - L\gamma_k^2)}, \quad k = 1, ..., N. \tag{2.3}$$

Then, under Assumption A1,

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

$$\frac{1}{L}\mathbb{E}[\|\nabla f(x_R)\|^2] \leq \frac{D_f^2 + \sigma^2 \sum_{k=1}^N \gamma_k^2}{\sum_{k=1}^N (2\gamma_k - L\gamma_k^2)}, \tag{2.4}$$

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

$$D_f := \left[\frac{2\,(f(x_1) - f^*)}{L}\right]^{\frac{1}{2}}, \tag{2.5}$$

and f^* denotes the optimal value of problem (1.1);

b) if, in addition, problem (1.1) is convex with an optimal solution x^*, then, for any $N \geq 1$,

$$\mathbb{E}[f(x_R) - f^*] \leq \frac{D_X^2 + \sigma^2 \sum_{k=1}^N \gamma_k^2}{\sum_{k=1}^N (2\gamma_k - L\gamma_k^2)}, \tag{2.6}$$

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

$$D_X := \|x_1 - x^*\|. \tag{2.7}$$

Proof. Display $\delta_k \equiv G(x_k, \xi_k) - \nabla f(x_k)$, $k \geq 1$. We first show part a). Using the assumption that $f \in \mathcal{C}_L^{1,1}(\mathbb{R}^n)$, (1.6) and (2.2), we have, for any $k = 1, \ldots, N$,

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\gamma_k^2 \|G(x_k, \xi_k)\|^2 \\
&= f(x_k) - \gamma_k \langle \nabla f(x_k), G(x_k, \xi_k) \rangle + \frac{L}{2}\gamma_k^2 \|G(x_k, \xi_k)\|^2 \\
&= f(x_k) - \gamma_k \|\nabla f(x_k)\|^2 - \gamma_k \langle \nabla f(x_k), \delta_k \rangle + \frac{L}{2}\gamma_k^2 \left[\|\nabla f(x_k)\|^2 + 2\langle \nabla f(x_k), \delta_k \rangle + \|\delta_k\|^2\right] \\
&= f(x_k) - \left(\gamma_k - \frac{L}{2}\gamma_k^2\right) \|\nabla f(x_k)\|^2 - \left(\gamma_k - L\gamma_k^2\right) \langle \nabla f(x_k), \delta_k \rangle + \frac{L}{2}\gamma_k^2 \|\delta_k\|^2.
\end{aligned} \tag{2.8}$$

Summing up the above inequalities and re-arranging the terms, we obtain

$$\sum_{k=1}^{N}\left(\gamma_k-\frac{L}{2}\gamma_k^2\right)\|\nabla f(x_k)\|^2 \le f(x_1)-f(x_{N+1})-\sum_{k=1}^{N}\left(\gamma_k-L\gamma_k^2\right)\langle\nabla f(x_k),\delta_k\rangle+\frac{L}{2}\sum_{k=1}^{N}\gamma_k^2\|\delta_k\|^2$$
$$\le f(x_1)-f^*-\sum_{k=1}^{N}\left(\gamma_k-L\gamma_k^2\right)\langle\nabla f(x_k),\delta_k\rangle+\frac{L}{2}\sum_{k=1}^{N}\gamma_k^2\|\delta_k\|^2, \quad (2.9)$$

where the last inequality follows from the fact that $f(x_{N+1}) \ge f^*$. Note that the search point x_k is a function of the history $\xi_{[k-1]}$ of the generated random process and hence is random. Taking expectations (with respect to $\xi_{[N]}$) on both sides of (2.9) and noting that under Assumption A1, $\mathbb{E}[\|\delta_k\|^2] \le \sigma^2$, and

$$\mathbb{E}[\langle\nabla f(x_k),\delta_k\rangle|\xi_{[k-1]}]=0, \quad (2.10)$$

we obtain

$$\sum_{k=1}^{N}\left(\gamma_k-\frac{L}{2}\gamma_k^2\right)\mathbb{E}_{\xi_{[N]}}\|\nabla f(x_k)\|^2 \le f(x_1)-f^*+\frac{L\sigma^2}{2}\sum_{k=1}^{N}\gamma_k^2 \quad (2.11)$$

Dividing both sides of the above inequality by $L\sum_{k=1}^{N}\left(\gamma_k-L\gamma_k^2/2\right)$ and noting that

$$\mathbb{E}[\|\nabla f(x_R)\|^2]=\mathbb{E}_{R,\xi_{[N]}}[\|\nabla f(x_R)\|^2]=\frac{\sum_{k=1}^{N}\left(2\gamma_k-L\gamma_k^2\right)\mathbb{E}_{\xi_{[N]}}\|\nabla f(x_k)\|^2}{\sum_{k=1}^{N}\left(2\gamma_k-L\gamma_k^2\right)},$$

we conclude

$$\frac{1}{L}\mathbb{E}[\|\nabla f(x_R)\|^2] \le \frac{1}{\sum_{k=1}^{N}(2\gamma_k-L\gamma_k^2)}\left[\frac{2\left(f(x_1)-f^*\right)}{L}+\sigma^2\sum_{k=1}^{N}\gamma_k^2\right],$$

which, in view of (2.5), clearly implies (2.4).

We now show that part b) holds. Display $\omega_k \equiv \|x_k-x^*\|$. First observe that, for any $k=1,\dots,N$,

$$\begin{aligned}\omega_{k+1}^2 &= \|x_k-\gamma_k G(x_k,\xi_k)-x^*\|^2\\ &= \omega_k^2-2\gamma_k\langle G(x_k,\xi_k),x_k-x^*\rangle+\gamma_k^2\|G(x_k,\xi_k)\|^2\\ &= \omega_k^2-2\gamma_k\langle\nabla f(x_k)+\delta_k,x_k-x^*\rangle+\gamma_k^2\left(\|\nabla f(x_k)\|^2+2\langle\nabla f(x_k),\delta_k\rangle+\|\delta_k\|^2\right).\end{aligned}$$

Moreover, in view of (1.8) and the fact that $\nabla f(x^*)=0$, we have

$$\frac{1}{L}\|\nabla f(x_k)\|^2 \le \langle\nabla f(x_k),x_k-x^*\rangle. \quad (2.12)$$

Combining the above two relations, we obtain, for any $k=1,\dots,N$,

$$\begin{aligned}\omega_{k+1}^2 &\le \omega_k^2-(2\gamma_k-L\gamma_k^2)\langle\nabla f(x_k),x_k-x^*\rangle-2\gamma_k\langle x_k-\gamma_k\nabla f(x_k)-x^*,\delta_k\rangle+\gamma_k^2\|\delta_k\|^2\\ &\le \omega_k^2-(2\gamma_k-L\gamma_k^2)[f(x_k)-f^*]-2\gamma_k\langle x_k-\gamma_k\nabla f(x_k)-x^*,\delta_k\rangle+\gamma_k^2\|\delta_k\|^2,\end{aligned}$$

where the last inequality follows from the convexity of $f(\cdot)$ and the fact that $\gamma_k \le 2/L$. Summing up the above inequalities and re-arranging the terms, we have

$$\sum_{k=1}^{N}(2\gamma_k-L\gamma_k^2)[f(x_k)-f^*] \le \omega_1^2-\omega_{N+1}^2-2\sum_{k=1}^{N}\gamma_k\langle x_k-\gamma_k\nabla f(x_k)-x^*,\delta_k\rangle+\sum_{k=1}^{N}\gamma_k^2\|\delta_k\|^2$$
$$\le D_X^2-2\sum_{k=1}^{N}\gamma_k\langle x_k-\gamma_k\nabla f(x_k)-x^*,\delta_k\rangle+\sum_{k=1}^{N}\gamma_k^2\|\delta_k\|^2,$$

where the last inequality follows from (2.7) and the fact that $\omega_{N+1} \geq 0$. The rest of the proof is similar to that of part a) and hence the details are skipped.

■

We now describe a possible strategy for the selection of the stepsizes $\{\gamma_k\}$ in the RSG method. For the sake of simplicity, let us assume that a constant stepsize policy is used, i.e., $\gamma_k = \gamma$, $k = 1, \ldots, N$, for some $\gamma \in (0, 2/L)$. Note that the assumption of constant stepsizes does not hurt the efficiency estimate of the RSG method. The following corollary of Theorem 2.1 is obtained by appropriately choosing the parameter γ.

COROLLARY 2.2. *Suppose that the stepsizes $\{\gamma_k\}$ are set to*

$$\gamma_k = \min\left\{\frac{1}{L}, \frac{\tilde{D}}{\sigma\sqrt{N}}\right\}, k = 1, \ldots, N, \tag{2.13}$$

for some $\tilde{D} > 0$. Also assume that the probability mass function $P_R(\cdot)$ is set to (2.3). Then, under Assumption A1, we have

$$\frac{1}{L}\mathbb{E}[\|\nabla f(x_R)\|^2] \leq \mathcal{B}_N := \frac{LD_f^2}{N} + \left(\tilde{D} + \frac{D_f^2}{\tilde{D}}\right)\frac{\sigma}{\sqrt{N}}, \tag{2.14}$$

where D_f is defined in (2.5). If, in addition, problem (1.1) is convex with an optimal solution x^*, then

$$\mathbb{E}[f(x_R) - f^*] \leq \frac{LD_X^2}{N} + \left(\tilde{D} + \frac{D_X^2}{\tilde{D}}\right)\frac{\sigma}{\sqrt{N}}, \tag{2.15}$$

where D_X is defined in (2.7).

Proof. Noting that by (2.13), we have

$$\begin{aligned}\frac{D_f^2 + \sigma^2\sum_{k=1}^N \gamma_k^2}{\sum_{k=1}^N (2\gamma_k - L\gamma_k^2)} &= \frac{D_f^2 + N\sigma^2\gamma_1^2}{N\gamma_1(2 - L\gamma_1)} \leq \frac{D_f^2 + N\sigma^2\gamma_1^2}{N\gamma_1} = \frac{D_f^2}{N\gamma_1} + \sigma^2\gamma_1 \\ &\leq \frac{D_f^2}{N}\max\left\{L, \frac{\sigma\sqrt{N}}{\tilde{D}}\right\} + \sigma^2\frac{\tilde{D}}{\sigma\sqrt{N}} \\ &\leq \frac{LD_f^2}{N} + \left(\tilde{D} + \frac{D_f^2}{\tilde{D}}\right)\frac{\sigma}{\sqrt{N}},\end{aligned}$$

which together with (2.4) then imply (2.14). Relation (2.15) follows similarly from the above inequality (with D_f replaced by D_X) and (2.6). ■

We now add a few remarks about the results obtained in Theorem 2.1 and Corollary 2.2. Firstly, as can be seen from (2.11), instead of randomly selecting a solution x_R from $\{x_1, \ldots, x_N\}$, another possibility would be to output the solution $\hat{x}_N$ such that

$$\|\nabla f(\hat{x}_N)\| = \min_{k=1,\ldots,N} \|\nabla f(x_k)\|. \tag{2.16}$$

We can show that $\mathbb{E}\|\nabla f(\hat{x}_N)\|$ goes to zero with similar rates of convergence as in (2.4) and (2.14). However, to use this strategy would require some extra computational effort to compute $\|\nabla f(x_k)\|$ for all $k = 1, \ldots, N$. Since $\|\nabla f(x_k)\|$ cannot be

computed exactly, to estimate them by using Monte-carlo simulation would incur additional approximation errors and raise some reliability issues. On the other hand, the above RSG method does not require any extra computational effort for estimating the gradients $\|\nabla f(x_k)\|$, $k = 1, \ldots, N$.

