SIAM J. OPTIM.
Vol. 30, No. 1, pp. 960–979

A SINGLE TIMESCALE STOCHASTIC APPROXIMATION METHOD FOR NESTED STOCHASTIC OPTIMIZATION*

SAEED GHADIMI†, ANDRZEJ RUSZCZYŃSKI‡, AND MENGDI WANG§

Abstract. We study constrained nested stochastic optimization problems in which the objective function is a composition of two smooth functions whose exact values and derivatives are not available. We propose a single timescale stochastic approximation algorithm, which we call the nested averaged stochastic approximation (NASA), to find an approximate stationary point of the problem. The algorithm has two auxiliary averaged sequences (filters) which estimate the gradient of the composite objective function and the inner function value. By using a special Lyapunov function, we show that the NASA achieves the sample complexity of $\mathcal{O}(1/\varepsilon^2)$ for finding an ε-approximate stationary point, thus outperforming all extant methods for nested stochastic approximation. Our method and its analysis are the same for both unconstrained and constrained problems, without any need of batch samples for constrained nonconvex stochastic optimization. We also present a simplified parameter-free variant of the NASA method for solving constrained single-level stochastic optimization problems, and we prove the same complexity result for both unconstrained and constrained problems.

Key words. stochastic approximation, compositional optimization, stochastic gradient, stochastic variational inequality, machine learning

AMS subject classifications. 90C15, 62L20

DOI. 10.1137/18M1230542

1. Introduction.

The main objective of this work is to propose a new recursive stochastic algorithm for constrained smooth composition optimization problems of the following form:

$$\min_{x\in X} \big\{F(x) = f(g(x))\big\}. \tag{1.1}$$

Here, the functions $f : \mathbb{R}^m \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are continuously differentiable, and the set $X \subseteq \mathbb{R}^n$ is convex and closed. We do not assume f, g, or F to be convex.

We focus on the simulation setting where neither the values nor the derivatives of f or g can be observed, but at any argument values $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$ we can obtain random estimates of $g(x)$, of the Jacobian $\nabla g(x)$, and of the gradient $\nabla f(u)$. Such situations occur in *stochastic composition optimization*, where we need to solve the problem

$$\min_{x\in X} \mathbb{E}\big[\varphi\big(\mathbb{E}[\psi(x;\zeta)];\xi\big)\big] \tag{1.2}$$

in which ζ and ξ are random vectors, and $\mathbb{E}$ denotes the expected value. In such situations, one can obtain samples $(\tilde{\xi}, \tilde{\zeta})$ of (ξ, ζ), and treat $\psi(x, \tilde{\zeta})$, $\nabla_x \psi(x, \tilde{\zeta})$, and $\nabla_u \varphi(u, \xi)$ as random estimates of $\mathbb{E}[\psi(x;\zeta)]$, $\nabla\mathbb{E}[\psi(x;\zeta)]$, and $\nabla\mathbb{E}\big[\varphi(u,\xi)\big]$, respectively. In this paper, we propose stochastic gradient-type methods for finding approx-

*Received by the editors December 4, 2018; accepted for publication (in revised form) December 20, 2019; published electronically March 17, 2020.
https://doi.org/10.1137/18M1230542
Funding: This work was supported by the AFOSR Grant FA9550-19-1-0203 608 and the NSF Awards CMMI-1653435 and DMS-1907522.

†Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08540 (sghadimi@princeton.edu).
‡Department of Management Science and Information Systems, Rutgers University, Piscataway, NJ 08854 (rusz@rutgers.edu).
§Department of Electrical Engineering and Center for Statistics and Machine Learning, Princeton University, Princeton, NJ 08540 (mengdiw@princeton.edu).

imate stationary points of problems of this type. We also derive sample complexity guarantees for these methods.

Stochastic composition problems of form (1.1)–(1.2) occur in many applications; we present three modern motivating examples.

Example 1.1 (stochastic variational inequality). We have a random mapping $H : \mathbb{R}^n \times \Omega \to \mathbb{R}^n$ on some probability space $(\Omega, \mathcal{F}, P)$ and a closed convex set X. The problem is to find $x \in X$ such that

$$\big\langle \mathbb{E}[H(x)], y - x \big\rangle \le 0 \quad \text{for all} \quad y \in X. \tag{1.3}$$

The reader is referred to the recent publications [13] and [17] for a discussion of the challenges associated with this problem and its applications (our use of the "$\le$" relation instead of the common "$\ge$" is only motivated by the ease with which we show the conversion to our formulation). We propose to convert problem (1.3) to the nested form (1.1) by defining the *lifted gap function* $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ as

$$f(x,h) = \max_{y \in X} \left\{ \langle h, y - x \rangle - \frac{1}{2}\|y - x\|^2 \right\}, \tag{1.4}$$

and the function $g : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n$ as $g(x) = \big(x, \mathbb{E}[H(x)]\big)$. In this case, we actually have access to the gradient of f, but the value and the Jacobian of g must be estimated. We do not require $E[H(x)]$ to be monotone. ∎

Example 1.2 (policy evaluation for Markov decision processes). For a Markov chain $\{X_0, X_1, \ldots\} \subset \mathcal{X}$ with an unknown transition operator P, a reward function $r : \mathcal{X} \mapsto \mathbb{R}$, and a discount factor $\gamma \in (0,1)$, we want to estimate the value function $V : \mathcal{X} \mapsto \mathbb{R}$ given by $V(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(X_t) \mid X_0 = x\right]$. For a finite space $\mathcal{X}$, the functions r and V may be viewed as vectors, and the following policy evaluation equation is satisfied:

$$V = r + \gamma P V.$$

As P is not known and $|\mathcal{X}|$ may be large, this system cannot be solved directly. To reduce the dimension of this problem, we employ a sketching matrix $S \in \mathbb{R}^{d \times |\mathcal{X}|}$ and a linear model for the value function $V(x) \approx \sum_{i=1}^{k} w_i \phi_i(x)$, where $\phi_1(\cdot), \ldots, \phi_k(\cdot)$ are given basis functions. Then we can formulate the residual minimization problem for the policy evaluation equation:

$$\min_{w \in \mathbb{R}^d} \mathbb{E}\left[\left\| S\left(\Phi w - r - \gamma\, \mathbb{E}[\hat{P}] \Phi w \right) \right\|^2 \right],$$

where Φ is the matrix with columns being the basis functions, $\hat{P}$ is a sample transition matrix (see [28] and the references therein), S is a sketching matrix that selects a random subset of $\{1, \ldots, d\}$ such that $\mathbb{E}[\|Sy\|^2] = y^T \Xi y$ corresponds to a weighted norm (this weight is often chosen to be the invariant distribution of P, which is not known but can be sampled from to construct the random S). In this case, we may define the outer function f as the weighted squared norm, and the inner function g as the linear mapping inside the norm. Neither of the functions has an easily available value or derivative, but their samples can be generated by simulation. ∎

Example 1.3 (low-rank matrix estimation). Let $\bar{X} \in \mathbb{R}^{n \times n}$ be an unknown matrix that we aim to approximate. One can sample from the unknown matrix and each sample returns a random matrix X such that $\mathbb{E}[X] = \bar{X}$. Let $k < n$ be a prespecified rank. The low-rank matrix estimation problem has the following form:

$$\min_{(U,V) \in S} \ell\left(\mathbb{E}[X] - UV^T \right).$$

In this problem, the unknowns U and V are $n \times k$ matrices, $S \subset \mathbb{R}^{n\times k} \times \mathbb{R}^{n\times k}$ is a bounded set, and $\ell : \mathbb{R}^{n\times n} \to \mathbb{R}$ is a loss function (e.g., the Frobenius norm). Low-rank matrix approximation finds wide applications including image analysis, topic models, recommendation systems, and Markov models (see [8] and the references therein). Our formulation is nonconvex. When data arrive sequentially, our method can be applied in an online fashion to find a stationary solution. ∎

Interest in stochastic approximation algorithms for problems of form (1.1) dates back to [6, Chap. V.4], where penalty functions for stochastic constraints and composite regression models were considered. There, and in the literature that followed, the main approach was to use two- or multiple-level stochastic recursive algorithms in different timescales. For problems of form (1.1) this amounts to using two step size sequences: one for updating the main decision variable x, and another one for filtering the value of the inner function g. The crucial requirement is that the outer method must be infinitely slower than the inner method, which decreases the convergence rate and creates practical difficulties. Sample complexity analysis and acceleration techniques of stochastic approximation methods with multiple timescales for solving problems of form (1.1) gained interest in recent years. We refer the readers to [27, 28, 30] for a detailed account of these techniques and existing results for the general nested composition optimization problem. Furthermore, a central limit theorem for the stochastic composition problem (1.1)–(1.2) has been established in [5]. It shows that the N-sample empirical optimal value of problem (1.2) converges to the true optimal value at a rate of $\mathcal{O}(1/\sqrt{N})$. The work [7] establishes large deviation bounds for the empirical optimal value.

In addition to the general solution methods studied in [5, 27, 28, 30], several notable special cases of the composition problem have been considered in the machine learning literature. In the case where f is convex and g is linear, one can solve (1.1) using duality-based methods via a Fenchel dual reformulation [2]. In the case where it is allowed to take minibatches, [1] proposed a sampling scheme to obtain unbiased sample gradients of F by forming randomly sized minibatches. In the case when f and g take the form of a finite sum and strong convexity holds, one can leverage the special structure to obtain linearly convergent algorithms; see, e.g., [18, 20]. To the authors' best knowledge, no method exists for solving (1.1) with general smooth f and g, which uses a single timescale stochastic approximation update and does not resort to minibatches. There is also no method for approximating stationary solutions that has provable complexity bounds, when the composition problem is constrained.

