Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

eg-VSSA: An Extragradient Variable Sample-size Stochastic Approximation Scheme: Error Analysis and Complexity Trade-offs

Afrooz Jalilzadeh
Uday V. Shanbhag

Department of Indust. and Manuf. Engg. Penn. State University,
310 Leonhard Building University Park, PA 16803, USA

ABSTRACT

Given a sampling budget M, stochastic approximation (SA) schemes for constrained stochastic convex programs generally utilize a single sample for each projection, requiring an effort of M projection operations, each of possibly significant complexity. We present an extragradient-based variable sample-size SA scheme (**eg-VSSA**) that uses N_k samples at step k where $\sum_k N_k \leq M$. We make the following contributions: (i) In strongly convex regimes, the expected error decays linearly in the number of projection steps; (ii) In convex settings, if the sample-size is increased at suitable rates and the steplength is optimally chosen, the error diminishes at $\mathscr{O}(1/K^{1-\delta_1})$ and $\mathscr{O}(1/\sqrt{M})$, requiring $\mathscr{O}(M^{1/(2-\delta_2)})$ steps, where K denotes the number of steps and $\delta_1, \delta_2 > 0$ can be made arbitrarily small. Preliminary numerics reveal that increasing sample-size schemes provide solutions of similar accuracy to SA schemes but with effort reduced by factors as high as 20.

1 INTRODUCTION

In this paper, we consider the solution of the following constrained convex stochastic optimization problem:

$$\min_{x \in X} \mathbb{E}[f(x, \xi(\omega))], \tag{Opt}$$

where $X \subseteq \mathbb{R}^n$, $\xi : \Omega \to \mathbb{R}^d$, $f : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ and $(\Omega, \mathscr{F}, \mathbb{P})$ denotes the associated probability space, and $\mathbb{E}[\bullet]$ denotes the expectation with respect to the probability measure $\mathbb{P}$. Stochastic approximation schemes (cf. Robbins and Monro (1951), Spall (2003)) have been useful in solving a large class of stochastic optimization problems. However, in constrained settings, traditional approaches utilize a single sample at every step, requiring as many projection steps as the sampling budget. In settings where the constraint set is not simple, computing estimators via such an approach proves to be computationally intensive.

We pursue an alternate approach building on prior work by Shanbhag and Blanchet (2015) on variable sample-size stochastic approximation schemes for strongly convex optimization. This avenue prescribed rules under which the sample-size can be increased in the presence of both constant and diminishing steplength sequences so as to recover linear convergence in terms of the number of projection steps. In this paper, extending the framework presented by Shanbhag and Blanchet (2015), we construct an **e**xtra**g**radient Variable Sample-size Stochastic Approximation scheme (**eg-VSSA**) scheme in which given a random $x_1 \in X$, N_k samples are utilized at the kth iterate in computing iterates y_{k+1} and x_{k+1} for $k \geq 1$:

$$\begin{aligned} y_{k+1} &:= \Pi_X(x_k - \gamma_k(\nabla_x f(x_k) + w'_{k,N_k})), \\ x_{k+1} &:= \Pi_X(x_k - \gamma_k(\nabla_x f(y_{k+1}) + w_{k,N_k})), \end{aligned} \tag{\textbf{eg-VSSA}}$$

where $\Pi_X(y)$ denotes the projection of y onto X, γ_k denotes the steplength, $\omega_{j,\bullet} = \{\omega_{j,1}, \ldots, \omega_{j,N_j}\}$, $w'_{N_k} \triangleq \sum_{j=1}^{N_k}(\nabla_x f(x_k, \omega'_{k,j}) - \nabla_x f(x_k))/N_k$, $w_{N_k} \triangleq \sum_{j=1}^{N_k}(\nabla_x f(y_{k+1}, \omega_{k,j}) - \nabla_x f(y_{k+1}))/N_k$, $\mathscr{F}'_k \triangleq \{x_1, \omega'_{1,\bullet}, \ldots, \omega'_{k,\bullet}\}$

and $\mathscr{F}_k \triangleq \{\omega_{1,\bullet},\ldots,\omega_{k,\bullet}\}$. The scheme terminates after K steps where K is the largest integer satisfying $\sum_{k=1}^{K} N_k \leq M$. Note that extragradient schemes require two projections per step but we consider it a single projection requiring twice the effort.