Secondly, observe that in the stepsize policy (2.13), we need to specify a parameter $\tilde{D}$. While the RSG method converges for any arbitrary $\tilde{D} > 0$, it can be easily seen from (2.14) and (2.15) that an optimal selection of $\tilde{D}$ would be D_f and D_X, respectively, for solving nonconvex and convex SP problems. With such selections, the bounds in (2.14) and (2.15), respectively, reduce to

$$\frac{1}{L}\mathbb{E}[\|\nabla f(x_R)\|^2] \leq \frac{LD_f^2}{N} + \frac{2D_f\sigma}{\sqrt{N}}. \tag{2.17}$$

and

$$\mathbb{E}[f(x_R) - f^*] \leq \frac{LD_X^2}{N} + \frac{2D_X\sigma}{\sqrt{N}}. \tag{2.18}$$

Note however, that the exact values of D_f or D_X are rarely known and one often need to set $\tilde{D}$ to a suboptimal value, e.g., certain upper bounds on D_f or D_X.

Thirdly, one possible drawback for the above RSG method is that one need to estimate L to obtain an upper bound on γ_k (see, e.g., (2.13)), which will also possibly affect the selection of P_R (see (2.3)). Note that similar requirements also exist for some deterministic first-order methods (e.g., gradient descent and Nesterov's accelerated gradient methods). While under the deterministic setting, one can somehow relax such requirements by using certain line-search procedures to enhance the practical performance of these methods, it is more difficult to devise similar line-search procedures for the stochastic setting, since the exact values of $f(x_k)$ and $\nabla f(x_k)$ are not available. It should be noted, however, that we do not need very accurate estimate for L in the RSG method. Indeed, it can be easily checked that the RSG method exhibits an $\mathcal{O}(1/\sqrt{N})$ rate of convergence if the stepsizes $\{\gamma_k\}$ are set to

$$\min\left\{\frac{1}{qL}, \frac{\tilde{D}}{\sigma\sqrt{N}}\right\}, \quad k = 1, \ldots, N$$

for any $q \in [1, \sqrt{N}]$. In other words, we can overestimate the value of L by a factor up to $\sqrt{N}$ and the resulting RSG method still exhibits similar rate of convergence. A common practice in stochastic optimization is to estimate L by using the stochastic gradients computed at a small number of trial points (see, e.g., [24, 19, 11, 10]). We have adopted such a strategy in our implementation of the RSG method as described in more details in the technical report associated with this paper [12]. It is also worth noting that, although in general the selection of P_R will depend on γ_k and hence on L, such a dependence is not necessary in some special cases. In particular, if the stepsizes $\{\gamma_k\}$ are chosen according to a constant stepsize policy (e.g., (2.13)), then R is uniformly distributed on $\{1, \ldots, N\}$.

Fourthly, it is interesting to note that the RSG method allows us to have a unified treatment for both nonconvex and convex SP problems in view of the specification of $\{\gamma_k\}$ and $P_R(\cdot)$ (c.f., (2.3) and (2.13)). Recall that the optimal rate of convergence for solving smooth convex SP problems is given by

$$\mathcal{O}\left(\frac{LD_X^2}{N^2} + \frac{D_X\sigma}{\sqrt{N}}\right).$$

This bound has been obtained by Lan [18] based on a stochastic counterpart of Nesterov's method [26, 27]. Comparing (2.18) with the above bound, the RSG method possesses a nearly optimal rate of convergence, since the second term in (2.18) is unimprovable while the first term in (2.18) can be much improved. Moreover, as shown by Cartis et al. [3], the first term in (2.17) for nonconvex problems is also unimprovable for gradient descent methods. It should be noted, however that the analysis in [3] applies only for gradient descent methods and does not show that the $\mathcal{O}(1/N)$ term is tight for all first-order methods.

Finally, observe that we can use different stepsize policy other than the constant one in (2.13). In particular, it can be shown that the RSG method with the following two stepsize policies will exhibit similar rates of convergence as those in Corollary 2.2.

- *Increasing stepsize policy*:

$$\gamma_k = \min\left\{\frac{1}{L}, \frac{\tilde{D}\sqrt{k}}{\sigma N}\right\}, k = 1, \ldots, N.$$

- *Decreasing stepsize policy*:

$$\gamma_k = \min\left\{\frac{1}{L}, \frac{\tilde{D}}{\sigma (kN)^{\frac{1}{4}}}\right\}, k = 1, \ldots, N.$$

Intuitively speaking, one may want to choose decreasing stepsizes which, according to the definition of $P_R(\cdot)$ in (2.3), can stop the algorithm earlier. On the other hand, as the algorithm moves forward and local information about the gradient gets better, choosing increasing stepsizes might be a better option. We expect that the practical performance of these stepsize policies will depend on each problem instance to be solved.

While Theorem 2.1 and Corollary 2.2 establish the expected convergence performance over many runs of the RSG method, we are also interested in the large-deviation properties for a single run of this method. In particular, we are interested in establishing its complexity for computing an (ϵ, Λ)-*solution* of problem (1.1), i.e., a point $\bar{x}$ satisfying $\text{Prob}\{\|\nabla f(\bar{x})\|^2 \leq \epsilon\} \geq 1 - \Lambda$ for some $\epsilon > 0$ and $\Lambda \in (0, 1)$. By using (2.14) and Markov's inequality, we have

$$\text{Prob}\left\{\|\nabla f(x_R)\|^2 \geq \lambda L \mathcal{B}_N\right\} \leq \frac{1}{\lambda}, \quad \forall \lambda > 0. \tag{2.19}$$

It then follows that the number of calls to $\mathcal{SFO}$ performed by the RSG method for finding an (ϵ, Λ)-solution, after disregarding a few constant factors, can be bounded by

$$\mathcal{O}\left\{\frac{1}{\Lambda\epsilon} + \frac{\sigma^2}{\Lambda^2\epsilon^2}\right\}. \tag{2.20}$$

The above complexity bound is rather pessimistic in terms of its dependence on Λ. We will investigate one possible way to significantly improve it in next subsection.

2.2. A two-phase randomized stochastic gradient method. In this section, we describe a variant of the RSG method which can considerably improve the complexity bound in (2.20). This procedure consists of two phases: an optimization phase used to generate a list of candidate solutions via a few independent runs of the

RSG method and a post-optimization phase in which a solution is selected from this candidate list.

A two-phase RSG (2-RSG) method

Input: Initial point x_1, number of runs S, iteration limit N, and sample size T.

Optimization phase:

For $s = 1, \ldots, S$

Call the RSG method with input x_1, iteration limit N, stepsizes $\{\gamma_k\}$ in (2.13) and probability mass function P_R in (2.3). Let $\bar{x}_s$ be the output of this procedure.

Post-optimization phase:

Choose a solution $\bar{x}^*$ from the candidate list $\{\bar{x}_1, \ldots, \bar{x}_S\}$ such that

$$\|g(\bar{x}^*)\| = \min_{s=1,\ldots,S} \|g(\bar{x}_s)\|, \quad g(\bar{x}_s) := \frac{1}{T}\sum_{k=1}^{T} G(\bar{x}_s, \xi_k), \tag{2.21}$$

where $G(x, \xi_k)$, $k = 1, \ldots, T$, are the stochastic gradients returned by the $\mathcal{SFO}$.

Observe that in (2.21), we define the best solution $\bar{x}^*$ as the one with the smallest value of $\|g(\bar{x}_s)\|$, $s = 1, \ldots, S$. Alternatively, one can choose $\bar{x}^*$ from $\{\bar{x}_1, \ldots, \bar{x}_S\}$ such that

$$\tilde{f}(\bar{x}^*) = \min_{1,\ldots,S} \tilde{f}(\bar{x}_s), \quad \tilde{f}(\bar{x}_s) = \frac{1}{T}\sum_{k=1}^{T} F(\bar{x}_s, \xi_k). \tag{2.22}$$

It should be noted that the 2-RSG method is different from a two-phase procedure for convex stochastic programming by Nesterov and Vial [30], where the average of $\bar{x}_1, \ldots, \bar{x}_S$ is chosen as the output solution.

In the 2-RSG method described above, the number of calls to the $\mathcal{SFO}$ are given by $S \times N$ and $S \times T$, respectively, for the optimization phase and post-optimization phase. Also note that we can possibly recycle the same sequence $\{\xi_k\}$ across all gradient estimations in the post-optimization phase of 2-RSG method. We will provide in Theorem 2.4 below certain bounds on S, N and T, to compute an (ϵ, Λ)-solution of problem (1.1).

We need the following results regarding the large deviations of vector valued martingales (see, e.g., Theorem 2.1 of [15]).

LEMMA 2.3. *Assume that we are given a polish space with Borel probability measure μ and a sequence of $\mathcal{F}_0 = \{\emptyset, \Omega\} \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots$ of σ-sub-algebras of Borel σ-algebra of Ω. Let $\zeta_i \in \mathbb{R}^n$, $i = 1, \ldots, \infty$, 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 $\mathbb{E}[\cdot|i]$, $i = 1, 2, \ldots$, 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 $\mathbb{E}\left\{\exp\left(\|\zeta_i\|^2/\sigma_i^2\right)|i-1\right\} \leq \exp(1)$ *almost surely for any* $i \geq 1$*, then*

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

We are now ready to describe the main convergence properties of the 2-RSG method. More specifically, Theorem 2.4.a) below shows the convergence rate of this algorithm for a given set of parameters (S, N, T), while Theorem 2.4.b) establishes the complexity of the 2-RSG method for computing an (ϵ, Λ)-solution of problem (1.1).

THEOREM 2.4. *Under Assumption A1, the following statements hold for the 2-RSG method applied to problem (1.1).*

a) Let $\mathcal{B}_N$ *be defined in (2.14). We have*

$$\text{Prob}\left\{\|\nabla f(\bar{x}^*)\|^2 \geq 2\left(4L\mathcal{B}_N + \frac{3\lambda\sigma^2}{T}\right)\right\} \leq \frac{S+1}{\lambda} + 2^{-S}, \quad \forall\, \lambda > 0; \tag{2.23}$$

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

$$S = S(\Lambda) := \lceil \log(2/\Lambda) \rceil, \tag{2.24}$$

$$N = N(\epsilon) := \left\lceil \max\left\{\frac{32L^2D_f^2}{\epsilon}, \left\lceil 32L\left(\tilde{D} + \frac{D_f^2}{\tilde{D}}\right)\frac{\sigma}{\epsilon}\right\rceil^2\right\}\right\rceil, \tag{2.25}$$

$$T = T(\epsilon, \Lambda) := \left\lceil \frac{24(S+1)\sigma^2}{\Lambda\epsilon}\right\rceil, \tag{2.26}$$

then the 2-RSG method can compute an (ϵ, Λ)*-solution of problem (1.1) after taking at most*

$$S(\Lambda)\ [N(\epsilon) + T(\epsilon, \Lambda)] \tag{2.27}$$

calls to the stochastic first-order oracle.