Our contributions are the following. First, we propose a new nested averaged stochastic approximation (NASA) algorithm for solving (1.1), which is qualitatively different from the earlier approaches. Its main idea, inspired by [25, 22, 23], is to lift the problem into a higher dimensional space, $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m$, where our objective is not only to find the optimal x, but also to find the gradient of F at the optimal point, and the value of g at this point. In this space, we construct an iterative method using one step size sequence, and we prove convergence by employing a specially tailored merit (Lyapunov) function. This leads to the first single-timescale stochastic approximation algorithm for the composition problem, and entails essential improvements over the earlier approaches.

Second, we show that with proper choice of the step size sequence, $\mathcal{O}(1/\varepsilon^2)$ observations are sufficient for the NASA algorithm to find a pair $(\bar{x}, \bar{z}) \in X \times \mathbb{R}^n$ satisfying $\mathbb{E}[V(\bar{x}, \bar{z})] \le \varepsilon$, where $\bar{z}$ is an estimate for $\nabla F(\bar{x})$, and $V(x, z)$ is an optimality measure generalizing $\|\nabla F(\bar{x})\|^2$ to constrained problems; see (2.6). This complexity bound is

consistent with the central limit theorem for composite risk functionals [5] and is better than the best-known complexity of $\mathcal{O}(1/\varepsilon^{2.25})$ obtained in [28] for smooth nested nonconvex stochastic optimization. In fact, our complexity bound for the two-level composition problem is of the same order as the complexity of the stochastic gradient method for general smooth one-level nonconvex stochastic optimization [9].

Third, our convergence analysis of the NASA method is the same for both unconstrained and constrained problems, allowing us to obtain the same complexity of $\mathcal{O}(1/\varepsilon^2)$ for the constrained case, without taking batches of samples per iteration. To the best of our knowledge, this is the first direct convergence analysis of a method for general stochastic nested problems of form (1.1) which avoids multiple samples per iteration to reduce the variance of the stochastic gradients. Hence, this property makes the NASA method attractive for online learning where the samples are received one by one.

Finally, we present a simplified variant of the NASA method for solving a class of single-level stochastic optimization problems, i.e., with $g(x) \equiv x$ in problem (1.1). This simplified version is a form of a constrained dual averaging method. We show that the sample (iteration) complexity of this algorithm is of the same order as that of the NASA method. Moreover, the step size schedule of this method, unlike almost all existing stochastic approximation algorithms, does not depend on any problem parameters or employ line-search procedures. Its rate of convergence is established without forming minibatches of samples, and is valid for both unconstrained and constrained problems. It should be mentioned that a similar complexity bound has recently been obtained in [4, 3] for finding an approximate stationary point (albeit with a different optimality measure) for nonsmooth, constrained, and nonconvex one-level stochastic optimization problems without taking minibatches of samples per iteration. Some online algorithms for constrained problems that use minibatches of samples choose their size to improve error complexity, while making a trade-off with sample complexity (see, e.g., [10, 11]). Hence, they may be more desirable in the cases where the projection onto the feasible set is computationally hard.

Notation. The optimal value of problem (1.1) is denoted by F^*. For any Lipschitz continuous function h, we use L_h to denote its Lipschitz constant. We use ∇h to denote the gradient (or Jacobian) of a scalar (or vector) function h.

2. The method. Our goal in this section is to propose a stochastic approximation algorithm for solving problem (1.1) where estimates of the gradient of f and the value and Jacobian of g are available through calls to a stochastic oracle.

The method generates three random sequences, namely, approximate solutions $\{x^k\}$, average gradients $\{z^k\}$, and average g-values $\{u^k\}$, defined on a certain probability space $(\Omega, \mathcal{F}, P)$. We let $\mathcal{F}_k$ be the σ-algebra generated by

$$\{x^0, \dots, x^k, z^0, \dots, z^k, u^0, \dots, u^k\}.$$

We also make the following assumption on the stochastic oracle.

Assumption 1. For each k, the stochastic oracle delivers random vectors $G^{k+1} \in \mathbb{R}^m$, $s^{k+1} \in \mathbb{R}^n$, and a random matrix $J^{k+1} \in \mathbb{R}^{m\times n}$, such that

$$\begin{aligned}
&\mathbb{E}[G^{k+1}|\mathcal{F}_k] = g(x^{k+1}), \qquad \mathbb{E}[\|G^{k+1} - g(x^{k+1})\|^2|\mathcal{F}_k] \le \sigma_G^2,\\
&\mathbb{E}[J^{k+1}|\mathcal{F}_k] = \nabla g(x^{k+1}), \qquad \mathbb{E}[\|J^{k+1}\|^2|\mathcal{F}_k] \le \sigma_J^2,\\
&\mathbb{E}[s^{k+1}|\mathcal{F}_k] = \nabla f(u^k), \qquad \mathbb{E}[\|s^{k+1}\|^2|\mathcal{F}_k] \le \sigma_s^2,
\end{aligned}$$

and J^{k+1} and s^{k+1} are conditionally independent, given $\mathcal{F}_k$.

The shift in indexing of x^k in the above assumption is due to the fact that x^{k+1} will be $\mathcal{F}_k$-measurable in our method. Our method proceeds as presented in Algorithm 2.1.

Algorithm 2.1 Nested averaged stochastic approximation (NASA).

Input: $x^0 \in X$, $z^0 \in \mathbb{R}^n$, $u^0 \in \mathbb{R}^m$, $a > 0$, $b > 0$, $\beta > 0$.

0. Set $k = 0$.

1. For an $\mathcal{F}_k$-measurable step size $\tau_k \in (0, 1/a]$, compute

$$y^k = \arg\min_{y \in X} \left\{ \langle z^k, y - x^k \rangle + \frac{\beta}{2} \|y - x^k\|^2 \right\} \tag{2.1}$$

and set

$$x^{k+1} = x^k + \tau_k (y^k - x^k). \tag{2.2}$$

2. Call the stochastic oracle to obtain s^{k+1} at u^k, G^{k+1} and J^{k+1} at x^{k+1}, and update the running averages as

$$z^{k+1} = (1 - a\tau_k) z^k + a\tau_k \left[J^{k+1}\right]^T s^{k+1}, \tag{2.3}$$

$$u^{k+1} = (1 - b\tau_k) u^k + b\tau_k G^{k+1}. \tag{2.4}$$

3. Increase k by one and go to step 1.

A few remarks are in order. First, the stochastic gradient $\left[J^{k+1}\right]^T s^{k+1}$ returned by the stochastic oracle is a biased estimator of the gradient of $F(x^{k+1})$. Hence, z^{k+1}, as a weighted average of these stochastic gradients, is also a biased estimator of $\nabla F(x^{k+1})$. However, the sum of the bias terms of the latter estimator grows slower than the former one, ensuring convergence of the algorithm (see Theorem 3.5). Second, u^{k+1} is also a biased estimator of $g(x^{k+1})$, whose error can be properly controlled and asymptotically driven to 0. Finally, convergence of Algorithm 2.1 depends on the choice of the sequence $\{\tau_k\}$ and the parameters a, b, and β, which will be specified in the next section.

We end this section with a brief review of the optimality conditions for problem (1.1) and their relation to the subproblem (2.1). The following fact is standard (e.g., [24, Thm. 3.24]).

THEOREM 2.1. *If a point $\hat{x} \in X$ is a local minimum of problem* (1.1)*, then*

$$-\nabla F(\hat{x}) \in \mathcal{N}_X(\hat{x}), \tag{2.5}$$

where $\mathcal{N}_X(\hat{x})$ denotes the normal cone to X at the point $\hat{x}$. If in addition the function $F(\cdot)$ is convex, then every point $\hat{x}$ satisfying (2.5) *is the global minimum of problem* (1.1).

Condition (2.5) is closely related to the subproblem (2.1). Denote (for $\beta > 0$)

$$\bar{y}(x, z, \beta) = \arg\min_{y \in X} \left\{ \langle z, y - x \rangle + \frac{\beta}{2} \|y - x\|^2 \right\}.$$

Elementary manipulation shows that

$$\bar{y}(x, z, \beta) = \Pi_X\Big(x - \frac{1}{\beta} z\Big),$$

where $\Pi_X(\cdot)$ is the operation of the orthogonal projection on the set X. The relation $-z \in \mathcal{N}_X(x)$ is equivalent to $\bar{y}(x, z, \beta) = x$. We will, therefore, use the function

$$V(x, z) = \|\bar{y}(x, z, 1) - x\|^2 + \|z - \nabla F(x)\|^2 \tag{2.6}$$

as a measure of violation of the optimality condition (2.5) by the primal-dual pair (x, z). In the unconstrained case where $X = \mathbb{R}^n$, we have $V(x, z) = \|z\|^2 + \|z - \nabla F(x)\|^2$.

The following lemma relates the optimality measure $V(x, z)$ to subproblem (2.1) for an arbitrary $\beta > 0$.

LEMMA 2.2. *For every $x \in X$ and every $\beta > 0$,*

$$\|\bar{y}(x, z, 1) - x\| \le \max(1, \beta)\, \|\bar{y}(x, z, \beta) - x\|. \tag{2.7}$$

Proof. To simplify notation, set $x = 0$, $y(\beta) = \bar{y}\big(x, z, \beta\big) = \Pi_X(-\frac{1}{\beta} z)$. By the characterization of the orthogonal projection,

$$\big\langle z + \tau y(\tau), \xi - y(\tau) \big\rangle \ge 0 \quad \text{for all} \quad \xi \in X, \quad \tau > 0. \tag{2.8}$$

Setting $\tau = \beta$ and $\xi = y(1)$ we obtain

$$\big\langle z + \beta y(\beta), y(1) - y(\beta) \big\rangle \ge 0.$$

Now we set $\tau = 1$ and $\xi = y(\beta)$ in (2.8) and get

$$\big\langle z + y(1), y(\beta) - y(1) \big\rangle \ge 0.$$

Adding these inequalities yields

$$\langle \beta y(\beta) - y(1), y(1) - y(\beta) \rangle \ge 0. \tag{2.9}$$

Consider two cases.