Research on SA schemes has considered exploring various extensions of the vanilla scheme from several standpoints: (i) *Choice of steplength sequences:* In this context, there has been an effort to develop optimal constant steplength schemes by Nemirovski et al. (2009) while self-tuned steplength rules that adapt to problem parameters have been presented by Yousefian, Nedić, and Shanbhag (2012); (ii) *Sample-size choices:* An error analysis for varying sample-size mini-batch schemes was provided by Ghadimi et al. (2016) when sampling budget was infinite. Techniques for increasing sample-sizes were examined by Friedlander and Schmidt (2012) and So and Zhou (2013) while Pasupathy et al. (2014) examined rates at which sample sizes should be raised to obtain rates of convergence in line with their deterministic counterparts. Recent work by Byrd et al. (2012) considers two-stage rules for determining sample sizes; (iii) *Extensions:* There have been a host of extensions of standard stochastic approximation schemes that have incorporated averaging (cf. Polyak and Juditsky (1992)), addressed nonsmoothness (cf. Yousefian, Nedić, and Shanbhag (2012)), variational inequality problems (cf. Juditsky et al. (2011)), and Nash gtames (cf. Koshal, Nedić, and Shanbhag (2013)). We extend extragradient methods, first presented by Korpolevich (1976, 1983), to a stochastic regime reliant on increasing or constant sample sizes. When $N_k = \infty$, we recover deterministic extragradient schemes while $N_k = 1$ leads to standard stochastic extragradient schemes (see Juditsky et al. (2011) and Yousefian, Nedić, and Shanbhag (2014)). Next, we outline the contributions in our paper.
(i) *Strongly convex stochastic optimization:* In Section 2, we propose precise update rules for sample sizes in the face of a finite budget M and derive finite-sample error bounds. Importantly, we show that the error decays geometrically (linearly) in terms of projection steps under these rules, akin to the result provided by Shanbhag and Blanchet (2015) for standard stochastic approximation.
(ii) *Convex stochastic optimization:* In section 3, we consider merely convex programs and derive error bounds for the expected sub-optimality of the averaged sequence in terms of M for constant and increasing sample sizes with either constant or diminishing steplength sequences. In particular, we observe that when $N_k = N_0 k^a$ and $\gamma_k = \gamma$ for all $k \geq 1$, $a \in [0,1)$, and γ is optimally chosen, then the expected sub-optimality diminishes at $\mathscr{O}(1/K^{(a+1)/2})$. In fact, when $a \to 1$, this rate tends to the canonical deterministic unaccelerated rate of convergence in terms of projection steps. Furthermore, the expected sub-optimality decays at the rate of $\mathscr{O}(\sqrt{N_0}/\sqrt{(1+a)M})$, (canonical rate in terms of sample-size) but requires approximately $((1+a)M)^{1/(1+a)}$ steps in contrast with M steps required by standard SA schemes. In Section 4, we draw further insights and discuss trade-offs between theoretical bounds on accuracy and computational complexity. Notably, naive naive batching schemes with $N_k = N$ lead to a degradation in the worst-case error by $\sqrt{N}$.
(iii) *Numerics:* Preliminary numerics discussed in Section 4 suggest promise; while standard schemes produce solutions with an empirical accuracy of 0.0058 in 1000 steps, increasing sample-size schemes produce solutions with accuracies of approximately 0.001 for both constant and diminishing steplengths in 54 steps. Notably, batching schemes display significantly poorer performance; with batch sizes of 10 and 100, such schemes produce solutions with error 0.12 and 0.011 in 100 and 10 steps, respectively.

2 STRONGLY CONVEX STOCHASTIC OPTIMIZATION

In this section, we show that in strongly convex regimes, we show that the mean-squared error diminishes at a geometric rate with the number of projection steps. Throughout, we employ the following assumptions.

Assumption 1

(a) The function $f(x)$ is continuously differentiable on an open set containing X, strongly convex with constant η, with a Lipschitz continuous gradient with constant L.

(b) For some $\nu, \nu' \in (0,\infty)$, $\mathbb{E}[w_k \mid \mathscr{F}_{k-1} \cup \mathscr{F}'_k] = 0$, $\mathbb{E}[w'_k \mid \mathscr{F}_{k-1} \cup \mathscr{F}'_{k-1}] = 0$, $\mathbb{E}[\|w_k\|^2 | \mathscr{F}_{k-1} \cup \mathscr{F}'_k] \leq \frac{\nu^2}{N_k}$, and $\mathbb{E}[\|w'_k\|^2 \mid \mathscr{F}_{k-1} \cup \mathscr{F}'_{k-1}] \leq \frac{(\nu')^2}{N_k}$.

(c) There exists positive scalar C such that for all $k \geq 1$, $\|x_k - x^*\| \leq C$.

In our analysis we exploit Lemma 1 for projection mappings.