Proof. We first show part a). Observe that by the definition of $\bar{x}^*$ in (2.21), we have

$$\begin{aligned}\|g(\bar{x}^*)\|^2 &= \min_{s=1,\ldots,S}\|g(\bar{x}_s)\|^2 = \min_{s=1,\ldots,S}\|\nabla f(\bar{x}_s) + g(\bar{x}_s) - \nabla f(\bar{x}_s)\|^2 \\ &\leq \min_{s=1,\ldots,S}\left\{2\|\nabla f(\bar{x}_s)\|^2 + 2\|g(\bar{x}_s) - \nabla f(\bar{x}_s)\|^2\right\} \\ &\leq 2\min_{s=1,\ldots,S}\|\nabla f(\bar{x}_s)\|^2 + 2\max_{s=1,\ldots,S}\|g(\bar{x}_s) - \nabla f(\bar{x}_s)\|^2,\end{aligned}$$

which implies that

$$\begin{aligned}\|\nabla f(\bar{x}^*)\|^2 \leq 2\|g(\bar{x}^*)\|^2 + 2\|\nabla f(\bar{x}^*) - g(\bar{x}^*)\|^2 &\leq 4\min_{s=1,\ldots,S}\|\nabla f(\bar{x}_s)\|^2 \\ &+ 4\max_{s=1,\ldots,S}\|g(\bar{x}_s) - \nabla f(\bar{x}_s)\|^2 + 2\|\nabla f(\bar{x}^*) - g(\bar{x}^*)\|^2. \end{aligned}\tag{2.28}$$

We now provide certain probabilistic upper bounds to the three terms in the right hand side of the above inequality. Firstly, using the fact that $\bar{x}_s$, $1 \leq s \leq S$, are independent and relation (2.19) (with $\lambda = 2$), we have

$$\text{Prob}\left\{\min_{s=1,\ldots,S}\|\nabla f(\bar{x}_s)\|^2 \geq 2L\mathcal{B}_N\right\} = \prod_{s=1}^{S}\text{Prob}\left\{\|\nabla f(\bar{x}_s)\|^2 \geq 2L\mathcal{B}_N\right\} \leq 2^{-S}. \tag{2.29}$$

Moreover, denoting $\delta_{s,k} = G(\bar{x}_s, \xi_k) - \nabla f(\bar{x}_s)$, $k = 1, \ldots, T$, we have $g(\bar{x}_s) - \nabla f(\bar{x}_s) = \sum_{k=1}^{T} \delta_{s,k}/T$. Using this observation, Assumption A1 and Lemma 2.3.a), we conclude that, for any $s = 1, \ldots, S$,

$$\text{Prob}\left\{\|g(\bar{x}_s) - \nabla f(\bar{x}_s)\|^2 \geq \frac{\lambda\sigma^2}{T}\right\} = \text{Prob}\left\{\|\sum_{k=1}^{T} \delta_{s,k}\|^2 \geq \lambda T\sigma^2\right\} \leq \frac{1}{\lambda}, \ \forall \lambda > 0,$$

which implies that

$$\text{Prob}\left\{\max_{s=1,\ldots,S} \|g(\bar{x}_s) - \nabla f(\bar{x}_s)\|^2 \geq \frac{\lambda\sigma^2}{T}\right\} \leq \frac{S}{\lambda}, \ \forall \lambda > 0, \tag{2.30}$$

and that

$$\text{Prob}\left\{\|g(\bar{x}^*) - \nabla f(\bar{x}^*)\|^2 \geq \frac{\lambda\sigma^2}{T}\right\} \leq \frac{1}{\lambda}, \ \forall \lambda > 0. \tag{2.31}$$

The result then follows by combining relations (2.28), (2.29), (2.30) and (2.31).

We now show that part b) holds. Since the 2-RSG method needs to call the RSG method S times with iteration limit $N(\epsilon)$ in the optimization phase, and estimate the gradients $g(\bar{x}_s)$, $s = 1, \ldots, S$ with sample size $T(\epsilon)$ in the post-optimization phase, the total number of calls to the stochastic first-order oracle is bounded by $S[N(\epsilon) + T(\epsilon)]$. It remains to show that $\bar{x}^*$ is an (ϵ, Λ)-solution of problem (1.1). Noting that by the definitions of $\mathcal{B}_N$ and $N(\epsilon)$, respectively, in (2.14) and (2.25), we have

$$\mathcal{B}_{N(\epsilon)} = \frac{LD_f^2}{N(\epsilon)} + \left(\tilde{D} + \frac{D_f^2}{\tilde{D}}\right)\frac{\sigma}{\sqrt{N(\epsilon)}} \leq \frac{\epsilon}{32L} + \frac{\epsilon}{32L} = \frac{\epsilon}{16L}.$$

Using the above observation, (2.26) and setting $\lambda = [2(S+1)]/\Lambda$ in (2.23), we have

$$4LB_{N(\epsilon)} + \frac{3\lambda\sigma^2}{T(\epsilon)} = \frac{\epsilon}{4} + \frac{\lambda\Lambda\epsilon}{8(S+1)} = \frac{\epsilon}{2},$$

which, together with relations (2.23) and (2.24), and the selection of λ, then imply that

$$\text{Prob}\left\{\|\nabla f(\bar{x}^*)\|^2 \geq \epsilon\right\} \leq \frac{\Lambda}{2} + 2^{-S} \leq \Lambda.$$

■

It is interesting to compare the complexity bound in (2.27) with the one in (2.20). In view of (2.24), (2.25) and (2.26), the complexity bound in (2.27), after disregarding a few constant factors, is equivalent to

$$\mathcal{O}\left\{\frac{\log(1/\Lambda)}{\epsilon} + \frac{\sigma^2}{\epsilon^2}\log\frac{1}{\Lambda} + \frac{\log^2(1/\Lambda)\sigma^2}{\Lambda\epsilon}\right\}. \tag{2.32}$$

The above bound can be considerably smaller than the one in (2.20) up to a factor of $1/\left[\Lambda^2\log(1/\Lambda)\right]$, when the second terms are the dominating ones in both bounds.

The following result shows that the bound (2.27) obtained in Theorem 2.4 can be further improved under certain light-tail assumption of $\mathcal{SFO}$.

COROLLARY 2.5. *Under Assumptions A1 and A2, the following statements hold for the 2-RSG method applied to problem (1.1).*

a) Let $\mathcal{B}_N$ is defined in (2.14). We have, $\forall\, \lambda > 0$,

$$\text{Prob}\left\{\|\nabla f(\bar{x}^*)\|^2 \geq 4\left[2L\mathcal{B}_N + 3(1+\lambda)^2\frac{\sigma^2}{T}\right]\right\} \leq (S+1)\exp(-\lambda^2/3) + 2^{-S}; \tag{2.33}$$

b) Let $\epsilon > 0$ and $\Lambda \in (0,1)$ be given. If S and N are set to $S(\Lambda)$ and $N(\epsilon)$ as in (2.24) and (2.25), respectively, and the sample size T is set to

$$T = T'(\epsilon,\Lambda) := \frac{24\sigma^2}{\epsilon}\left[1 + \left(3\ln\frac{2(S+1)}{\Lambda}\right)^{\frac{1}{2}}\right]^2, \tag{2.34}$$

then the 2-RSG method can compute an (ϵ,Λ)-solution of problem (1.1) in at most

$$S(\Lambda)\left[N(\epsilon) + T'(\epsilon,\Lambda)\right] \tag{2.35}$$

calls to the stochastic first-order oracle.

Proof. We provide the proof of part a) only, since part b) follows immediately from part a) and an argument similar to the one used in the proof of Theorem 2.4.b). Denoting $\delta_{s,k} = G(\bar{x}_s,\xi_k) - \nabla f(\bar{x}_s)$, $k = 1,\ldots,T$, we have $g(\bar{x}_s) - \nabla f(\bar{x}_s) = \sum_{k=1}^T \delta_{s,k}/T$. Using this observation, Assumption A2 and Lemma 2.3.b), we conclude that, for any $s = 1,\ldots,S$ and $\lambda > 0$,

$$\begin{aligned}&\text{Prob}\left\{\|g(\bar{x}_s) - \nabla f(\bar{x}_s)\|^2 \geq 2(1+\lambda)^2\tfrac{\sigma^2}{T}\right\}\\ &= \text{Prob}\left\{\|\textstyle\sum_{k=1}^T \delta_{s,k}\| \geq \sqrt{2T}(1+\lambda)\sigma\right\} \leq \exp(-\lambda^2/3),\end{aligned}$$

which implies that

$$\text{Prob}\left\{\max_{s=1,\ldots,S}\|g(\bar{x}_s) - \nabla f(\bar{x}_s)\|^2 \geq 2(1+\lambda)^2\frac{\sigma^2}{T}\right\} \leq S\exp(-\lambda^2/3),\ \forall\lambda > 0. \tag{2.36}$$

and that

$$\text{Prob}\left\{\|g(\bar{x}^*) - \nabla f(\bar{x}^*)\|^2 \geq 2(1+\lambda)^2\frac{\sigma^2}{T}\right\} \leq \exp(-\lambda^2/3),\ \forall\lambda > 0. \tag{2.37}$$

The result in part a) then follows by combining relations (2.28), (2.29), (2.36) and (2.37). ∎

In view of (2.24), (2.25) and (2.34), the bound in (2.35), after disregarding a few constant factors, is equivalent to

$$\mathcal{O}\left\{\frac{\log(1/\Lambda)}{\epsilon} + \frac{\sigma^2}{\epsilon^2}\log\frac{1}{\Lambda} + \frac{\log^2(1/\Lambda)\sigma^2}{\epsilon}\right\}. \tag{2.38}$$

Clearly, the third term of the above bound is significantly smaller than the corresponding one in (2.32) by a factor of $1/\Lambda$.

3. Stochastic zeroth-order methods. Our problem of interest in this section is problem (1.1) with f given in (1.4), i.e.,

$$f^* := \inf_{x\in\mathbb{R}^n}\left\{f(x) := \int_\Xi F(x,\xi)dP(\xi)\right\}. \tag{3.1}$$

Moreover, we assume that $F(x,\xi) \in \mathcal{C}_L^{1,1}(\mathbb{R}^n)$ almost surely, which clearly implies $f(x) \in \mathcal{C}_L^{1,1}(\mathbb{R}^n)$. Our goal in this section is to specialize the RSG and 2-RSG method, respectively, in Subsections 3.1 and 3.2, to deal with the situation when only stochastic zeroth-order information of f is available.

3.1. The randomized stochastic gradient free method. Throughout this section, we assume that f is represented by a *stochastic zeroth-order oracle ($\mathcal{SZO}$)*. More specifically, at the k-th iteration, x_k and ξ_k being the input, the $\mathcal{SZO}$ outputs the quantity $F(x_k,\xi_k)$ such that the following assumption holds:

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

$$\mathbb{E}[F(x_k,\xi_k)] = f(x_k). \tag{3.2}$$

To exploit zeroth-order information, we consider a smooth approximation of the objective function f. It is well-known (see, e.g., [35], [6] and [43]) that the convolution of f with any nonnegative, measurable and bounded function $\psi : \mathbb{R}^n \to \mathbb{R}$ satisfying $\int_{\mathbb{R}^n} \psi(u)\,du = 1$ is an approximation of f which is at least as smooth as f. One of the most important examples of the function ψ is the probability density function. Here, we use the Gaussian distribution in the convolution. Let u be n-dimensional standard Gaussian random vector and $\mu > 0$ be the smoothing parameter. Then, a smooth approximation of f is defined as

$$f_\mu(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} \int f(x+\mu u) e^{-\frac{1}{2}\|u\|^2}\,du = \mathbb{E}_u[f(x+\mu u)]. \tag{3.3}$$

The following result due to Nesterov [29] describes some properties of $f_\mu(\cdot)$.

THEOREM 3.1. *The following statements hold for any $f \in \mathcal{C}_L^{1,1}$.*

a) The gradient of f_μ given by

$$\nabla f_\mu(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} \int \frac{f(x+\mu u) - f(x)}{\mu} u e^{-\frac{1}{2}\|u\|^2}\,du, \tag{3.4}$$

is Lipschitz continuous with constant L_μ such that $L_\mu \leq L$;

b) For any $x \in \mathbb{R}^n$,

$$|f_\mu(x) - f(x)| \leq \frac{\mu^2}{2} Ln, \tag{3.5}$$

$$\|\nabla f_\mu(x) - \nabla f(x)\| \leq \frac{\mu}{2} L(n+3)^{\frac{3}{2}}; \tag{3.6}$$

c) For any $x \in \mathbb{R}^n$,

$$\frac{1}{\mu^2}\mathbb{E}_u[\{f(x+\mu u) - f(x)\}^2 \|u\|^2] \leq \frac{\mu^2}{2} L^2 (n+6)^3 + 2(n+4)\|\nabla f(x)\|^2. \tag{3.7}$$

It immediately follows from (3.6) that

$$\|\nabla f_\mu(x)\|^2 \leq 2\|\nabla f(x)\|^2 + \frac{\mu^2}{2} L^2 (n+3)^3, \tag{3.8}$$

$$\|\nabla f(x)\|^2 \leq 2\|\nabla f_\mu(x)\|^2 + \frac{\mu^2}{2} L^2 (n+3)^3. \tag{3.9}$$

Moreover, denoting

$$f_\mu^* := \min_{x\in\mathbb{R}^n} f_\mu(x), \tag{3.10}$$

we conclude from (3.5) that $|f_\mu^* - f^*| \le \mu^2 Ln/2$ and hence that

$$-\mu^2 Ln \le [f_\mu(x) - f_\mu^*] - [f(x) - f^*] \le \mu^2 Ln. \tag{3.11}$$

Below we modify the RSG method in subsection (2.1) to use stochastic zeroth-order rather than first-order information for solving problem (3.1).