Case 1. $\beta \ge 1$. Inequality (2.9) implies that

$$(\beta - 1)\|y(\beta)\| = \|\beta y(\beta) - y(\beta)\| \ge \|y(1) - y(\beta)\|.$$

By the triangle inequality and the last relation,

$$\|y(1)\| \le \|y(\beta)\| + \|y(1) - y(\beta)\| \le \beta \|y(\beta)\|,$$

which proves our claim in this case.

Case 2. $0 < \beta \le 1$. From (2.9) we obtain

$$(1 - \beta)\langle y(\beta) - y(1), y(1) \rangle \ge \beta \|y(\beta) - y(1)\|^2 \ge 0.$$

Therefore, $\|y(\beta)\| \ge \|y(1)\|$ and our claim is true in this case as well. ∎

Note that our measure of nonoptimality in (2.6) is an upper bound for the squared norm of the gradient, when $X = \mathbb{R}^n$. For the constrained case, it can also be related to the existing ones in the literature. To do so, we need to view other algorithms in the primal-dual space. For example, for the proximal point mapping used in [4, 3],

$$\hat{y} = \arg\min_{y \in X} \left\{ F(y) + \frac{1}{2}\|y - x\|^2 \right\},$$

the squared distance $\|\hat{y} - x\|^2$ is employed as a measure of nonoptimality (the use of a parameter γ there can be dealt with by minor adjustments). By optimality conditions of the above subproblem, we have $\hat{y} = \bar{y}(x, \hat{z}, 1)$ with $\hat{z} = \nabla F(\hat{y})$. If we view the method of [4] as generating primal-dual pairs of form $(x, \hat{z})$, we obtain

$$\|\hat{y} - x\|^2 \le V(x, \hat{z}) = \|\hat{y} - x\|^2 + \|\hat{z} - \nabla F(x)\|^2 \le (1 + L_{\nabla F}^2)\|\hat{y} - x\|^2.$$

It follows that both optimality criteria would be equivalent in the primal-dual space, if $\hat{z}$ were observed. In the (accelerated) projected gradient method of [10, 11], the optimality criterion is the squared distance $\|\tilde{y} - x\|^2$, where

$$\tilde{y} = \arg\min_{y \in X} \left\{ \langle \nabla F(x), y \rangle + \frac{1}{2}\|y - x\|^2 \right\}.$$

Evidently, $\tilde{y} = \bar{y}(x, \tilde{z}, 1)$ and if we could see the dual vector $\tilde{z} = \nabla F(x)$ we would obtain

$$\|\tilde{y} - x\|^2 = V(x, \tilde{z}) = \|\tilde{y} - x\|^2 + \|\tilde{z} - \nabla F(x)\|^2.$$

It should be mentioned that while the above $\hat{z}$ and $\tilde{z}$ are not computable under the stochastic setting, the vector z_k defined in (2.3) is computed every iteration and can be used as an online estimate of $\nabla F(x)$.

3. Convergence analysis. In this section, we provide convergence analysis of Algorithm 2.1. To do so, we need the following assumption.

Assumption 2. The functions f and g and their derivatives are Lipschitz continuous.

This immediately implies that the gradient of the composite function F is Lipschitz continuous.

LEMMA 3.1. *Under Assumption* 2*, the gradient of the function* F *defined in* (1.1) *is Lipschitz continuous with* $L_{\nabla F} := L_g^2 L_{\nabla f} + L_f L_{\nabla g}$.

Proof. Let $x, \hat{x} \in X$ be given. Then, by the chain rule we have

$$\begin{aligned}\|\nabla F(x) - \nabla F(\hat{x})\| &= \|\nabla g(x)^\top \nabla f(g(x)) - \nabla g(\hat{x})^\top \nabla f(g(\hat{x}))\| \\ &\le \|\nabla g(x)\|\|\nabla f(g(x)) - \nabla f(g(\hat{x}))\| + \|\nabla f(g(\hat{x}))\|\|\nabla g(x) - \nabla g(\hat{x})\| \\ &\le (L_g^2 L_{\nabla f} + L_f L_{\nabla g})\|x - \hat{x}\|. \qquad \square\end{aligned}$$

The next result about the subproblem employed at step 2 of Algorithm 2.1 will be used in our convergence analysis.

LEMMA 3.2. *Let* $\eta(x, z)$ *be the optimal value of subproblem* (2.1) *for any* (x, z)*, i.e.,*

$$\eta(x, z) = \min_{y \in X} \left\{ \langle z, y - x \rangle + \frac{\beta}{2}\|y - x\|^2 \right\}. \tag{3.1}$$

Then the gradient of η *w.r.t.* (x, z) *is Lipschitz continuous with the constant*

$$L_{\nabla \eta} = 2\sqrt{(1 + \beta)^2 + (1 + \tfrac{1}{2\beta})^2}.$$

Proof. Let $\bar{y}(x, z) \in X$ be the solution of (3.1). Since the solution is unique, the partial derivatives of the optimal value function η are given by

$$\nabla_x \eta(x, z) = -z + \beta(x - \bar{y}(x, z)), \quad \nabla_z \eta(x, z) = \bar{y}(x, z) - x.$$

Hence, for any (x,z) and $(\hat{x},\hat{z})$, we have

$$\begin{aligned}\|\nabla\eta(x,z)-\nabla\eta(\hat{x},\hat{z})\| &\le \|\nabla_x\eta(x,z)-\nabla_x\eta(\hat{x},\hat{z})\|+\|\nabla_z\eta(x,z)-\nabla_z\eta(\hat{x},\hat{z})\|\\ &\le 2(1+\beta)\|x-\hat{x}\|+(2+1/\beta)\|z-\hat{z}\|\le L_{\nabla\eta}\|(x,z)-(\hat{x},\hat{z})\|,\end{aligned}$$

where the inequalities follow from the nonexpansiveness of the projection operator and the Cauchy–Schwarz inequality, respectively. ∎

The proof of convergence of Algorithm 2.1 follows from the analysis of the following merit function:

$$W(x,z,u)=a(F(x)-F^*)-\eta(x,z)+\frac{\gamma}{2}\|g(x)-u\|^2, \tag{3.2}$$

where $\gamma>0$ and $\eta(x,z)$ is the optimal value of subproblem (3.1).

LEMMA 3.3. *Let $\{x^k,z^k,y^k,u^k\}_{k\ge0}$ be the sequence generated by Algorithm* 2.1. *Also assume that Assumption* 2 *holds, and*

$$2(a\beta-c)(\gamma b-2c)\ge L_g^2(aL_{\nabla f}+\gamma)^2 \tag{3.3}$$

for some positive constants c and γ. Then

$$c\sum_{k=0}^{N-1}\tau_k\left(\|d^k\|^2+\|g(x^k)-u^k\|^2\right)\le W(x^0,z^0,u^0)+\sum_{k=0}^{N-1}r^{k+1}\qquad\forall N\ge1, \tag{3.4}$$

where, for any $k\ge0$,

$$\begin{aligned}d^k&=y^k-x^k,\\ r^{k+1}&=\frac{\tau_k^2}{2}\Big([aL_{\nabla F}+L_{\nabla\eta}+\gamma L_g^2+2aL_g^2L_{\nabla f}]\|d^k\|^2+b^2\|g(x^{k+1})-G^{k+1}\|^2\Big)\\ &\quad+\tau_k\Big(\gamma b(1-b\tau_k)\langle g(x^{k+1})-u^k,\Delta_k^g\rangle+a\langle d^k,\Delta_k^F\rangle\Big)+\frac{L_{\nabla\eta}}{2}\|z^{k+1}-z^k\|^2,\\ \Delta_k^g&=g(x^{k+1})-G^{k+1},\qquad \Delta_k^F:=\nabla g(x^{k+1})^\top\nabla f(u^k)-\left[J^{k+1}\right]^\top s^{k+1}.\end{aligned} \tag{3.5}$$

Proof. We estimate the decrease of the three terms of the function $W(x,z,u)$ in iteration k.

1. Due to Assumption 2 and in view of Lemma 3.1, we have

$$F(x^k)-F(x^{k+1})\ge\langle\nabla F(x^{k+1}),x^k-x^{k+1}\rangle-\frac{L_{\nabla F}}{2}\|x^k-x^{k+1}\|^2.$$

After rearranging the terms and using (2.2), we obtain

$$F(x^{k+1})-F(x^k)\le\tau_k\langle\nabla F(x^{k+1}),d^k\rangle+\frac{L_{\nabla F}\tau_k^2}{2}\|d^k\|^2. \tag{3.6}$$

2. By (2.2), (2.3), and Lemma 3.2, we have

$$\begin{aligned}\eta(x^k,z^k)-\eta(x^{k+1},z^{k+1})&\le\langle z^k+\beta(y^k-x^k),x^{k+1}-x^k\rangle-\langle y^k-x^k,z^{k+1}-z^k\rangle\\ &\quad+\frac{L_{\nabla\eta}}{2}\left[\|x^{k+1}-x^k\|^2+\|z^{k+1}-z^k\|^2\right]\\ &=\tau_k\langle(1+a)z^k+\beta d^k,d^k\rangle-a\tau_k\langle d^k,\left[J^{k+1}\right]^Ts^{k+1}\rangle\\ &\quad+\frac{L_{\nabla\eta}}{2}\left[\|x^{k+1}-x^k\|^2+\|z^{k+1}-z^k\|^2\right].\end{aligned} \tag{3.7}$$