Lemma 1 (Bertsekas 2003) Let $X \subseteq \mathbb{R}^n$ be a nonempty closed and convex set. Then the following hold: (a) $\|\Pi_X[u]-\Pi_X[v]\| \le \|u-v\|$ for all $u,v \in \mathbb{R}^n$; (b)$(\Pi_X[u]-u)^T(x-\Pi_X[u]) \ge 0$ for all $u \in \mathbb{R}^n$ and $x \in X$.

The following result is specialized from Lemma 3 in (Yousefian, Nedić, and Shanbhag 2014).

Lemma 2 Consider the **eg-VSSA** scheme. Then we have the following for any $y \in X$ and for all $k \ge 1$.

$$\begin{aligned}\|x_{k+1}-y\|^2 \le \|x_k-y\|^2 &- (1-3L^2\gamma_k^2)\|y_{k+1}-x_k\|^2 \\ &+2\gamma_k(\nabla_x f(y_{k+1})+w_{k,N_k})^T(y-y_{k+1})+3\gamma_k^2(\|w_{k,N_k}\|^2+\|w'_{k,N_k}\|^2).\end{aligned} \tag{1}$$

Proof. Choose $y \in X$ and an arbitrary index $k \ge 1$. Hence, we have the following:

$$\|x_{k+1}-y\|^2 = \|x_{k+1}-x_k+x_k-y\|^2 = \|x_{k+1}-x_k\|^2+\|x_k-y\|^2+2(x_{k+1}-x_k)^T(x_k-y).$$

By adding and subtracting x_{k+1}, we obtain the following:

$$\begin{aligned}\|x_{k+1}-y\|^2 &= \|x_{k+1}-x_k\|^2+\|x_k-y\|^2+2(x_{k+1}-x_k)^T(x_k-x_{k+1})+2(x_{k+1}-x_k)^T(x_{k+1}-y) \\ &= \|x_k-y\|^2-\|x_{k+1}-x_k\|^2+2\underbrace{(x_{k+1}-x_k)^T(x_{k+1}-y)}_{\text{Term (a)}}.\end{aligned} \tag{2}$$

Since $x_{k+1} = \Pi_X(x_k-\gamma_k(\nabla_x f(y_{k+1})+w_{k,N_k}))$, by Lemma 1(b), the following holds:

$$\begin{aligned}0 &\le (x_{k+1}-(x_k-\gamma_k(\nabla_x f(y_{k+1})+w_{k,N_k})))^T(y-x_{k+1}) \\ &= (x_{k+1}-x_k)^T(y-x_{k+1})+\gamma_k(\nabla_x f(y_{k+1})+w_{k,N_k})^T(y-x_{k+1}), \text{ for all } y \in X.\end{aligned} \tag{3}$$

This implies that term (a) can be written as $(x_{k+1}-x_k)^T(x_{k+1}-y) \le \gamma_k(\nabla_x f(y_{k+1})+w_{k,N_k})^T(y-x_{k+1})$. Therefore, (2) can be expressed as follows:

$$\begin{aligned}\|x_{k+1}-y\|^2 &\le \|x_k-y\|^2-\|x_{k+1}-x_k\|^2+2\gamma_k(\nabla_x f(y_{k+1})+w_{k,N_k})^T(y-x_{k+1}) \\ &= \|x_k-y\|^2-\|x_{k+1}-x_k+y_{k+1}-y_{k+1}\|^2+2\gamma_k(\nabla_x f(y_{k+1})+w_{k,N_k})^T(y-x_{k+1}) \\ &= \|x_k-y\|^2-\|x_{k+1}-y_{k+1}\|^2-\|y_{k+1}-x_k\|^2 \\ &\quad -2\underbrace{(x_{k+1}-y_{k+1})^T(y_{k+1}-x_k)}_{\text{Term (b)}}+2\gamma_k(\nabla_x f(y_{k+1})+w_{k,N_k})^T(y-x_{k+1}).\end{aligned} \tag{4}$$

Since $y_{k+1} = \Pi_X(x_k-\gamma_k(\nabla_x f(x_k)+w'_{k,N_k}))$, by invoking Lemma 1(b), Term (b) can be bounded as follows:

$$\begin{aligned}0 &\le (y_{k+1}-(x_k-\gamma_k(\nabla_x f(x_k)+w'_{k,N_k})))^T(x_{k+1}-y_{k+1}) \\ &= (y_{k+1}-x_k)^T(x_{k+1}-y_{k+1})+\gamma_k(\nabla_x f(x_k)+w'_{k,N_k})^T(x_{k+1}-y_{k+1}).\end{aligned}$$