A randomized stochastic gradient free (RSGF) method

Input: Initial point x_1, iteration limit N, stepsizes $\{\gamma_k\}_{k\ge1}$, probability mass function $P_R(\cdot)$ supported on $\{1,\dots,N\}$.

Step 0. Let R be a random variable with probability mass function P_R.

Step $k = 1,\dots,R$. Generate u_k by Gaussian random vector generator and call the stochastic zeroth-order oracle for computing $G_\mu(x_k,\xi_k,u_k)$ given by

$$G_\mu(x_k,\xi_k,u_k) = \frac{F(x_k+\mu u_k,\xi_k) - F(x_k,\xi_k)}{\mu} u_k. \tag{3.12}$$

Set

$$x_{k+1} = x_k - \gamma_k G_\mu(x_k,\xi_k,u_k). \tag{3.13}$$

Output x_R.

Note that the esimator $G_\mu(x_k,\xi_k,u_k)$ of $\nabla f_\mu(x_k)$ in (3.12) was suggested by Nesterov in [29]. Indeed, by (3.4) and Assumption A3, we have

$$\mathbb{E}_{\xi,u}[G_\mu(x,\xi,u)] = \mathbb{E}_u\left[\mathbb{E}_\xi[G_\mu(x,\xi,u)|u]\right] = \nabla f_\mu(x), \tag{3.14}$$

which implies that $G_\mu(x,\xi,u)$ is an unbiased estimator of $\nabla f_\mu(x)$. Hence, if the variance $\tilde\sigma^2 \equiv \mathbb{E}_{\xi,u}[\|G_\mu(x,\xi,u) - \nabla f_\mu(x)\|^2]$ is bounded, we can directly apply the convergence results in Theorem 2.1 to the above RSGF method. However, there still exist a few problems in this approach. Firstly, we do not know an explicit expression of the bound $\tilde\sigma^2$. Secondly, this approach does not provide any information regarding how to appropriately specify the smoothing parameter μ. The latter issue is critical for the implementation of the RSGF method.

By applying the approximation results in Theorem 3.1 to the functions $F(\cdot,\xi_k)$, $k = 1,\dots,N$, and using a slightly different convergence analysis than the one in Theorem 2.1, we are able to obtain much refined convergence results for the above RSGF method.

THEOREM 3.2. *Suppose that the stepsizes $\{\gamma_k\}$ and the probability mass function $P_R(\cdot)$ in the RSGF method are chosen such that $\gamma_k < 1/[2(n+4)L]$ and*

$$P_R(k) := \mathrm{Prob}\{R=k\} = \frac{\gamma_k - 2L(n+4)\gamma_k^2}{\sum_{k=1}^N [\gamma_k - 2L(n+4)\gamma_k^2]}, \quad k = 1,...,N. \tag{3.15}$$

Then, under Assumptions A1 and A3,

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

$$\begin{array}{l} \frac{1}{L}\mathbb{E}[\|\nabla f(x_R)\|^2] \leq \frac{1}{\sum_{k=1}^N \left[\gamma_k - 2L(n+4)\gamma_k^2\right]} \left[D_f^2 + 2\mu^2(n+4)\right. \\ \left. \left(1 + L(n+4)^2 \sum_{k=1}^N (\frac{\gamma_k}{4} + L\gamma_k^2)\right) + 2(n+4)\sigma^2 \sum_{k=1}^N \gamma_k^2\right], \end{array} \tag{3.16}$$

where the expectation is taken with respect to R, $\xi_{[N]}$ and $u_{[N]}$, and D_f is defined in(2.5);

b) if, in addition, problem (3.1) is convex with an optimal solution x^*, then, for any $N \geq 1$,

$$\begin{array}{l} \mathbb{E}[f(x_R) - f^*] \leq \frac{1}{2\sum_{k=1}^N \left[\gamma_k - 2(n+4)L\gamma_k^2\right]} \left[D_X^2 + 2\mu^2 L(n+4)\right. \\ \left. \sum_{k=1}^N \left[\gamma_k + L(n+4)^2\gamma_k^2\right] + 2(n+4)\sigma^2 \sum_{k=1}^N \gamma_k^2\right], \end{array} \tag{3.17}$$

where the expectation is taken with respect to R, $\xi_{[N]}$ and $u_{[N]}$, and D_X is defined in (2.7).

Proof. Let $\zeta_k \equiv (\xi_k, u_k)$, $k \geq 1$, $\zeta_{[N]} := (\zeta_1, ..., \zeta_N)$, and $\mathbb{E}_{\zeta[N]}$ denote the expectation w.r.t. $\zeta_{[N]}$. Also denote $\Delta_k \equiv G_\mu(x_k, \xi_k, u_k) - \nabla f_\mu(x_k) \equiv G_\mu(x_k, \zeta_k) - \nabla f_\mu(x_k)$, $k \geq 1$. Using the fact that $f \in \mathcal{C}_L^{1,1}(\mathbb{R}^n)$, Theorem 3.1.a), (1.6) and (3.13), we have, for any $k = 1, \ldots, N$,

$$\begin{array}{rcl} f_\mu(x_{k+1}) & \leq & f_\mu(x_k) - \gamma_k \langle \nabla f_\mu(x_k), G_\mu(x_k, \zeta_k)\rangle + \frac{L}{2}\gamma_k^2 \|G_\mu(x_k, \zeta_k)\|^2 \\ & = & f_\mu(x_k) - \gamma_k \|\nabla f_\mu(x_k)\|^2 - \gamma_k \langle \nabla f_\mu(x_k), \Delta_k\rangle + \frac{L}{2}\gamma_k^2 \|G_\mu(x_k, \zeta_k)\|^2. \end{array} \tag{3.18}$$

Summing up these inequalities, re-arranging the terms and noting that $f_\mu^* \leq f_\mu(x_{N+1})$, we obtain

$$\sum_{k=1}^N \gamma_k \|\nabla f_\mu(x_k)\|^2 \leq f_\mu(x_1) - f_\mu^* - \sum_{k=1}^N \gamma_k \langle \nabla f_\mu(x_k), \Delta_k\rangle + \frac{L}{2}\sum_{k=1}^N \gamma_k^2 \|G_\mu(x_k, \zeta_k)\|^2. \tag{3.19}$$

Now, observe that by (3.14),

$$\mathbb{E}[\langle \nabla f_\mu(x_k), \Delta_k\rangle | \zeta_{[k-1]}] = 0. \tag{3.20}$$

and that by the assumption $F(\cdot, \xi_k) \in \mathcal{C}_L^{1,1}(\mathbb{R}^n)$, (3.7) (with $f = F(\cdot, \xi_k)$), and (3.12),

$$\begin{array}{rcl} \mathbb{E}[\|G_\mu(x_k, \zeta_k)\|^2 | \zeta_{[k-1]}] & \leq & 2(n+4)\mathbb{E}[\|G(x_k, \xi_k)\|^2 | \zeta_{[k-1]}] + \frac{\mu^2}{2}L^2(n+6)^3 \\ & \leq & 2(n+4)\left[\mathbb{E}[\|\nabla f(x_k)\|^2 | \zeta_{[k-1]}] + \sigma^2\right] + \frac{\mu^2}{2}L^2(n+6)^3, \end{array} \tag{3.21}$$

where the second inequality follows from Assumption A1. Taking expectations with respect to $\zeta_{[N]}$ on both sides of (3.19) and using the above two observations, we obtain

$$\begin{array}{l} \sum_{k=1}^N \gamma_k \mathbb{E}_{\zeta[N]}\left[\|\nabla f_\mu(x_k)\|^2\right] \leq f_\mu(x_1) - f_\mu^* \\ \quad + \frac{L}{2}\sum_{k=1}^N \gamma_k^2 \left\{2(n+4)\left[\mathbb{E}_{\zeta[N]}[\|\nabla f(x_k)\|^2] + \sigma^2\right] + \frac{\mu^2}{2}L^2(n+6)^3\right\}. \end{array}$$

The above conclusion together with (3.8) and (3.11) then imply that

$$\begin{array}{l} \sum_{k=1}^N \gamma_k \left[\mathbb{E}_{\zeta[N]}[\|\nabla f(x_k)\|^2] - \frac{\mu^2}{2}L^2(n+3)^3\right] \leq 2\left[f(x_1) - f^*\right] + 2\mu^2 Ln \\ +2L(n+4)\sum_{k=1}^N \gamma_k^2 \mathbb{E}_{\zeta[N]}[\|\nabla f(x_k)\|^2] + \left[2L(n+4)\sigma^2 + \frac{\mu^2}{2}L^3(n+6)^3\right]\sum_{k=1}^N \gamma_k^2. \end{array} \tag{3.22}$$

By re-arranging the terms and simplifying the constants, we have

$$\begin{aligned}
&\textstyle\sum_{k=1}^N \left\{\left[\gamma_k - 2L(n+4)\gamma_k^2\right] \mathbb{E}_{\zeta[N]}[\|\nabla f(x_k)\|^2]\right\} \\
&\leq 2\left[f(x_1) - f^*\right] + 2L(n+4)\sigma^2 \textstyle\sum_{k=1}^N \gamma_k^2 + 2\mu^2 Ln + \frac{\mu^2}{2} L^2 \sum_{k=1}^N \left[(n+3)^3\gamma_k + L(n+6)^3\gamma_k^2\right] \\
&\leq 2\left[f(x_1) - f^*\right] + 2L(n+4)\sigma^2 \textstyle\sum_{k=1}^N \gamma_k^2 + 2\mu^2 L(n+4)\left[1 + L(n+4)^2 \sum_{k=1}^N (\frac{\gamma_k}{4} + L\gamma_k^2)\right].
\end{aligned} \tag{3.23}$$

Dividing both sides of the above inequality by $\sum_{k=1}^N \left[\gamma_k - 2L(n+4)\gamma_k^2\right]$ and noting that

$$\mathbb{E}[\|\nabla f(x_R)\|^2] = \mathbb{E}_{R,\zeta[N]}[\|\nabla f(x_R)\|^2] = \frac{\sum_{k=1}^N \left\{\left[\gamma_k - 2L(n+4)\gamma_k^2\right] \mathbb{E}_{\zeta[N]}\|\nabla f(x_k)\|^2\right\}}{\sum_{k=1}^N \left[\gamma_k - 2L(n+4)\gamma_k^2\right]},$$

we obtain (3.16).