Due to the optimality condition of subproblem (2.1), we have $\langle z^k+\beta(y^k-x^k), y-y^k\rangle \geq 0$ for all $y \in X$, which together with the choice of $y = x^k$ implies that

$$\langle z^k, d^k\rangle + \beta\|d^k\|^2 \leq 0. \tag{3.8}$$

Combining the last relation with (3.7), we obtain

$$\begin{aligned}\eta(x^k, z^k) - \eta(x^{k+1}, z^{k+1}) \leq & -a\beta\tau_k\|d^k\|^2 - a\tau_k\langle d^k, \nabla g(x^{k+1})^\top \nabla f(u^k)\rangle \\ & + a\tau_k\langle d^k, \nabla g(x^{k+1})^\top \nabla f(u^k) - \left[J^{k+1}\right]^\top s^{k+1}\rangle \\ & + \frac{L_{\nabla\eta}}{2}\left[\|x^{k+1}-x^k\|^2 + \|z^{k+1}-z^k\|^2\right].\end{aligned} \tag{3.9}$$

3. By (2.4) we have

$$\begin{aligned}\|g(x^{k+1}) - u^{k+1}\|^2 = & (b\tau_k)^2\|g(x^{k+1}) - G^{k+1}\|^2 + (1-b\tau_k)^2\|g(x^{k+1}) - u^k\|^2 \\ & + 2b\tau_k(1-b\tau_k)\langle g(x^{k+1}) - G^{k+1}, g(x^{k+1}) - u^k\rangle,\end{aligned}$$

$$\|g(x^{k+1}) - u^{k+1}\|^2 \leq \|g(x^k) - u^k\|^2 + L_g^2\tau_k^2\|d^k\|^2 + 2L_g\tau_k\|d^k\|\|g(x^k) - u^k\|, \tag{3.10}$$

where the last inequality follows from (2.2) and the Lipschitz continuity of g. Moreover, using (2.2) and Assumption 2, we obtain

$$\langle d^k, \nabla F(x^{k+1}) - \nabla g(x^{k+1})^\top \nabla f(u^k)\rangle \leq L_g L_{\nabla f}\left[\tau_k L_g\|d^k\|^2 + \|d^k\|\|g(x^k) - u^k\|\right]. \tag{3.11}$$

4. The overall estimate is obtained by combining (3.2), (3.6), (3.9), (3.10), and (3.11):

$$\begin{aligned}& W(x^{k+1}, z^{k+1}, u^{k+1}) - W(x^k, z^k, u^k) \\ & \leq -\tau_k\Big(a\beta\|d^k\|^2 + \frac{\gamma b}{2}\|g(x^k) - u^k\|^2 - (aL_gL_{\nabla f} + \gamma L_g)\|d^k\|\|g(x^k) - u^k\|\Big) + r^{k+1},\end{aligned}$$

where r^{k+1} is defined in (3.5). Hence, when condition (3.3) holds, we have

$$W(x^{k+1}, z^{k+1}, u^{k+1}) - W(x^k, z^k, u^k) \leq -c\tau_k\left(\|d^k\|^2 + \|g(x^k) - u^k\|^2\right) + r^{k+1}.$$

Observe that $\eta(x, z) \leq 0$ for any (x, z), due to (3.8). Therefore, $W(x, z, u) \geq 0$ for all (x, z, u). Summing up the above inequalities for all k, we obtain (3.4). ∎

As a consequence of the above result, we can provide upper bounds for the sequences generated by Algorithm 2.1.

PROPOSITION 3.4. *Let $\{x^k, z^k, y^k, u^k\}_{k\geq 0}$ be the sequence generated by Algorithm 2.1 and Assumption 1 holds. Then*

(a) *if $\tau_0 = 1/a$, we have*

$$\beta^2\mathbb{E}\left[\|d^k\|^2|\mathcal{F}_{k-1}\right] \leq \mathbb{E}\left[\|z^k\|^2|\mathcal{F}_{k-1}\right] \leq \sigma_J^2\sigma_s^2 \qquad \forall k \geq 1;$$

(b) *if Assumption 2 also holds and $a\tau_k \leq 1/\sqrt{2}$ for all $k \geq 1$, we have*

$$\begin{aligned}\sum_{k=0}^{\infty}\mathbb{E}[\|z^{k+1} - z^k\|^2|\mathcal{F}_k] &\leq 2\left[\|z^0\|^2 + 24a^2\sigma_J^2\sigma_s^2\sum_{k=0}^{\infty}\tau_k^2\right], \\ \sum_{k=0}^{\infty}\mathbb{E}[r^{k+1}|\mathcal{F}_k] &\leq \sigma^2\sum_{k=0}^{\infty}\tau_k^2,\end{aligned} \tag{3.12}$$

where

(3.13)
$$\sigma^2 = \frac{1}{2}\Big([L_{\nabla F} + L_{\nabla\eta} + \gamma L_g^2 + 2aL_g^2 L_{\nabla f}]\frac{\sigma_J^2\sigma_s^2}{\beta^2} + b^2\sigma_g^2 + 4L_{\nabla\eta}\big[\|z^0\|^2 + 24a^2\sigma_J^2\sigma_s^2\big]\Big).$$

Proof. We first show part (a). The first inequality follows immediately from (3.8) and the Cauchy–Schwarz inequality. Also, defining

$$\Gamma_1 := \begin{cases} 1, & \tau_0 = 1/a, \\ 1 - a\tau_0, & \tau_0 < 1/a, \end{cases} \qquad \Gamma_k := \Gamma_1 \prod_{i=1}^{k-1}(1 - a\tau_i) \quad \forall k \geq 2, \tag{3.14}$$

and noting (2.3), we obtain

$$\frac{z^1}{\Gamma_1} = \frac{(1-a\tau_0)z^0}{\Gamma_1} + \frac{a\tau_0}{\Gamma_1}\big[J^1\big]^\top s^1, \qquad \frac{z^{k+1}}{\Gamma_{k+1}} = \frac{z^k}{\Gamma_k} + \frac{a\tau_k}{\Gamma_{k+1}}\big[J^{k+1}\big]^\top s^{k+1} \quad \forall k \geq 1.$$

Summing up the above inequalities and assuming that $\tau_0 = 1/a$, we obtain, for any $k \geq 1$,

$$z^k = \sum_{i=0}^{k-1} \alpha_{i,k}\big[J^{i+1}\big]^\top s^{i+1}, \quad \alpha_{i,k} = \frac{a\tau_i}{\Gamma_{i+1}}\Gamma_k, \quad \sum_{i=0}^{k-1}\alpha_{i,k} = 1, \tag{3.15}$$

where the last equality follows from the fact that

$$\sum_{i=0}^{k-1}\frac{a\tau_i}{\Gamma_{i+1}} = \frac{a\tau_0}{\Gamma_1} + \sum_{i=1}^{k-1}\frac{a\tau_i}{(1-a\tau_i)\Gamma_i} = \frac{a\tau_0}{\Gamma_1} + \sum_{i=1}^{k-1}\left(\frac{1}{\Gamma_{i+1}} - \frac{1}{\Gamma_i}\right) = \frac{1}{\Gamma_k} - \frac{1-a\tau_0}{\Gamma_1}.$$

Therefore, noting that $\|\cdot\|^2$ is a convex function and using Assumption 1, we conclude that

$$\mathbb{E}\big[\|z^k\|^2\big|\mathcal{F}_{k-1}\big] \leq \sum_{i=0}^{k-1}\alpha_{i,k}\mathbb{E}\big[\|J^{i+1}\|^2\big|\mathcal{F}_i\big]\mathbb{E}\big[\|s^{i+1}\|^2\big|\mathcal{F}_i\big] \leq \sigma_J^2\sigma_s^2\sum_{i=0}^{k-1}\alpha_{i,k} = \sigma_J^2\sigma_s^2.$$

We now show part (b). By (2.3), the above estimate, and assuming that $\tau_0 = 1/a$, we have

$$\begin{aligned}
\mathbb{E}\big[\|z^1 - z^0\|^2\big|\mathcal{F}_0\big] &\leq 2\Big(\|z^0\|^2 + \mathbb{E}\big[\|[J^1]^\top s^1\|^2\big|\mathcal{F}_0\big]\Big) \leq 2\Big(\|z^0\|^2 + \sigma_J^2\sigma_s^2\Big), \\
\mathbb{E}\big[\|z^{k+1} - z^k\|^2\big|\mathcal{F}_k\big] &\leq \frac{2a^2\tau_k^2}{(1-a\tau_k)^2}\Big(\mathbb{E}\big[\|z^{k+1}\|^2\big|\mathcal{F}_k\big] + \mathbb{E}\big[\|[J^{k+1}]^\top s^{k+1}\|^2\big|\mathcal{F}_k\big]\Big) \\
&\leq \frac{4a^2\sigma_J^2\sigma_s^2\tau_k^2}{(1-a\tau_k)^2} \qquad \forall k \geq 1,
\end{aligned}$$

implying that

$$\sum_{k=0}^{\infty}\mathbb{E}[\|z^{k+1} - z^k\|^2|\mathcal{F}_k] \leq 2\left[\|z^0\|^2 + \sigma_J^2\sigma_s^2\left(1 + 2a^2\sum_{k=1}^{\infty}\big(\frac{\tau_k}{1-a\tau_k}\big)^2\right)\right].$$

Combining the above inequality with the fact that $\frac{1}{(1-a\tau_k)^2} \leq 12$ due to the assumption that $a\tau_k \leq 1/\sqrt{2}$, we obtain the first inequality in (b). Finally, due to the equation

$$\mathbb{E}\big[\langle g(x^{k+1}) - u^k, \Delta_k^g\rangle\big|\mathcal{F}_k\big] = \mathbb{E}\big[\langle d^k, \Delta_k^F\rangle\big|\mathcal{F}_k\big] = 0,$$

the second inequality in (b) follows from the first one in (a) and (3.5). ∎

We are now ready to estimate the quality of the iterates generated by Algorithm 2.1. In view of Lemma 2.2, we can bound the optimality measure at iteration k as follows:

$$V(x^k, z^k) \le \max(1, \beta^2)\, \|d^k\|^2 + \|z^k - \nabla F(x^k)\|^2. \tag{3.16}$$