We now show part b). Denote $\omega_k \equiv \|x_k - x^*\|$. First observe that, for any $k = 1, \ldots, N$,

$$\begin{aligned}
\omega_{k+1}^2 &= \|x_k - \gamma_k G_\mu(x_k, \zeta_k) - x^*\|^2 \\
&= \omega_k^2 - 2\gamma_k \langle \nabla f_\mu(x_k) + \Delta_k, x_k - x^* \rangle + \gamma_k^2 \|G_\mu(x_k, \zeta_k)\|^2.
\end{aligned}$$

and hence that

$$\omega_{N+1}^2 = \omega_1^2 - 2\sum_{k=1}^N \gamma_k \left\langle \nabla f_\mu(x_k), x_k - x^* \right\rangle - 2\sum_{k=1}^N \gamma_k \left\langle \Delta_k, x_k - x^* \right\rangle + \sum_{k=1}^N \gamma_k^2 \left\| G_\mu(x_k, \zeta_k) \right\|^2.$$

Taking expectation w.r.t. $\zeta_{\zeta[N]}$ on both sides of the above equality, using relation (3.21) and noting that by (3.14), $\mathbb{E}[\langle \Delta_k, x_k - x^* \rangle | \zeta_{[k-1]}] = 0$, we obtain

$$\begin{aligned}
\mathbb{E}_{\zeta[N]}[\omega_{N+1}^2] &\leq \omega_1^2 - 2\sum_{k=1}^N \gamma_k \, \mathbb{E}_{\zeta[N]}[\langle \nabla f_\mu(x_k), x_k - x^* \rangle] + 2(n+4) \sum_{k=1}^N \gamma_k^2 \, \mathbb{E}_{\zeta[N]}[\|\nabla f(x_k)\|^2] \\
&\quad + \left[2(n+4)\sigma^2 + \frac{\mu^2}{2} L^2 (n+6)^3\right] \sum_{k=1}^N \gamma_k^2 \\
&\leq \omega_1^2 - 2\sum_{k=1}^N \gamma_k \, \mathbb{E}_{\zeta[N]}[f_\mu(x_k) - f_\mu(x^*)] + 2(n+4)L \sum_{k=1}^N \gamma_k^2 \, \mathbb{E}_{\zeta[N]}[f(x_k) - f^*] \\
&\quad + \left[2(n+4)\sigma^2 + \frac{\mu^2}{2} L^2 (n+6)^3\right] \sum_{k=1}^N \gamma_k^2 \\
&\leq \omega_1^2 - 2\sum_{k=1}^N \gamma_k \, \mathbb{E}_{\zeta[N]}\left[f(x_k) - f^* - \mu^2 Ln\right] + 2(n+4)L \sum_{k=1}^N \gamma_k^2 \, \mathbb{E}_{\zeta[N]}[f(x_k) - f^*] \\
&\quad + \left[2(n+4)\sigma^2 + \frac{\mu^2}{2} L^2 (n+6)^3\right] \sum_{k=1}^N \gamma_k^2,
\end{aligned}$$

where the second inequality follows from (2.12) and the convexity of f_μ, and the last inequality follows from (3.5). Re-arranging the terms in the above inequality, using

the facts that $\omega_{N+1}^2 \geq 0$ and $f(x_k) \geq f^*$, and simplifying the constants, we have

$$\begin{aligned}
&2\sum_{k=1}^{N}\left[\gamma_k - 2(n+4)L\gamma_k^2)\right]\mathbb{E}_{\zeta[N]}[f(x_k)-f^*] \\
&\leq 2\sum_{k=1}^{N}\left[\gamma_k - (n+4)L\gamma_k^2)\right]\mathbb{E}_{\zeta[N]}[f(x_k)-f^*] \\
&\leq \omega_1^2 + 2\mu^2 L(n+4)\sum_{k=1}^{N}\gamma_k + 2(n+4)\left[L^2\mu^2(n+4)^2+\sigma^2\right]\sum_{k=1}^{N}\gamma_k^2.
\end{aligned}$$

The rest of proof is similar to part a) and hence the details are skipped. ∎

Similarly to the RSG method, we can specialize the convergence results in Theorem 3.2 for the RSGF method with a constant stepsize policy.

COROLLARY 3.3. *Suppose that the stepsizes $\{\gamma_k\}$ are set to*

$$\gamma_k = \frac{1}{\sqrt{n+4}}\min\left\{\frac{1}{4L\sqrt{n+4}}, \frac{\tilde{D}}{\sigma\sqrt{N}}\right\}, \quad k=1,\ldots,N, \tag{3.24}$$

for some $\tilde{D} > 0$. Also assume that the probability mass function $P_R(\cdot)$ is set to (3.15) and μ is chosen such that

$$\mu \leq \frac{D_f}{(n+4)\sqrt{2N}} \tag{3.25}$$

where D_f and D_X are defined in (2.5) and (2.7), respectively. Then, under Assumptions A1 and A3, we have

$$\frac{1}{L}\mathbb{E}[\|\nabla f(x_R)\|^2] \leq \bar{\mathcal{B}}_N := \frac{12(n+4)LD_f^2}{N} + \frac{4\sigma\sqrt{n+4}}{\sqrt{N}}\left(\tilde{D} + \frac{D_f^2}{\tilde{D}}\right). \tag{3.26}$$

If, in addition, problem (3.1) is convex with an optimal solution x^* and μ is chosen such that

$$\mu \leq \frac{D_X}{\sqrt{(n+4)}},$$

then,

$$\mathbb{E}[f(x_R) - f^*] \leq \frac{5L(n+4)D_X^2}{N} + \frac{2\sigma\sqrt{n+4}}{\sqrt{N}}\left(\tilde{D} + \frac{D_X^2}{\tilde{D}}\right). \tag{3.27}$$

Proof. We prove (3.26) only since relation (3.27) can be shown by using similar arguments. First note that by (3.24), we have

$$\gamma_k \leq \frac{1}{4(n+4)L}, \quad k=1,\ldots,N, \tag{3.28}$$

$$\sum_{k=1}^{N}\left[\gamma_k - 2L(n+4)\gamma_k^2\right] = N\gamma_1\left[1-2L(n+4)\gamma_1\right] \geq \frac{N\gamma_1}{2}. \tag{3.29}$$

Therefore, using the above inequalities and (3.16), we obtain

$$\begin{aligned}\frac{1}{L}\mathbb{E}[\|\nabla f(x_R)\|^2] &\le \frac{2D_f^2+4\mu^2(n+4)}{N\gamma_1}+\mu^2L(n+4)^3+4(n+4)\left[\mu^2L^2(n+4)^2+\sigma^2\right]\gamma_1\\ &\le \frac{2D_f^2+4\mu^2(n+4)}{N}\max\left\{4L(n+4),\frac{\sigma\sqrt{(n+4)N}}{\tilde{D}}\right\}\\ &\quad+\mu^2L(n+4)^2\left[(n+4)+1\right]+\frac{4\sqrt{n+4}\tilde{D}\sigma}{\sqrt{N}},\end{aligned}$$

which, in view of (3.25), then implies that

$$\begin{aligned}\frac{1}{L}\mathbb{E}[\|\nabla f(x_R)\|^2] &\le \frac{2D_f^2}{N}\left[1+\frac{1}{(n+4)N}\right]\left[4L(n+4)+\frac{\sigma\sqrt{(n+4)N}}{\tilde{D}}\right]\\ &\quad+\frac{LD_f^2}{2N}\left[(n+4)+1\right]+\frac{4\sqrt{n+4}\tilde{D}\sigma}{\sqrt{N}}\\ &= \frac{LD_f^2}{N}\left[\frac{17(n+4)}{2}+\frac{8}{N}+\frac{1}{2}\right]+\frac{2\sigma\sqrt{n+4}}{\sqrt{N}}\left[\frac{D_f^2}{\tilde{D}}\left(1+\frac{1}{(n+4)N}\right)+2\tilde{D}\right]\\ &\le \frac{12L(n+4)D_f^2}{N}+\frac{4\sigma\sqrt{n+4}}{\sqrt{N}}\left(\tilde{D}+\frac{D_f^2}{\tilde{D}}\right).\end{aligned}$$

■

A few remarks about the results obtained in Corollary 2.2 are in order. Firstly, similar to the RSG method, we use the same selection of stepsizes $\{\gamma_k\}$ and probability mass function $P_R(\cdot)$ in RSGF method for both convex and nonconvex SP problems. In particular, in view of (3.26), the iteration complexity of the RSGF method for finding an ϵ-solution of problem (3.1) can be bounded by $\mathcal{O}(n/\epsilon^2)$. Moreover, in view of (3.27), if the problem is convex, a solution $\bar{x}$ satisfying $\mathbb{E}[f(\bar{x})-f^*]\le\epsilon$ can also be found in $\mathcal{O}(n/\epsilon^2)$ iterations. This result has a weaker dependence (by a factor of n) than the one established by Nesterov for solving general nonsmooth convex SP problems (see page 17 of [29]). This improvement is obtained since we are dealing with a more special class of SP problems. Also, note that in the case of $\sigma=0$, the iteration complexity of the RSGF method reduces to $\mathcal{O}(n/\epsilon)$ which is is similar to the one obtained by Nesterov [29] for the derivative free random search method when applied to both smooth convex and nonconvex deterministic problems.

Secondly, we need to specify $\tilde{D}$ for the stepsize policy in (3.24). According to (3.26) and (3.27), an optimal selection of $\tilde{D}$ would be D_f and D_X, respectively, for the nonconvex and convex case. With such selections, the bounds in (3.26) and (3.27), respectively, reduce to

$$\frac{1}{L}\mathbb{E}[\|\nabla f(x_R)\|^2]\le\frac{12(n+4)LD_f^2}{N}+\frac{8\sqrt{n+4}D_f\sigma}{\sqrt{N}},\tag{3.30}$$

$$\mathbb{E}[f(x_R)-f^*]\le\frac{5L(n+4)D_X^2}{N}+\frac{4\sqrt{n+4}D_X\sigma}{\sqrt{N}}.\tag{3.31}$$

Similarly to the RSG method, we can establish the complexity of the RSGF method for finding an (ϵ,Λ)-solution of problem (3.1) for some $\epsilon>0$ and $\Lambda\in(0,1)$.

More specifically, by using (3.26) and Markov's inequality, we have

$$\text{Prob}\left\{\|\nabla f(x_R)\|^2 \geq \lambda L\bar{\mathcal{B}}_N\right\} \leq \frac{1}{\lambda}, \quad \forall \lambda > 0, \tag{3.32}$$

which implies that the total number of calls to the $\mathcal{SZO}$ performed by the RSGF method for finding an (ϵ, Λ)-solution of (3.1) can be bounded by

$$\mathcal{O}\left\{\frac{nL^2D_f^2}{\Lambda\epsilon} + \frac{nL^2}{\Lambda^2}\left(\tilde{D} + \frac{D_f^2}{\tilde{D}}\right)^2\frac{\sigma^2}{\epsilon^2}\right\}. \tag{3.33}$$

We will investigate a possible approach to improve the above complexity bound in next subsection.

3.2. A two-phase randomized stochastic gradient free method. In this section, we modify the 2-RSG method to improve the complexity bound in (3.33) for finding an (ϵ, Λ)-solution of problem (3.1).

A two-phase RSGF (2-RSGF) method

Input: Initial point x_1, number of runs S, iteration limit N, and sample size T.

Optimization phase:

For $s = 1, \ldots, S$

Call the RSGF method with input x_1, iteration limit N, stepsizes $\{\gamma_k\}$ in (3.24), probability mass function P_R in (3.15), and the smoothing parameter μ satisfying (3.25). Let $\bar{x}_s$ be the output of this procedure.

Post-optimization phase:

Choose a solution $\bar{x}^*$ from the candidate list $\{\bar{x}_1, \ldots, \bar{x}_S\}$ such that

$$\|g_\mu(\bar{x}^*)\| = \min_{s=1,\ldots,S}\|g_\mu(\bar{x}_s)\|, \quad g_\mu(\bar{x}_s) := \frac{1}{T}\sum_{k=1}^{T} G_\mu(\bar{x}_s, \xi_k, u_k), \tag{3.34}$$

where $G_\mu(x, \xi, u)$ is defined in (3.12).