THEOREM 3.5. *Suppose Assumptions* 1 *and* 2 *are satisfied and let*

$$\{x^k, z^k, y^k, u^k\}_{k\ge 0}$$

be the sequence generated by Algorithm 2.1. *Moreover, assume that the parameters are chosen so that* (3.3) *holds and step sizes* $\{\tau_k\}$ *are deterministic and satisfy*

$$\sum_{i=k+1}^{N} \tau_i \Gamma_i \le \bar{c}\Gamma_{k+1} \quad \forall k \ge 0 \ \text{ and } \ \forall N \ge 1, \ \text{ and } \ a\tau_k \le 1/\sqrt{2} \quad \forall k \ge 1, \ \ \tau_0 = \frac{1}{a}, \tag{3.17}$$

where Γ_k *is defined in* (3.14), *and* $\bar{c}$ *is a positive constant. Then*

(a) *for every* $N \ge 2$, *we have*

$$\sum_{k=1}^{N} \tau_k \mathbb{E}\big[\|\nabla F(x^k) - z^k\|^2 \big| \mathcal{F}_{k-1}\big] \le a\bar{c}\left(\frac{1}{c}\max(L_1, L_2)\sigma^2 + 2a\sigma_J^2\sigma_s^2\right)\left(\sum_{k=0}^{N-1}\tau_k^2\right) + \frac{a\bar{c}}{c}\max(L_1, L_2) W(x^0, z^0, u^0), \tag{3.18}$$

where

$$L_1 := \frac{2L_{\nabla F}^2}{a^2} + 4L_g^4 L_{\nabla f}^2, \qquad L_2 := 4L_g^2 L_{\nabla f}^2; \tag{3.19}$$

(b) *as a consequence, we have*

$$\begin{aligned}\mathbb{E}\big[V(x^R, z^R)\big] &\le \frac{1}{\sum_{k=1}^{N-1}\tau_k}\left\{a\bar{c}\left(\frac{1}{c}\big[\max(L_1, L_2) + \max(1, \beta^2)\big]\sigma^2 + 2a\sigma_J^2\sigma_s^2\right)\left(\sum_{k=0}^{N-1}\tau_k^2\right)\right. \\ &\quad \left. + \frac{1}{c}\Big(a\bar{c}\max(L_1, L_2) + \max(1, \beta^2)\Big) W(x^0, z^0, u^0)\right\},\end{aligned} \tag{3.20}$$

where the expectation is taken with respect to all random sequences generated by the method and an independent random integer number $R \in \{1, \dots, N-1\}$, *whose probability distribution is given by*

$$P[R = k] = \frac{\tau_k}{\sum_{j=1}^{N-1}\tau_j}; \tag{3.21}$$

(c) *moreover, if* $a = b = 1$, *if the regularization coefficient is equal to*

$$\beta = \left(\frac{(1+\alpha)^2}{\alpha} L_g^2 + \frac{\alpha}{4}\right) L_{\nabla f} \tag{3.22}$$

for some $\alpha > 0$, *and if the step sizes are equal to*

$$\tau_0 = 1, \qquad \tau_k \equiv \frac{1}{\sqrt{N}} \qquad \forall k = 1, \dots, N-1, \tag{3.23}$$

then

(3.24)
$$\mathbb{E}\big[V(x^R, z^R)\big] \le \frac{4}{\sqrt{N}-1}\left(\frac{2}{\alpha L_{\nabla F}}\big[\max(L_1, L_2) + \max(1, \beta^2)\big]\big[W(x^0, z^0, u^0) + \sigma^2\big] + \sigma_f^2\sigma_s^2\right)$$

and

$$\mathbb{E}\big[\|g(x^R) - u^R\|^2\big] \le \frac{W(x^0, z^0, u^0) + \sigma^2}{\alpha L_{\nabla F}(\sqrt{N}-1)}. \tag{3.25}$$

Proof. We first show part (a). By (2.3), we have

$$\begin{aligned}&\nabla F(x^{k+1}) - z^{k+1}\\ &= (1 - a\tau_k)[\nabla F(x^k) - z^k + \nabla F(x^{k+1}) - \nabla F(x^k)] + a\tau_k[\nabla F(x^{k+1}) - [J^{k+1}]^\top s^{k+1}].\end{aligned}$$

Dividing both sides of the above inequality by Γ_{k+1}, summing them up, noting the fact that $\tau_0 = 1/a$, similarly to (3.15), we obtain

$$\nabla F(x^k) - z^k = \sum_{i=0}^{k-1} \alpha_{i,k}\left[e_i + \Delta_i^F\right] \qquad \forall k \ge 1$$

with

$$e_i := \frac{(1 - a\tau_i)}{a\tau_i}\big[\nabla F(x^{i+1}) - \nabla F(x^i)\big] + \nabla F(x^{i+1}) - \nabla g(x^{i+1})^\top \nabla f(u^i), \tag{3.26}$$

where Δ_i^F is defined in (3.5). Hence,

$$\nabla F(x^{k-1}) - z^{k-1} = \sum_{i=0}^{k-2} \alpha_{i,k-1}\left[e_i + \Delta_i^F\right] = \frac{\Gamma_{k-1}}{\Gamma_k}\sum_{i=0}^{k-2} \alpha_{i,k}\left[e_i + \Delta_i^F\right],$$

which together with (3.26) implies that

$$\begin{aligned}&\|\nabla F(x^k) - z^k\|^2\\ &= \Big\|\frac{\Gamma_k}{\Gamma_{k-1}}\big[\nabla F(x^{k-1}) - z^{k-1}\big] + \alpha_{k-1,k}\left[e_{k-1} + \Delta_{k-1}^F\right]\Big\|^2\\ &= \big\|(1 - a\tau_{k-1})\big[\nabla F(x^{k-1}) - z^{k-1}\big] + a\tau_{k-1}\left[e_{k-1} + \Delta_{k-1}^F\right]\big\|^2\\ &= \big\|(1 - a\tau_{k-1})\big[\nabla F(x^{k-1}) - z^{k-1}\big] + a\tau_{k-1}e_{k-1}\big\|^2 + a^2\tau_{k-1}^2\big\|\Delta_{k-1}^F\big\|^2\\ &\quad + 2a\tau_{k-1}\big\langle(1 - a\tau_{k-1})\big[\nabla F(x^{k-1}) - z^{k-1}\big] + a\tau_{k-1}e_{k-1}, \Delta_{k-1}^F\big\rangle\\ &\le (1 - a\tau_{k-1})\big\|\nabla F(x^{k-1}) - z^{k-1}\big\|^2 + a\tau_{k-1}\big\|e_{k-1}\big\|^2 + a^2\tau_{k-1}^2\|\Delta_{k-1}^F\big\|^2\\ &\quad + 2a\tau_{k-1}\big\langle(1 - a\tau_{k-1})[\nabla F(x^{k-1}) - z^{k-1}] + a\tau_{k-1}e_{k-1}, \Delta_{k-1}^F\big\rangle,\end{aligned}$$

where the inequality follows from the convexity of $\|\cdot\|^2$. Dividing both sides of the above inequality by Γ_k, using (3.14), summing all the resulting inequalities, and noting the facts that $\tau_0 = 1/a$, and $\tau_k \le 1/a$, we obtain

$$\|\nabla F(x^k) - z^k\|^2 \le \Gamma_k\left[\sum_{i=0}^{k-1}\left(\frac{a\tau_i}{\Gamma_{i+1}}\|e_i\|^2 + \frac{a^2\tau_i^2}{\Gamma_{i+1}}\|\Delta_i^F\|^2 + \frac{2a\tau_i\delta_i}{\Gamma_{i+1}}\right)\right], \tag{3.27}$$

where

$$\delta_i := \langle (1-a\tau_i)[\nabla F(x^i) - z^i] + a\tau_i e_i, \Delta_i^F \rangle.$$

Now, using (2.2), (3.11), with a view to Lemma 3.1, we obtain

$$\|e_i\|^2 \le \frac{2L_{\nabla F}^2(1-a\tau_i)^2}{a^2}\|d^i\|^2 + 4L_g^2 L_{\nabla f}^2\Big[\tau_i^2 L_g^2\|d^i\|^2 + \|g(x^i)-u^i\|^2\Big],$$

which together with (3.19) implies that

(3.28)

$$\begin{aligned}\sum_{k=1}^{N}\left(\tau_k\Gamma_k\sum_{i=0}^{k-1}\frac{a\tau_i}{\Gamma_{i+1}}\|e_i\|^2\right) &\le \sum_{k=1}^{N}\left(\tau_k\Gamma_k\sum_{i=0}^{k-1}\frac{a\tau_i}{\Gamma_{i+1}}\left[L_1\|d^i\|^2 + L_2\|g(x^i)-u^i\|^2\right]\right)\\ &= a\sum_{k=0}^{N-1}\left\{\frac{\tau_k}{\Gamma_{k+1}}\Big[\sum_{i=k+1}^{N}\tau_i\Gamma_i\Big]\Big[L_1\|d^k\|^2 + L_2\|g(x^k)-u^k\|^2\Big]\right\}\\ &\le a\bar{c}\sum_{k=0}^{N-1}\left\{\tau_k\left[L_1\|d^k\|^2 + L_2\|g(x^k)-u^k\|^2\right]\right\},\end{aligned}$$

where the last inequality follows from the condition (3.17). In a similar way,

(3.29)

$$\sum_{k=1}^{N}\left(\tau_k\Gamma_k\sum_{i=0}^{k-1}\frac{a^2\tau_i^2}{\Gamma_{i+1}}\|\Delta_i^F\|^2\right) = a^2\sum_{k=0}^{N-1}\frac{\tau_k^2}{\Gamma_{k+1}}\Big(\sum_{i=k+1}^{N}\tau_i\Gamma_i\Big)\|\Delta_k^F\|^2 \le \bar{c}a^2\sum_{k=0}^{N-1}\tau_k^2\|\Delta_k^F\|^2.$$

Moreover, under Assumption 1 we have

$$\mathbb{E}[\|\Delta_k^F\|^2|\mathcal{F}_k] \le 2\mathbb{E}[\left[J^{k+1}\right]^\top s^{k+1}|\mathcal{F}_k] \le 2\sigma_J^2\sigma_s^2, \qquad \mathbb{E}[\delta_k|\mathcal{F}_k] = 0.$$

Therefore, by taking the conditional expectation of both sides of (3.27), noting (3.28), (3.29), the above inequality, and Lemma 3.3, we obtain (3.18). Part (b) then follows from (3.16) and the facts that

$$\begin{aligned}&c\sum_{k=0}^{N-1}\tau_k\mathbb{E}[\|d^k\|^2|\mathcal{F}_k] \le W(x^0,z^0,u^0) + \sigma^2\sum_{k=0}^{N-1}\tau_k^2,\\ &\mathbb{E}\big[V(x^R,z^R)\big] = \frac{\sum_{k=1}^{N-1}\tau_k\mathbb{E}\big[V(x^k,z^k)\big]}{\sum_{k=1}^{N-1}\tau_k},\end{aligned}$$

due to (3.4), (3.12), and (3.21).

To show part (c), observe that condition (3.3) is satisfied by (3.22) and the choice of $\gamma = 4c = \alpha L_{\nabla f}$. Also by (3.14) and (3.23), we have

$$\begin{aligned}&\sum_{k=1}^{N-1}\tau_k \ge \sqrt{N}-1, \qquad \sum_{k=0}^{N-1}\tau_k^2 \le 2, \qquad \Gamma_k = \Big(1-\frac{1}{\sqrt{N}}\Big)^{k-1},\\ &\sum_{i=k+1}^{N}\tau_i\Gamma_i = \Big(1-\frac{1}{\sqrt{N}}\Big)^{k}\frac{1}{\sqrt{N}}\sum_{i=0}^{N-k-1}\Big(1-\frac{1}{\sqrt{N}}\Big)^{i} \le \Big(1-\frac{1}{\sqrt{N}}\Big)^{k},\end{aligned}$$

which ensures (3.17) with $\bar{c}=1$ and, hence, together with (3.21), implies (3.24). ∎

We now add a few remarks about the above result. First, the estimate (3.24) implies that to find an approximate stationary point $(\bar{x}, \bar{z})$ of problem (1.1) satisfying $\mathbb{E}[V(\bar{x}, \bar{z})] \leq \varepsilon$, Algorithm 2.1 requires at most $\mathcal{O}(1/\varepsilon^2)$ iterations (stochastic gradients), which is better than $\mathcal{O}(1/\varepsilon^{2.25})$ obtained in [28] for unconstrained nonconvex stochastic optimization. Our complexity bound indeed matches the sample complexity of the stochastic gradient method for the general (single-level) smooth nonconvex optimization problem [9]. It is also consistent with the central limit theorem for composite risk functionals [5], because the objective value gap is essentially proportional to the squared gradient norm at an approximate stationary point (in the unconstrained case). Second, Theorem 3.5 provides the same convergence rate of Algorithm 2.1 for both constrained and unconstrained problems: we can still get sample complexity of $\mathcal{O}(1/\varepsilon^2)$ with taking only one sample per iteration in the constrained case. To the best of our knowledge, this is the first direct convergence analysis for constrained nonconvex stochastic optimization providing the aforementioned sample complexity. Third, (3.25) provides not only accuracy bounds for the approximate stationary point x^R but also squared error bounds for estimating the exact value of the inner function, $g(x^R)$, by the second running average, u^R. As a result, Algorithm 2.1 provides not only accurate approximations to the stationary solution but also reliable estimates of the gradients. Finally, note that Assumption 1 implies that derivatives of f and g are bounded. Hence, to establish the results of Theorem 3.5, we can relax Assumption 2 and require Assumption 1 together with Lipschitz continuity of the derivatives of f and g.

For the stochastic variational inequality (SVI) problem of Example 1.1, our method has significantly lower complexity and faster convergence than the approach of [13]. Most of the literature on stochastic methods for SVI requires monotonicity or even strong monotonicity of the operator $H(\cdot)$ (see, e.g., [14, 17] and the references within). The authors in [13] consider SVI with operators $\mathbb{E}[H(\cdot)]$ satisfying a weaker pseudomonotonicity assumption. With the use of a variance reduction technique by increasing sizes of minibatches, a generalization of the extragradient method originating in [16] was developed, with the oracle complexity $\mathcal{O}(\varepsilon^{-2})$, which is matched by our results. The recent manuscript [19] considers a special case of SVI, a stochastic saddle point problem, with weakly convex-concave functions. By employing a proximal point scheme, the resulting SVI is converted to a monotone one, and approximately solved by a stochastic algorithm. The resulting two-level scheme has the rate of convergence $\mathcal{O}(\varepsilon^{-2})$, if acceleration techniques are used. Our approach does not make any monotonicity assumption and achieves the oracle complexity of $\mathcal{O}(\varepsilon^{-2})$ as well. This is due to the use of a one-level method for a special merit function (3.2), involving the lifted gap function (1.4) as one of its components (the other being a projection mapping using its gradient).

We also have the following asymptotic convergence result.

THEOREM 3.6. *Assume that the sequence of step sizes satisfies*

$$\sum_{k=1}^{\infty} \tau_k = +\infty \quad a.s., \qquad \mathbb{E} \sum_{k=1}^{\infty} \tau_k^2 < \infty. \tag{3.30}$$

Then a constant $\bar{a} > 0$ exists such that, for all $a \in (0, \bar{a})$, with probability 1, every accumulation point (x^*, z^*, u^*) of the sequence $\{x^k, z^k, u^k)$ generated by Algorithm 2.1 satisfies the conditions

$$z^* = \big[\nabla g(x^*)\big]^T \nabla f(u^*), \quad u^* = g(x^*), \quad -z^* \in \mathcal{N}_X(x^*).$$

Proof. Note that the sequence $\{r^k\}$ defined in (3.5) is adapted to $\{\mathcal{F}_k\}$ and summable almost surely under Assumption 1 and (3.30). Therefore, (3.4) implies that almost surely

$$\lim_{k\to\infty} \inf \|d^k\| = 0, \qquad \lim_{k\to\infty} \inf \|g(x^k) - u^k\| = 0.$$

Using the techniques of [23, Lem. 8], we can then show that with probability 1, $d^k \to 0$ and $g(x^k) - u^k \to 0$. Following analysis similar to [23], we prove that each convergent subsequence of $\{(x^k, z^k, u^k)\}$ converges to a stationary point of problem (1.1), the corresponding gradient of F, and the value of g. ∎

In the convergence theorem, we allow the step sizes to be random (but adapted to the filtration $\{\mathcal{F}_k\}$), because earlier experience indicates that adaptation of the step sizes, based on the information gathered along the path, greatly improves performance of stochastic subgradient methods [25]. We plan to explore this avenue for composition optimization in our future research.

4. Dual averaging with constraints. Although our main interest is in composite optimization, we also provide a simplified variant of Algorithm 2.1 for solving a single-level stochastic optimization. Our techniques allow the same convergence analysis for both constrained and unconstrained problems which removes the necessity of forming minibatches of samples per iteration for constrained problems (see, e.g., [11, 10]). Moreover, this variant of the NASA method, different from the existing stochastic-approximation–type methods, is a parameter-free algorithm in the sense that its step size policy does not depend on any problem parameters and allows for random (history dependent) step sizes. This algorithmic feature is more important under the stochastic setting since estimating problem parameters becomes more difficult.

Throughout this section, we assume that the inner function g in (1.1) is the identity map, i.e., $g(x) \equiv x$, and only noisy information about f is available. In this case, our problem is reduced to

$$\min_{x\in X} f(x), \tag{4.1}$$

and it is easy to verify that

$$G = g, \qquad J = I, \qquad \sigma_G = 0, \qquad \sigma_J = 1, \qquad L_g = 1, \qquad L_{\nabla g} = 0. \tag{4.2}$$

Moreover, Algorithm 2.1 can be simplified as follows.

Algorithm 4.1 The averaged stochastic approximation method.

Replace step 2 of Algorithm 2.1 with the following:

2′. Call the stochastic oracle to obtain s^{k+1} at x^k and update the "running average" as

$$z^{k+1} = (1 - a\tau_k)z^k + a\tau_k s^{k+1}. \tag{4.3}$$

The above algorithm differs from Algorithm 2.1 in two aspects. First, stochastic approximation of ∇g is replaced by its exact value, the identity matrix. Second, the averaged sequence in (2.4) is not required and u^k is simply set to x^k, the exact value of $g(x^k)$.

The resulting method belongs to the class of algorithms with direction averaging (multistep) methods. The literature on these methods for unconstrained problems is very rich. They were initiated in [26] and developed and analyzed in [12, 15, 21] and other works. Recently, these methods play a role in machine learning, under the name of *dual averaging methods* (see [29] and the references therein). In all these versions, two timescales were essential for convergence. The first single-timescale method was proposed in [25], with convergence analysis based on a different Lyapunov function suitable for unconstrained problems. Our version is related to [22, 23], where a similar approach to constrained problems was proposed, albeit without rate of convergence estimates. We may remark here that the version of [22, 23] calculates s^{k+1} at x^{k+1} rather than at x^k at step $2'$, which is essential for nonsmooth weakly convex $f(\cdot)$. For smooth functions both ways are possible and can be analyzed in the same way with minor adjustments.