The main convergence properties of the 2-RSGF method are summarized in Theorem 3.4. More specifically, Theorem 3.4.a) establishes the rate of convergence of the 2-RSGF method with a given set of parameters (S, N, T), while Theorem 3.4.b) shows the complexity of this method for finding an (ϵ, Λ)-solution of problem (3.1).

THEOREM 3.4. *Under Assumptions A1 and A3, the following statements hold for the 2-RSGF method applied to problem (3.1).*

a) Let $\bar{\mathcal{B}}_N$ be defined in (3.26). We have

$$\begin{aligned}&\text{Prob}\left\{\|\nabla f(\bar{x}^*)\|^2 \geq 8L\bar{\mathcal{B}}_N + \tfrac{3(n+4)L^2D_f^2}{2N} + \tfrac{24(n+4)\lambda}{T}\left[L\bar{\mathcal{B}}_N + \tfrac{(n+4)L^2D_f^2}{N} + \sigma^2\right]\right\}\\ &\leq \tfrac{S+1}{\lambda} + 2^{-S}, \quad \forall\, \lambda > 0;\end{aligned} \tag{3.35}$$

b) Let $\epsilon > 0$ and $\Lambda \in (0, 1)$ be given. If S is set to $S(\Lambda)$ as in (2.24), and the iteration limit N and sample size T, respectively, are set to

$$N = \hat{N}(\epsilon) := \max\left\{\frac{12(n+4)(6LD_f)^2}{\epsilon}, \left[72L\sqrt{n+4}\left(\tilde{D} + \frac{D_f^2}{\tilde{D}}\right)\frac{\sigma}{\epsilon}\right]^2\right\}, \tag{3.36}$$

$$T = \hat{T}(\epsilon, \Lambda) := \frac{24(n+4)(S+1)}{\Lambda}\max\left\{1, \frac{6\sigma^2}{\epsilon}\right\}, \tag{3.37}$$

then the 2*-RSGF method can compute an* (ϵ, Λ)*-solution of problem (3.1) after taking at most*

$$2S(\Lambda)\left[\hat{N}(\epsilon) + \hat{T}(\epsilon, \Lambda)\right] \tag{3.38}$$

calls to the $\mathcal{SZO}$*.*

Proof. First, observe that by (3.6), (3.25) and (3.26), we have

$$\|\nabla f_\mu(x) - \nabla f(x)\|^2 \leq \frac{\mu^2}{4} L^2 (n+3)^3 \leq \frac{(n+4)L^2 D_f^2}{8N}. \tag{3.39}$$

Using this observation and the definition of $\bar{x}^*$ in (3.34), we obtain

$$\begin{aligned}
\|g_\mu(\bar{x}^*)\|^2 &= \min_{s=1,\ldots,S} \|g_\mu(\bar{x}_s)\|^2 = \min_{s=1,\ldots,S} \|\nabla f(\bar{x}_s) + g_\mu(\bar{x}_s) - \nabla f(\bar{x}_s)\|^2 \\
&\leq \min_{s=1,\ldots,S} \left\{2\left[\|\nabla f(\bar{x}_s)\|^2 + \|g_\mu(\bar{x}_s) - \nabla f(\bar{x}_s)\|^2\right]\right\} \\
&\leq \min_{s=1,\ldots,S} \left\{2\left[\|\nabla f(\bar{x}_s)\|^2 + 2\|g_\mu(\bar{x}_s) - \nabla f_\mu(\bar{x}_s)\|^2 + 2\|\nabla f_\mu(\bar{x}_s) - \nabla f(\bar{x}_s)\|^2\right]\right\} \\
&\leq 2 \min_{s=1,\ldots,S} \|\nabla f(\bar{x}_s)\|^2 + 4 \max_{s=1,\ldots,S} \|g_\mu(\bar{x}_s) - \nabla f(\bar{x}_s)\|^2 + \frac{(n+4)L^2 D_f^2}{2N},
\end{aligned}$$

which implies that

$$\begin{aligned}
\|\nabla f(\bar{x}^*)\|^2 &\leq 2\|g_\mu(\bar{x}^*)\|^2 + 2\|\nabla f(\bar{x}^*) - g_\mu(\bar{x}^*)\|^2 \\
&\leq 2\|g_\mu(\bar{x}^*)\|^2 + 4\|\nabla f_\mu(\bar{x}^*) - g_\mu(\bar{x}^*)\|^2 + 4\|\nabla f(\bar{x}^*) - \nabla f_\mu(\bar{x}^*)\|^2 \\
&\leq 4 \min_{s=1,\ldots,S} \|\nabla f(\bar{x}_s)\|^2 + 8 \max_{s=1,\ldots,S} \|g_\mu(\bar{x}_s) - \nabla f(\bar{x}_s)\|^2 + \frac{(n+4)L^2 D_f^2}{N} \\
&\quad + 4\|\nabla f_\mu(\bar{x}^*) - g_\mu(\bar{x}^*)\|^2 + 4\|\nabla f(\bar{x}^*) - \nabla f_\mu(\bar{x}^*)\|^2 \\
&\leq 4 \min_{s=1,\ldots,S} \|\nabla f(\bar{x}_s)\|^2 + 8 \max_{s=1,\ldots,S} \|g_\mu(\bar{x}_s) - \nabla f(\bar{x}_s)\|^2 \\
&\quad + 4\|\nabla f_\mu(\bar{x}^*) - g_\mu(\bar{x}^*)\|^2 + \frac{3(n+4)L^2 D_f^2}{2N},
\end{aligned} \tag{3.40}$$

where the last inequality also follows from (3.39). We now provide certain probabilistic bounds on the individual terms in the right hand side of the above inequality. Using (3.32) (with $\lambda = 2$), we obtain

$$\text{Prob}\left\{\min_{s=1,\ldots,S} \|\nabla f(\bar{x}_s)\|^2 \geq 2L\bar{\mathcal{B}}_N\right\} = \prod_{s=1}^{S} \text{Prob}\left\{\|\nabla f(\bar{x}_s)\|^2 \geq 2L\bar{\mathcal{B}}_N\right\} \leq 2^{-S}. \tag{3.41}$$

Moreover, denote $\Delta_{s,k} = G_\mu(\bar{x}_s, \xi_k, u_k) - \nabla f_\mu(\bar{x}_s)$, $k = 1, \ldots, T$. Note that, similar to (3.21), we have

$$\begin{aligned}
\mathbb{E}[\|G_\mu(\bar{x}_s, \xi_k, u_k)\|^2] &\leq 2(n+4)[\mathbb{E}[\|G(\bar{x}_s, \xi)\|^2] + \frac{\mu^2}{2} L^2 (n+6)^3 \\
&\leq 2(n+4)[\mathbb{E}[\|\nabla f(\bar{x}_s)\|^2] + \sigma^2] + 2\mu^2 L^2 (n+4)^3.
\end{aligned}$$

It then follows from the previous inequality, (3.25) and (3.26) that

$$\begin{aligned}\mathbb{E}[\|\Delta_{s,k}\|^2] &= \mathbb{E}[\|G_\mu(\bar{x}_s,\xi_k,u_k)-\nabla f_\mu(\bar{x}_s)\|^2] \le \mathbb{E}[\|G_\mu(\bar{x}_s,\xi_k,u_k)\|^2] \\ &\le 2(n+4)\left[L\bar{\mathcal{B}}_N+\sigma^2\right]+2\mu^2L^2(n+4)^3 \\ &\le 2(n+4)\left[L\bar{\mathcal{B}}_N+\sigma^2+\frac{L^2D_f^2}{2N}\right] =: \mathcal{D}_N. \end{aligned} \tag{3.42}$$

Noting that $g_\mu(\bar{x}_s)-\nabla f_\mu(\bar{x}_s)=\sum_{k=1}^T \Delta_{s,k}/T$, we conclude from (3.42), Assumption A1 and Lemma 2.3.a) that, for any $s=1,\ldots,S$,

$$\text{Prob}\left\{\|g_\mu(\bar{x}_s)-\nabla f_\mu(\bar{x}_s)\|^2 \ge \frac{\lambda\mathcal{D}_N}{T}\right\} = \text{Prob}\left\{\|\sum_{k=1}^T \Delta_{s,k}\|^2 \ge \lambda T\mathcal{D}_N\right\} \le \frac{1}{\lambda},\ \forall\lambda>0,$$

which implies that

$$\text{Prob}\left\{\max_{s=1,\ldots,S}\|g_\mu(\bar{x}_s)-\nabla f_\mu(\bar{x}_s)\|^2 \ge \frac{\lambda\mathcal{D}_N}{T}\right\} \le \frac{S}{\lambda},\ \forall\lambda>0. \tag{3.43}$$

and that

$$\text{Prob}\left\{\|g_\mu(\bar{x}^*)-\nabla f_\mu(\bar{x}^*)\|^2 \ge \frac{\lambda\mathcal{D}_N}{T}\right\} \le \frac{1}{\lambda},\ \forall\lambda>0. \tag{3.44}$$

The result then follows by combining relations (3.40), (3.41),(3.42), (3.43) and (3.44).

We now show part b) holds. Clearly, the total number of calls to $\mathcal{SZO}$ in the 2-RSGF method is bounded by $2S[\hat{N}(\epsilon)+\hat{T}(\epsilon)]$. It then suffices to show that $\bar{x}^*$ is an (ϵ,Λ)-solution of problem (3.1). Noting that by the definitions of $\bar{\mathcal{B}}(N)$ and $\hat{N}(\epsilon)$, respectively, in (3.26) and (3.36), we have

$$\bar{\mathcal{B}}_{\hat{N}(\epsilon)} = \frac{12(n+4)LD_f^2}{\hat{N}(\epsilon)} + \frac{4\sigma\sqrt{n+4}}{\sqrt{\hat{N}(\epsilon)}}\left(\tilde{D}+\frac{D_f^2}{\tilde{D}}\right) \le \frac{\epsilon}{36L}+\frac{\epsilon}{18L} = \frac{\epsilon}{12L}.$$

Hence, we have

$$8L\bar{B}_{\hat{N}(\epsilon)} + \frac{3(n+4)L^2D_f^2}{2\hat{N}(\epsilon)} \le \frac{2\epsilon}{3}+\frac{\epsilon}{288} \le \frac{17\epsilon}{24}.$$

Moreover, by setting $\lambda=[2(S+1)]/\Lambda$ and using (3.36) and (3.37), we obtain

$$\begin{aligned}\frac{24(n+4)\lambda}{T}\left[L\bar{\mathcal{B}}_{\hat{N}(\epsilon)}+\frac{(n+4)L^2D_f^2}{\hat{N}(\epsilon)}+\sigma^2\right] &\le \frac{24(n+4)\lambda}{T}\left(\frac{\epsilon}{12}+\frac{\epsilon}{432}+\sigma^2\right) \\ &\le \frac{\epsilon}{12}+\frac{\epsilon}{432}+\frac{\epsilon}{6} \le \frac{7\epsilon}{24}.\end{aligned}$$

Using these two observations and relation (3.35) with $\lambda=[2(S+1)]/\Lambda$, we conclude that

$$\begin{aligned}\text{Prob}\left\{\nabla f(\bar{x}^*)\|^2 \ge \epsilon\right\} \le \text{Prob}\Bigg\{&\|\nabla f(\bar{x}^*)\|^2 \ge 8L\bar{\mathcal{B}}_{\hat{N}(\epsilon)}+\frac{3(n+4)L^2D_f^2}{2\hat{N}(\epsilon)} \\ &+ \frac{24(n+4)\lambda}{T}\left[L\bar{\mathcal{B}}_{\hat{N}(\epsilon)}+\frac{(n+4)L^2D_f^2}{\hat{N}(\epsilon)}+\sigma^2\right]\Bigg\} \\ \le \frac{S+1}{\lambda}+2^{-S} &= \Lambda.\end{aligned}$$

■

Observe that in the view of (2.24), (3.36) and (3.37), the total number of calls to $\mathcal{SZO}$ performed by the 2-RSGF method can be bounded by

$$\mathcal{O}\left\{\frac{nL^2D_f^2\log(1/\Lambda)}{\epsilon}+nL^2\left(\tilde{D}+\frac{D_f^2}{\tilde{D}}\right)^2\frac{\sigma^2}{\epsilon^2}\log\frac{1}{\Lambda}+\frac{n\log^2(1/\Lambda)}{\Lambda}\left(1+\frac{\sigma^2}{\epsilon}\right)\right\}. \tag{3.45}$$

The above bound is considerably smaller than the one in (3.33), up to a factor of $\mathcal{O}\left(1/[\Lambda^2\log(1/\Lambda)]\right)$, when the second terms are the dominating ones in both bounds.