Convergence analysis of Algorithm 4.1 follows directly from that of Algorithm 2.1 by simplifying the definition of the merit function as

$$W(x,z) = a(f(x) - f^*) - \eta(x,z), \tag{4.4}$$

exactly as used in [22, 23]. We then have the following result.

LEMMA 4.1. *Let $\{x^k, z^k, y^k\}_{k\ge 0}$ be the sequence generated by Algorithm* 4.1. *Also assume that function f has Lipschitz continuous gradient. Then*

(a) *for any $N \ge 2$, we have*

$$\beta \sum_{k=0}^{N-1} \tau_k \|d^k\|^2 \le W(x^0,z^0) + \sum_{k=0}^{N-1} r^{k+1}, \tag{4.5}$$

where, for any $k \ge 0$,

$$\begin{aligned} &d^k = y^k - x^k, \qquad \Delta_k^f = \nabla f(x^k) - s^{k+1}, \\ &r^{k+1} = \frac{1}{2}(3aL_{\nabla f} + L_{\nabla\eta})\tau_k^2\|d^k\|^2 + a\tau_k\langle d^k, \Delta_k^f\rangle + \frac{1}{2}L_{\nabla\eta}\|z^{k+1} - z^k\|^2; \end{aligned} \tag{4.6}$$

(b) *if, in addition, Assumption* 1 *holds along with* (4.2), *$a\tau_k \le 1/\sqrt{2}$ for all $k \ge 1$, and $\tau_0 = 1/a$, we have*

$$\begin{aligned} &\beta^2\mathbb{E}[\|d^k\|^2|\mathcal{F}_{k-1}]] \le \mathbb{E}[\|z^k\|^2|\mathcal{F}_{k-1}]] \le \sigma_s^2 \qquad \forall k \ge 1, \\ &\sum_{k=0}^{\infty} \mathbb{E}[\|z^{k+1} - z^k\|^2|\mathcal{F}_k] \le 2\left[\|z^0\|^2 + 24a^2\sigma_s^2\sum_{k=0}^{\infty}\tau_k^2\right], \\ &\sum_{k=0}^{\infty} \mathbb{E}[r^{k+1}|\mathcal{F}_k] \le \sigma^2\sum_{k=0}^{\infty}\tau_k^2, \quad \sigma^2 = \frac{(3aL_{\nabla f} + L_{\nabla\eta})\sigma_s^2}{2\beta^2} + 2L_{\nabla\eta}\big[\|z^0\|^2 + 24a^2\sigma_s^2\big]. \end{aligned} \tag{4.7}$$

Proof. Multiplying (3.6) by a, summing it up with (3.9), noting (4.2), (4.4), and the fact that

$$a\tau_k\langle d^k, \nabla f(x^{k+1}) - \nabla f(x^k)\rangle \le aL_{\nabla f}\tau_k^2\|d^k\|^2,$$

we obtain

$$W(x^{k+1},z^{k+1}) - W(x^k,z^k) \le -a\beta\tau_k\|d^k\|^2 + r^{k+1}. \tag{4.8}$$

The remainder of the proof is similar to that of Lemma 4.1 and Proposition 3.4; hence, we skip the details. ∎

Using the above results, we can provide the main convergence property of Algorithm 4.1.

THEOREM 4.2. *Let $\{x^k, z^k, y^k\}_{k\geq 0}$ be the sequence generated by Algorithm* 4.1 *with $a = 1$. Moreover, assume that Assumption* 1 *holds along with* (4.2)*, and the step sizes are set to* (3.23)*. Then, for a random R distributed according to* (3.21)*, we have*

$$\mathbb{E}\big[V(x^R, z^R)\big] \leq \frac{1}{\sqrt{N}-1}\left(\frac{1}{\beta}(\max(1,\beta^2) + L_{\nabla f}^2)\big[W(x^0,z^0) + 2\sigma^2\big] + 4\sigma_s^2\right), \tag{4.9}$$

where $V(x,z)$ and σ^2 are, respectively, defined in (2.6) *and* (4.7)*.*

Proof. Similarly to (3.26), we have

$$\nabla F(x^k) - z^k = \sum_{i=0}^{k-1} \alpha_{i,k}\left[e_i + \Delta_i^f\right], \quad e_i := \frac{\nabla f(x^{i+1}) - \nabla f(x^i)}{a\tau_i}, \tag{4.10}$$

which, together with the Lipschitz continuity of ∇f and (2.2), imply that

$$\|e_i\|^2 \leq \frac{L_{\nabla f}^2 \|d^i\|^2}{a^2}. \tag{4.11}$$

In view of Lemma 4.1, the rest of the proof is similar to that of Theorem 3.5. ∎

It is worth noting that unlike Algorithm 2.1, the regularization coefficient β in Algorithm 4.1, due to (4.8), can be set to any positive constant number to achieve the sample (iteration) complexity of $\mathcal{O}(1/\varepsilon^2)$. Such a result has not been obtained before for a parameter-free algorithm for smooth nonconvex stochastic optimization. Moreover, Algorithm 4.1, similarly to Algorithm 2.1, outputs a pair (x^R, z^R), where z^R is an accurate estimate of $\nabla f(x^R)$ without taking any additional samples. This is important for both unconstrained and constrained problems, where one can use the quantity $\max(1,\beta)\|y^k - x^k\|$ as an online certificate of the quality of the current solution; see Lemma 2.2.

Note that the convergence result of Theorem 4.2 is established under the boundedness assumption of the second moment of the stochastic gradient. In the remainder of this section, we modify the convergence analysis of Algorithm 4.1 under a relaxed assumption that only variance of the stochastic gradient is bounded. This assumption, which is common in the literature on smooth stochastic optimization, is stated as follows.

Assumption 3. For each k, the stochastic oracle delivers a random vector $s^{k+1} \in \mathbb{R}^n$ such that

$$\mathbb{E}[s^{k+1}|\mathcal{F}_k] = \nabla f(x^k), \qquad \mathbb{E}[\|s^{k+1} - \nabla f(x^k)\|^2|\mathcal{F}_k] \leq \hat{\sigma}_s^2.$$

LEMMA 4.3. *Let $\{x^k, z^k, y^k\}_{k\geq 0}$ be the sequence generated by Algorithm* 4.1*. Also assume that the function f has a Lipschitz continuous gradient. Then*

(a) *for any $N \geq 2$, we have*

$$\sum_{k=1}^{N-1} \tau_k \left(\beta - \tfrac{(3aL_{\nabla f} + L_{\nabla \eta})\tau_k}{2}\right)\|d^k\|^2 \leq W(x^0,z^0) + \sum_{k=0}^{N-1} \hat{r}^{k+1}, \tag{4.12}$$

where, for any $k \geq 0$,

$$\hat{r}^{k+1} = \big\langle a\tau_k d^k - a^2\tau_k^2 L_{\nabla\eta}(\nabla F(x^k) - z^k), \Delta_k^f\big\rangle + \frac{1}{2}a^2\tau_k^2 L_{\nabla\eta}\left[\|\nabla F(x^k) - z^k\|^2 + \|\Delta_k^f\|^2\right]; \tag{4.13}$$

(b) *if, in addition, Assumption* 3 *holds and step sizes* $\{\tau_k\}$ *are chosen such that*

$$\tau_0 = 1/a, \qquad \sum_{i=k+1}^{N} \tau_i^2 \Gamma_i \le \hat{c}\tau_k \Gamma_{k+1} \qquad \forall k \ge 0 \quad \text{and} \quad \forall N \ge 2, \tag{4.14}$$

where Γ_k *is defined in* (3.14) *and* $\hat{c}$ *is a positive constant, we have*

$$\sum_{k=0}^{N-1} \tau_k \left(\beta - \frac{1}{2}(3aL_{\nabla f} + L_{\nabla\eta} + a\hat{c}L_{\nabla\eta}L_{\nabla f}^2)\tau_k\right) \mathbb{E}[\|d^k\|^2|\mathcal{F}_k] \le W(x^0, z^0) + \frac{L_{\nabla\eta}}{2}\|\nabla F(x^0) - z^0\|^2 + \frac{1}{2}a^2(L_{\nabla\eta} + 2\hat{c})\hat{\sigma}_s^2 \sum_{k=0}^{N-1} \tau_k^2. \tag{4.15}$$

Proof. To show part (a), note that by (2.3), (4.2), and (4.13), we have

$$\|z^{k+1} - z^k\|^2 = a^2\tau_k^2\left[\|\nabla F(x^k) - z^k\|^2 + \|\Delta_k^f\|^2 - 2\langle \nabla F(x^k) - z^k, \Delta_k^f\rangle\right],$$

which together with (4.5) and in view of (4.13) implies (4.12). To show part (b), note that by (4.10), (4.11), and, similarly to the proof of Theorem 3.5, part (a), we have

$$\begin{aligned}&\sum_{k=1}^{N} \tau_k^2 \|\nabla F(x^k) - z^k\|^2 \\ &\le \hat{c}\sum_{k=0}^{N-1} \tau_k^2 \left(\frac{L_{\nabla f}^2\|d^k\|^2}{a} + a^2\tau_k\|\Delta_k^f\|^2 + 2a\langle \nabla F(x^{k+1}) - z^k - a\tau_k[\nabla F(x^k) - z^k], \Delta_k^f\rangle\right).\end{aligned}$$

Taking conditional expectation from both sides of the above inequality and using (4.13) under Assumption 3, with the choice of $\tau_0 = 1/a$, we obtain

$$\begin{aligned}&\sum_{k=0}^{N-1} \mathbb{E}[r^{k+1}|\mathcal{F}_k] \\ &\le \frac{a^2(L_{\nabla\eta} + 2\hat{c})\hat{\sigma}_s^2}{2}\sum_{k=0}^{N-1}\tau_k^2 + \frac{a\hat{c}L_{\nabla\eta}L_{\nabla f}^2}{2}\sum_{k=0}^{N-1}\tau_k^2\mathbb{E}[\|d^k\|^2|\mathcal{F}_{k-1}] + \frac{L_{\nabla\eta}}{2}\|\nabla F(x^0) - z^0\|^2\end{aligned}$$

(with the notation of $\mathcal{F}_{-1} \equiv \mathcal{F}_0$), which together with (4.12) implies (4.15). ∎

We can now specialize the convergence rate of Algorithm 4.1 by properly choosing the step size policies.

THEOREM 4.4. *Let* $\{x^k, z^k, y^k\}_{k\ge 0}$ *be the sequence generated by Algorithm* 4.1*, the gradient of* $f(\cdot)$ *be Lipschitz continuous, and step sizes set to* (3.23)*. If Assumption* 3 *holds and*

$$\beta \ge \frac{2(3L_{\nabla f} + L_{\nabla\eta} + \hat{c}L_{\nabla\eta}L_{\nabla f}^2)}{3}, \tag{4.16}$$

then, for a random R *distributed according to* (3.21)*, we have*

$$\mathbb{E}\big[V(x^R, z^R)\big] \le \frac{1}{\sqrt{N} - 1}\left(\frac{6(\max(1, \beta^2) + L_{\nabla f}^2)}{\beta}\Big[W(x^0, z^0) + \frac{L_{\nabla\eta}}{2}\|\nabla F(x^0) - z^0\|^2 + (L_{\nabla\eta} + 2)\hat{\sigma}_s^2\Big] + 2\sigma_s^2\right). \tag{4.17}$$

Proof. First, note that by the choice of step sizes in (3.23), condition (4.14) is satisfied with $\hat{c} = 1$. Moreover, by (4.15) and (4.16), we have

$$\sum_{k=0}^{N-1} \tau_k \mathbb{E}[\|d^k\|^2 | \mathcal{F}_{k-1}] \leq \frac{6}{\beta} \left[W(x^0, z^0) + \frac{L_{\nabla\eta}}{2} \|\nabla F(x^0) - z^0\|^2 + (L_{\nabla\eta} + 2)\hat{\sigma}_s^2 \right],$$

$$\sum_{k=1}^{N} \tau_k \mathbb{E}[\|\nabla F(x^k) - z^k\|^2 | \mathcal{F}_{k-1}] \leq L_{\nabla f}^2 \sum_{k=0}^{N-1} \tau_k \mathbb{E}[\|d^k\|^2 | \mathcal{F}_k] + 2\hat{\sigma}_s^2.$$

Combining the above relations with (3.16), we obtain (4.17). ∎

While the rate of convergence of Algorithm 4.1 in (4.17) is of the same order as in (4.9), the former is obtained under a relaxed assumption on the outputs of the stochastic oracle, as stated in Assumption 3. However, in this case, the regularization coefficient β depends on the problem parameters (like in other algorithms for smooth stochastic optimization).

We also have the following asymptotic convergence result.

THEOREM 4.5. *Assume that the sequence of step sizes satisfy* (3.30). *Then a constant* $\bar{a} > 0$ *exists such that, for all* $a \in (0, \bar{a})$, *with probability* 1, *every accumulation point* (x^*, z^*) *of the sequence* $\{x^k, z^k\}$ *generated by Algorithm* 4.1 *satisfies the conditions*

$$z^* = \nabla f(x^*), \quad -z^* \in \mathcal{N}_X(x^*).$$

Proof. The analysis follows from [23]. ∎

5. Concluding remarks. We have presented a single-timescale stochastic approximation method for smooth nested optimization problems. We showed that the sample complexity bound of this method for finding an approximate stationary point of the problem is of the same order as that of the best-known bound for the stochastic gradient method for single level stochastic optimization problems. Furthermore, our convergence analysis is the same for both unconstrained and constrained cases and does not require batches of samples per iteration. We also presented a simplified parameter-free variant of the NASA method for single-level problems, which enjoys the same complexity bound, regardless of the existence of constraints.

REFERENCES

[1] J. BLANCHET, D. GOLDFARB, G. IYENGAR, F. LI, AND C. ZHOU, *Unbiased Simulation for Optimizing Stochastic Function Compositions*, preprint, https://arxiv.org/abs/1711.07564, 2017.

[2] B. DAI, N. HE, Y. PAN, B. BOOTS, AND L. SONG, *Learning from conditional distributions via dual embeddings*, in Artificial Intelligence and Statistics, 2017, pp. 1458–1467.

[3] D. DAVIS AND D. DRUSVYATSKIY, *Stochastic model-based minimization of weakly convex functions*, SIAM J. Optim., 29 (2019), pp. 207–239.

[4] D. DAVIS AND B. GRIMMER, *Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems*, SIAM J. Optim., 29 (2019), pp. 1908–1930.

[5] D. DENTCHEVA, S. PENEV, AND A. RUSZCZYŃSKI, *Statistical estimation of composite risk functionals and risk optimization problems*, Ann. Inst. Statist. Math., 69 (2017), pp. 737–760.

[6] YU. M. ERMOLIEV, *Methods of Stochastic Programming*, Nauka, Moscow, 1976.

[7] YU. M. ERMOLIEV AND V. I. NORKIN, *Sample average approximation method for compound stochastic optimization problems*, SIAM J. Optim., 23 (2013), pp. 2231–2263.

[8] R. GE, F. HUANG, C. JIN, AND Y. YUAN, *Escaping from saddle points—Online stochastic gradient for tensor decomposition*, in Conference on Learning Theory, MLR Press, Cambridge, MA, 2015, pp. 797–842.

[9] S. Ghadimi and G. Lan, *Stochastic first- and zeroth-order methods for nonconvex stochastic programming*, SIAM J. Optim., 23 (2013), pp. 2341–2368.
[10] S. Ghadimi and G. Lan, *Accelerated gradient methods for nonconvex nonlinear and stochastic programming*, Math. Program., 156 (2016), pp. 59–99.
[11] S. Ghadimi, G. Lan, and H. Zhang, *Mini-batch stochastic approximation methods for constrained nonconvex stochastic programming*, Math. Program., 155 (2016), pp. 267–305.
[12] A. M. Gupal and L. G. Bazhenov, *Stochastic analog of the conjugate gradient method*, Cybernetics (Kiev), 8 (1972), pp. 138–140.
[13] A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson, *Extragradient method with variance reduction for stochastic variational inequalities*, SIAM J. Optim., 27 (2017), pp. 686–724.
[14] A. Juditsky, A. Nemirovski, and C. Tauvel, *Solving variational inequalities with stochastic mirror-prox algorithm*, Stoch. Syst., 1 (2011), pp. 17–58.
[15] A. P. Korostelev, *On multi-step procedures of stochastic optimization*, Avtomat. i Telemekh., (5) 1981, pp. 82–90.
[16] G. M. Korpelevich, *The extragradient method for finding saddle points and other problems*, Matecon, 12 (1976), pp. 747–756.
[17] J. Koshal, A. Nedic, and U. V. Shanbhag, *Regularized iterative stochastic approximation methods for stochastic variational inequality problems*, IEEE Trans. Automat. Control, 58 (2013), pp. 594–609.
[18] X. Lian, M. Wang, and J. Liu, *Finite-sum composition optimization via variance reduced gradient descent*, in Artificial Intelligence and Statistics, 2017, pp. 1159–1167.
[19] Q. Lin, M. Liu, H. Rafique, and T. Yang, *Solving Weakly-Convex-Weakly-Concave Saddle-Point Problems as Weakly-Monotone Variational Inequality*, preprint, https://arxiv.org/abs/1810.10207, 2018.
[20] L. Liu, J. Liu, and D. Tao, *Duality-free methods for stochastic composition optimization*, IEEE Trans. Neural Netw. Learn. Syst., 30 (2019), pp. 1205–1217.
[21] B. T. Polyak, *Comparison of the convergence rates for single-step and multi-step optimization algorithms in the presence of noise*, Eng. Cybernet., 15 (1977), pp. 6–10.
[22] A. Ruszczyński, *A method of feasible directions for solving nonsmooth stochastic programming problems*, in Stochastic Programming, F. Archetti, G. Di Pillo, and M. Lucertini, eds., Springer, Berlin, 1986, pp. 258–271.
[23] A. Ruszczyński, *A linearization method for nonsmooth stochastic programming problems*, Math. Oper. Res., 12 (1987), pp. 32–49.
[24] A. Ruszczyński, *Nonlinear Optimization*, Princeton University Press, Princeton, NJ, 2006.
[25] A. Ruszczyński and W. Syski, *Stochastic approximation method with gradient averaging for unconstrained problems*, IEEE Trans. Automat. Control, 28 (1983), pp. 1097–1105.
[26] Ya. Z. Tsypkin, *Fundamentals of the Theory of Learning Systems*, Nauka, Moscow, 1970.
[27] M. Wang, E. X. Fang, and B. Liu, *Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions*, Math. Program., 161 (2017), pp. 419–449.
[28] M. Wang, J. Liu, and E. X. Fang, *Accelerating stochastic composition optimization*, J. Mach. Learn. Res., 18 (2017), pp. 1–23.
[29] L. Xiao, *Dual averaging methods for regularized stochastic learning and online optimization*, J. Machine Learn. Res., 11 (2010), pp. 2543–2596.
[30] S. Yang, M. Wang, and E. X. Fang, *Multilevel stochastic gradient methods for nested compositon optimization*, SIAM J. Optim., 29 (2019), pp. 616–659.