4. Numerical results. In this section, we present the results from our preliminary computational experiments where we implemented the developed stochastic first- and zeroth-order methods for solving three different test problems, namely, a stochastic convex least square problem, a stochastic nonconvex support vector machine problem and a simulation-based inventory optimization problem.

4.1. Preliminaries. *Algorithmic schemes.* The RSG method is implemented as described in Subsection 2.1 where the stepsizes $\{\gamma_k\}$ and probability mass function $P_R(\cdot)$ are set to (2.13) and (2.3), respectively. To implement the 2-RSG method, we take $S = 5$ independent runs of the RSG method to compute 5 candidate solutions in the optimization phase. We then use an i.i.d. sample of size $T = N/2$ in the post-optimization phase to estimate the gradients or function values at these candidate solutions and then choose the best one, $\bar{x}^*$, according to (2.21) or (2.22). Finally, the solution quality at $\bar{x}^*$ is evaluated by using another i.i.d. sample of size $K >> N$.

We also implement another variant of the 2-RSG method, namely, 2-RSG-V method. Similar to the 2-RSG method, the 2-RSG-V method consists of two phases. In the optimization phase, instead of terminating the RSG method by using a random count R, we terminate the algorithm by using a fixed number of iterations, say NS (see the discussions before Theorem 1). We then randomly pick up S solutions from the generated trajectory according to $P_R(\cdot)$ defined in (2.3), instead of running the RSG method S times. The post-optimization phase of the 2-RSG-V method is the same as the one of the 2-RSG method. Note that, in the 2-RSG-V method, unlike the 2-RSG method, the S candidate solutions are not independent. Therefore, we cannot provide the large-deviation results similar to the 2-RSG method. However, we observe from our computational results that this two-phase procedure can also significantly improve the large-deviation properties of the RSG method. Moreover, if the number of iterations used in the optimization phase of the 2-RSG-V method is about the same as the one in the 2-RSG method, i.e., $NS \approx N \times S$, then, in view of (2.14), the solutions computed by the 2-RSG-V method should be more accurate than those generated by the latter method.

In addition to the above methods, we also report the solutions obtained by taking the average of the trajectory of the 2-RSG-V method. This approach is essentially the mirror descent SA (MD-SA) method in [24, 18]. All these methods are implemented in MATLAB R2009b.

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. In particular, for the first two problems in our numerical experiment, 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. 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 (3.12) and take the average of them for each point, say $\bar{G}_\mu(x_i),\ \ i = 1, ...10$, as an approximation of the true gradient. Finally, we consider the maximum of $\frac{\|\bar{G}_\mu(x_i)-\bar{G}_\mu(x_j)\|}{\|x_i-x_j\|}$ for all pairs of i and j as an estimation of L. Also, we compute the stochastic gradients of the objective functions in our three problems at 200 randomly selected points with the same samples used in the estimation of L, and then take the maximum variance of the stochastic gradients over these points as an estimation of σ^2. Moreover, we estimate the parameter $\tilde{D} = D_f$ by (2.5). More specifically, since all the test problems considered in this section have nonnegative optimal values, i.e., $f^* \geq 0$, we have $D_f \leq (2f(x_1)/L)^{\frac{1}{2}}$, where x_1 denotes the initial point of the algorithms.

Notation in the tables.

- NS denotes the maximum number of calls to the $\mathcal{SFO}$ performed in the optimization phase the above methods. For example, $NS = 1,000$ has the following implications.
  - For the RSG method, the iteration limit N is set to $1,000$;
  - For the 2-RSG method, since $S = 5$, we set $N = 200$ so that totally at most $1,000$ calls to the $\mathcal{SFO}$ will be performed;
  - For the 2-RSG-V method, we run the RSG method for $NS = 1,000$ iterations and randomly select $S = 5$ solutions from the trajectory;
  - For the MD-SA method, we run the RSG method for $NS = 1,000$ iterations and take the average of the iterates as the output.
- $\bar{x}^*$ is the output solution of the above methods.
- *Mean* and *Var.* represent, respectively, the average and variance of the results obtained over different runs of each method.

4.2. Convex least square problem. In our first experiment, we consider the least square problem given by

$$\min_{x\in\mathbb{R}^n} \left\{ f(x) := \mathbb{E}_{u,v}[(\langle x, u\rangle - v)^2] \right\}. \tag{4.1}$$

Here, we assume that u is a sparse vector uniformly distributed in the hypercube $[0, 1]^n$ and v is obtained by $v = \langle \bar{x}, u\rangle + \xi$, where $\xi \sim N(0, \tilde{\sigma}^2)$ is the random noise independent of u and the coefficient $\bar{x}$ defines the true linear relationship between u and v. These types of problems have many applications in data fitting models (see, e.g., [21]).

We consider three different problem sizes, i.e., $n = 100, 500$ and $1,000$, and two different noise levels, i.e., $\tilde{\sigma} = 0.1$ and 1.0. Also, we set the data sparsity to 5%, implying that only five percent entries of u are nonzero for each data point (u, v). Also, the initial point is set to $x_1 = (0, ..., 0)^T \in \mathbb{R}^n$ for all problem instances.

Since the problem is convex, we use (2.22) to choose $\bar{x}^*$ in the post-optimization phase of the 2-RSG and 2-RSG-V method. Table 1 shows the mean and variance of $\tilde{f}(\bar{x}^*)$ over 20 runs of each algorithm. The following conclusions can be drawn from our experiments. Firstly, the RSG method is not very stable in the sense that there is a lot of variability over 20 runs of this algorithm. Secondly, both 2-RSG and 2-RSG-V can significantly improve the variability of the RSG method. Moreover, for a given fixed NS, the solution quality (in term of the $\tilde{f}(\bar{x}^*)$) of the 2-RSG-V method is significantly better than that of 2-RSG method. The reason is that, for

TABLE 4.1
Estimated function values $f(\bar{x}^*)$ for the least square problem ($K = 75,000$)

NS		RSG	2-RSG	2-RSG-V	MD-SA	RSG	2-RSG	2-RSG-V	MD-SA
		$n = 100, \tilde{\sigma} = 0.1$				$n = 100, \tilde{\sigma} = 1$			
1000	mean	0.1871	0.1118	0.0213	0.0139	1.481	1.269	1.103	1.088
	var.	1.34e-1	8.78e-4	2.16e-5	4.45e-7	3.41e-1	1.19e-2	2.91e-3	3.76e-4
5000	mean	0.0598	0.0223	0.0109	0.0104	1.207	1.086	1.040	1.026
	var.	1.97e-2	1.20e-4	2.83e-7	1.84e-8	1.82e-1	3.74e-4	4.77e-4	3.56e-5
25000	mean	0.0372	0.0111	0.0102	0.0102	1.043	1.037	1.014	1.010
	var.	1.11e-2	3.11e-6	5.10e-9	8.78e-9	2.95e-3	9.88e-5	1.74e-5	2.11e-5
		$n = 500, \tilde{\sigma} = 0.1$				$n = 500, \tilde{\sigma} = 1$			
1000	mean	1.719	1.798	0.4077	0.2910	3.469	3.066	1.755	1.591
	var.	1.05e+1	1.71e-2	1.40e-2	5.33e-4	3.26e+1	1.87e-2	2.24e-2	3.83e-3
5000	mean	2.613	0.3614	0.0343	0.0266	1.400	1.638	1.193	1.151
	var.	7.31e+1	7.45e-3	2.51e-4	3.11e-6	5.34e-1	8.46e-3	7.82e-4	1.03e-3
25000	mean	1.459	0.0346	0.0109	0.0105	1.114	1.187	1.064	1.041
	var.	3.80e+1	9.72e-4	1.37e-7	3.50e-9	8.72e-3	6.15e-3	1.25e-4	8.83e-5
		$n = 1000, \tilde{\sigma} = 0.1$				$n = 1000, \tilde{\sigma} = 1$			
1000	mean	4.679	4.647	1.341	1.164	3.139	5.897	2.749	2.537
	var.	9.33e+1	7.52e-2	7.26e-2	4.15e-3	2.36e-1	5.26e-2	4.43e-2	1.45e-2
5000	mean	2.963	1.296	0.2336	0.1218	2.089	2.705	1.459	1.336
	var.	1.03e+2	4.08e-2	1.55e-2	9.53e-5	4.27e-1	3.85e-2	1.66e-2	2.62e-4
25000	mean	1.444	0.2247	0.0157	0.0121	1.340	1.374	1.1095	1.067
	var.	3.66e+1	1.94e-2	2.28e-5	1.01e-7	7.09e-2	3.62e-3	6.41e-4	5.61e-5

fixed NS, 5 times more iterations are being used in the 2-RSG-V method to generate new solutions in the trajectory. Finally, the solution quality of 2-RSG-V method is comparable to that of the MD-SA method in most cases.

4.3. Nonconvex support vector machine problem. In our second experiment, we consider the support vector machine problem with a nonconvex sigmoid loss function, i.e.,

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) := \mathbb{E}_{u,v} \left[1 - \tanh\left(v \langle x, u \rangle \right) \right] + \lambda \|x\|_2^2 \right\}, \tag{4.2}$$

for some $\lambda > 0$ [23]. Here, we assume that each data point (u, v) is drawn from the uniform distribution on $\{0,1\}^n \times \{-1,1\}$, where $u \in \mathbb{R}^n$ is the feature vector and $v \in \{-1,1\}$ denotes the corresponding label. Moreover, we assume that u is sparse with 5% nonzero components and $v = \text{sign}\left(\langle \bar{x}, u \rangle\right)$ for some $\bar{x} \in \mathbb{R}^n$. In this experiment, we set $\lambda = 0.01$ and consider three different problem sizes, i.e., $n = 100$, 500 and 1000. Also, the initial point is set to $x_1 = 5 * \bar{x}_1$, where $\bar{x}_1$ is drawn from the uniform distribution over $[0,1]^n$. Table 4.2 reports the mean and variance of $\|g(\bar{x}^*)\|^2$ over 20 runs of the algorithms. Note that since the problem is nonconvex, we use (2.21) to choose $\bar{x}^*$ in the post-optimization phase of the 2-RSG and 2-RSG-V method. Moreover, due to the non-convexity of the problem, there is no guarantee that MD-SA will converge to a stationary point. In order to further assess the quality of the generated solutions, we also report in Table 4.3 the misclassification error evaluated at the obtained classifier $\bar{x}^*$, i.e.,

$$er(\bar{x}^*) := \frac{|\{i : v_i \neq \text{sign}(\langle \bar{x}^*, u_i \rangle), i = 1, ..., K\}|}{K}. \tag{4.3}$$

As it can be seen from Table 4.2, similarly to the convex case, the performance of the RSG method can be much improved by the 2-RSG method, which can be further

TABLE 4.2
Estimated gradients $\|\nabla f(\bar{x}^*)\|^2$ for the support vector machine problem ($K = 75,000$)

NS		RSG	2-RSG	2-RSG-V	MD-SA
$n = 100$					
1000	mean	0.0834	0.0821	0.0184	0.0625
	var.	9.73e-3	3.77e-4	5.06e-5	8.75e-8
5000	mean	0.0404	0.0212	0.0035	0.0208
	var.	8.04e-3	1.29e-4	3.33e-6	2.93e-8
25000	mean	0.0016	0.0027	0.0021	0.0087
	var.	4.58e-6	1.04e-6	2.15e-5	9.14e-9
$n = 500$					
1000	mean	0.8165	0.7423	0.2769	0.5837
	var.	1.91e-1	1.50e-2	6.90e-3	5.44e-9
5000	mean	0.2865	0.2630	0.0510	0.2137
	var.	1.43e-1	1.36e-2	2.54e-3	3.37e-8
25000	mean	0.2741	0.0549	0.0068	0.0525
	var.	2.92e-1	2.75e-3	1.64e-4	2.70e-8
$n = 1000$					
1000	mean	1.392	1.428	0.5557	1.5353
	var.	9.49e-1	2.49e-2	2.75e-2	2.09e-9
5000	mean	0.4661	0.5995	0.0778	0.6804
	var.	3.39e-1	6.10e-2	3.96e-3	3.98e-9
25000	mean	0.3875	0.0566	0.0092	0.1807
	var.	5.94e-1	4.45e-4	2.28e-5	3.86e-8

TABLE 4.3
Average misclassification error $er(\bar{x}^*)\%$ at $\bar{x}^*$

NS	RSG	2-RSG	2-RSG-V	MD-SA
$n = 100$				
1000	44.37	48.99	35.44	47.38
5000	23.45	37.32	11.69	37.61
25000	7.58	9.70	8.06	21.73
$n = 500$				
1000	51.50	51.62	50.69	51.48
5000	47.04	50.59	43.32	49.77
25000	30.69	43.30	13.60	43.86
$n = 1000$				
1000	48.70	48.78	48.55	48.84
5000	47.56	48.58	45.36	48.56
25000	33.21	44.96	18.31	46.40

improved by the 2-RSG-V method for a given NS. However, unlike the convex case, the performance of the MD-SA can be significantly worse than that of the 2-RSG-V up to orders of magnitude. Moreover, as can be seen from Table 4.3, the 2-RSG-V method can significantly outperform the MD-SA method in terms of the average misclassification error with the increment of the sample size NS.

4.4. Simulation-based inventory optimization problem. In the last experiment, we consider the classical (s, S) inventory problem. More specifically, we consider the following simple case study in [42]. Widgets company carries inventory of one item. Customers arrive according to Poisson distribution with mean 10 persons per day and they demand 1, 2, 3, and 4 widgets 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

TABLE 4.4
Estimated gradients $\|\nabla f_\mu(\bar{x}^*)\|^2$ *for the inventory problem* ($K = 10,000$)

$x_1 = (s_1, S_1)$	NS		RSGF	2-RSGF	2-RSGF-V	MD-SA-GF
$(10, 100)$	1000	mean	0.1720	0.0184	0.0251	0.1204
		var.	9.11e-2	1.24e-4	4.96e-4	9.10e-2
	5000	mean	0.0627	0.0162	0.0092	0.3001
		var.	1.76e-1	4.22e-4	1.34e-4	3.35e-1
$(50, 100)$	1000	mean	0.8543	0.7795	0.5170	0.6648
		var.	2.20e-2	3.96e-2	1.16e-1	8.24e-2
	5000	mean	0.5900	0.4428	0.5558	0.7777
		var.	7.07e-2	1.11e-1	7.16e-2	1.12e-1
$(10, 50)$	1000	mean	0.8467	0.5707	0.3924	0.5322
		var.	2.94e-1	2.65e-1	1.32e-1	1.25e-1
	5000	mean	0.7340	0.3318	0.0784	0.1991
		var.	1.21e-0	2.05e-1	4.51e-2	2.17e-2

the inventory up to S. Also, the lead time 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

$$\min_{(s,S)\in\mathbb{R}^2} \mathbb{E}\left[\text{daily inventory cost}\right] + \rho\left\{\left[\max\left(0,-s\right)\right]^2 + \left[\max\left(0,s-S\right)\right]^2\right\}, \qquad (4.4)$$

where $\rho > 0$ is the penalization parameter. In particular, ρ is set to 100 in our experiments.

We implement the RSGF method as described in Subsection 3.1, where the stepsizes $\{\gamma_k\}$ and probability mass function $P_R(\cdot)$ are set to (3.24) and (3.15), respectively. Moreover, we compute the value of the first term in (4.4) by simulating the inventory system over 100 days, while the gradient of the second term is known exactly. The other zeroth-order methods are implemented similarly to their corresponding first-order methods as described in Subsection 4.1. Note that in the post-optimization phase of the 2-RSGF and 2-RSGF-V methods, $\bar{x}^*$ is chosen according to (3.34). Also, the smoothing parameter μ satisfying (3.25) is set to 0.0025 for all these zeroth-order methods.

Table 4.4 reports the mean and variance of $\|g_\mu(\bar{x}^*)\|^2$ over 10 runs of these methods with different initial solutions x_1. Similar to the results for the nonconvex support vector machine problem, the solution quality (in term of $\|g_\mu(\bar{x}^*)\|^2$) of the RSGF method is not as good as the 2-RSGF method and, for a given NS, the 2-RSGF-V method outperforms the 2-RSGF method in many cases. Moreover, in most cases the 2-RSGF-V method can significantly outperform the MD-SA-GF method (i.e., the gradient free version of the MD-SA method described in Subsection 3.1). Table 4.5 shows the corresponding average daily inventory costs for the solutions, $\bar{x}^*$, computed by these algorithms. The best solution given by $(s, S) = (23.7, 64.5)$ with an estimated average daily inventory cost \$118.47 has been obtained by running the 2-RSG-V method starting from the initial solution $(10, 50)$.

5. Concluding remarks. In this paper, we present a class of new SA methods for solving the classical unconstrained NLP problem with noisy first-order information. We establish a few new complexity results regarding the computation of an ϵ-solution

TABLE 4.5
Average daily inventory costs

$x_1 = (s_1, S_1)$	NS	RSGF	2-RSGF	2-RSGF-V	MD-SA-GF
$(10, 100)$	1000	129.11	129.39	127.64	129.10
	5000	127.44	128.03	129.27	127.17
$(50, 100)$	1000	137.09	134.58	130.32	131.86
	5000	130.93	130.71	129.04	130.54
$(10, 50)$	1000	139.12	125.16	126.79	123.73
	5000	124.12	122.66	121.50	121.37

for solving this class of problems and show that they are nearly optimal whenever the problem is convex. Moreover, we introduce a post-optimization phase in order to improve the large-deviation properties of the RSG method. These procedures, along with their complexity results, are then specialized for simulation-based optimization problems when only stochastic zeroth-order information is available. In addition, we show that the complexity for gradient-free methods for smooth convex SP can have a much weaker dependence on the dimension n than that for more general nonsmooth convex SP. Promising numerical results are also reported for the developed algorithms.

REFERENCES

[1] S. Andradóttir. A review of simulation optimization techniques. *Proceedings of the 1998 Winter Simulation Conference*, pages 151–158.

[2] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the complexity of steepest descent, newtons and regularized newtons methods for nonconvex unconstrained optimization. *SIAM Journal on Optimization*, 20(6):2833–2852, 2010.

[3] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the oracle complexity of first-order and derivative-free algorithms for smooth nonconvex minimization. *SIAM Journal on Optimization*, 22:66–86, 2012.

[4] K.L. Chung. On a stochastic approximation method. *Annals of Mathematical Statistics*, pages 463–483, 1954.

[5] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to Derivative-Free Optimization*. SIAM, Philadelphia, 2009.

[6] J. C. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22:674–701, 2012.

[7] M. Fu. Optimization for simulation: Theory vs. practice. *INFORMS Journal on Computing*, 14:192–215, 2002.

[8] M.C. Fu. Gradient estimation. In S. G. Henderson and B. L. Nelson, editors, *Handbooks in Operations Research and Management Science: Simulation*, page 575616. Elsevier.

[9] R. Garmanjani and L. N. Vicente. Smoothing and worst-case complexity for direct-search methods in nonsmooth optimization. *IMA Journal of Numerical Analysis*, 2012. to appear.

[10] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. Technical report, 2010. *SIAM Journal on Optimization* (under second-round review).

[11] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, I: a generic algorithmic framework. *SIAM Journal on Optimization*, 22:1469–1492, 2012.

[12] S. Ghadimi and G. Lan. 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. *E-print available at:* `http://www.optimization-online.org`.

[13] P. Glasserman. *Gradient Estimation via Perturbation Analysis*. Kluwer Academic Publishers, Boston,Massachusetts, 2003.

[14] A. Juditsky, A. Nazin, A. B. Tsybakov, and N. Vayatis. Recursive aggregation of estimators via the mirror descent algorithm with average. *Problems of Information Transmission*, 41:n.4, 2005.

[15] A. Juditsky and A. 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.

[16] A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *Annals of Statistics*, 36:2183–2206, 2008.

[17] A. J. Kleywegt, A. Shapiro, and T. Homem de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12:479–502, 2001.

[18] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

[19] G. Lan, A. Nemirovski, and A. Shapiro. Validation analysis of mirror descent stochastic approximation method. *Mathematical Programming*, 134:425–458, 2012.

[20] P. LÉcuyer. A unified view of the IPA, SF, and LR gradient estimation techniques. *Management Science*, 36(11):1364–1383, 1990.

[21] T. Leeuwen, M. Schmidt, M. Freidlander, and F. Herrmann. A hybrid stochastic-deterministic optimization method for waveform inversion. Manuscript, University of British Columbia, Vancouver, January 2011.

[22] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *In ICML*, pages 689–696, 2009.

[23] L. Mason, J. Baxter, P. Bartlett, and M. Frean. Boosting algorithms as gradient descent in function space. *Proc. NIPS*, 12:512–518, 1999.

[24] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19:1574–1609, 2009.

[25] A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley, XV, 1983.

[26] Y. E. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983.

[27] Y. E. Nesterov. *Introductory Lectures on Convex Optimization: a basic course*. Kluwer Academic Publishers, Massachusetts, 2004.

[28] Y. E. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120:221–259, 2006.

[29] Y. E. Nesterov. Random gradient-free minimization of convex functions. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, January 2010.

[30] Y. E. Nesterov and J. P. Vial. Confidence level solutions for stochastic programming. 2000.

[31] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer-Verlag, New York, USA, 1999.

[32] B.T. Polyak. New stochastic approximation type procedures. *Automat. i Telemekh.*, 7:98–107, 1990.

[33] B.T. Polyak and A.B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control and Optimization*, 30:838–855, 1992.

[34] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[35] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis, ser. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998.

[36] R.Y. Rubinstein and A. Shapiro. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. John Wiley & Sons, 1993.

[37] A. Sartenaer S. Gratton and Ph. L. Toint. Recursive trust-region methods for multiscale nonlinear optimization. *SIAM Journal on Optimization*, 19:414–444, 2008.

[38] J. Sacks. Asymptotic distribution of stochastic approximation. *Annals of Mathematical Statistics*, 29:373–409, 1958.

[39] A. Shapiro. Monte carlo sampling methods. In A. Ruszczyński and A. Shapiro, editors, *Stochastic Programming*. North-Holland Publishing Company, Amsterdam, 2003.

[40] J.C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. John Wiley, Hoboken, NJ, 2003.

[41] L. N. Vicente. Worst case complexity of direct search. *EURO Journal on Computational Optimization*, 2012. to appear.

[42] R. P. Sadowski W. D. Kelton and D. T. Sturrock. *Simulation with Arena*. McGraw-Hill, New York, fourth edition, 2007.

[43] F. Yousefian, A. Nedic, and U. V. Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48:56–67, 2012.