This implies that term (b) can be bounded as $-(x_{k+1}-y_{k+1})^T(y_{k+1}-x_k) \le \gamma_k(\nabla_x f(x_k)+w'_{k,N_k})^T(x_{k+1}-y_{k+1})$, allowing for the rewriting of (4):

$$\begin{aligned}\|x_{k+1}-y\|^2 &\le \|x_k-y\|^2-\|x_{k+1}-y_{k+1}\|^2-\|y_{k+1}-x_k\|^2+2\gamma_k(\nabla_x f(x_k)+w'_{k,N_k})^T(x_{k+1}-y_{k+1}) \\ &\quad +2\gamma_k(\nabla_k f(y_{k+1})+w_{k,N_k})^T(y-x_{k+1}) \\ &= \|x_k-y\|^2-\|x_{k+1}-y_{k+1}\|^2-\|y_{k+1}-x_k\|^2+2\gamma_k(\nabla_x f(y_{k+1})+w_{k,N_k})^T(y-y_{k+1}) \\ &\quad +2\gamma_k((\nabla_x f(x_k)+w'_{k,N_k})-(\nabla_x f(y_{k+1}+w_{k,N_k}))^T(x_{k+1}-y_{k+1}),\end{aligned} \tag{5}$$

where the equality follows from adding and subtracting y_{k+1}. Then by employing $2a^Tb - \|a\|^2 \leq \|b\|^2$ for any $a,b \in \mathbb{R}^n$ for $a = -(x_{k+1}-y_{k+1})$ and $2a^Tb = 2\gamma_k((\nabla_x f(x_k)+w'_{k,N_k}) - (\nabla_x f(y_{k+1})+w_{k,N_k}))^T(x_{k+1}-y_{k+1})$, we obtain the following by using the triangle inequality:

$$\begin{aligned}\|x_{k+1}-y\|^2 &\leq \|x_k-y\|^2 - \|y_{k+1}-x_k\|^2 + 2\gamma_k(\nabla_x f(y_{k+1})+w_{k,N_k})^T(y-y_{k+1}) \\ &\quad + \gamma_k^2\|\nabla_x f(y_{k+1})+w_{k,N_k} - \nabla_x f(x_k)+w'_{k,N_k}\|^2 \\ &\leq \|x_k-y\|^2 - \|y_{k+1}-x_k\|^2 + 2\gamma_k(\nabla_x f(y_{k+1})+w_{k,N_k})^T(y-y_{k+1}) \\ &\quad + \gamma_k^2(\|\nabla_x f(y_{k+1}) - \nabla_x f(x_k)\| + \|w_{k,N_k}\| + \|w'_{k,N_k}\|)^2.\end{aligned} \tag{6}$$

By using the Lipschitz continuity of $\nabla f(x)$ and by recalling that for any $a_1,\ldots,a_m \in \mathbb{R}$ and any integer $m \geq 2$, we have $(a_1+\ldots+a_m)^2 \leq m(a_1^2+\ldots+a_m^2)$, the required result follows.

$$\begin{aligned}\|x_{k+1}-y\|^2 &\leq \|x_k-y\|^2 - \|y_{k+1}-x_k\|^2 + 2\gamma_k(\nabla_x f(y_{k+1})+w_{k,N_k})^T(y-y_{k+1}) \\ &\quad + \gamma_k^2\left(3L^2\|y_{k+1}-x_k\|^2 + 3\|w_{k,N_k}\|^2 + 3\|w'_{k,N_k}\|^2\right) \\ &= \|x_k-y\|^2 - (1-3L^2\gamma_k^2)\|y_{k+1}-x_k\|^2 + 2\gamma_k(\nabla_x f(y_{k+1})+w_{k,N_k})^T(y-y_{k+1}) \\ &\quad + 3\gamma_k^2(\|w_{k,N_k}\|^2 + \|w'_{k,N_k}\|^2).\end{aligned}$$

∎

Lemma 3 Consider the **eg-VSSA** scheme. Suppose that Assumption 1 holds. Then, for all $k \geq 1$, $\mathbb{E}\|x_{k+1}-x^*\|^2 \leq q_k\mathbb{E}\|x^*-x_k\|^2 + 3\gamma_k^2\left(\frac{\nu^2+(\nu')^2}{N_k}\right)$ where γ_k and q_k satisfy the following:

$$\gamma_k \leq \min\left\{\frac{1}{\eta}, \frac{-\eta+\sqrt{\eta^2+3L^2}}{3L^2}\right\} \text{ and } q_k \triangleq (1-\gamma_k\eta). \tag{7}$$

Proof. By the strong monotonicity of the gradient map $\nabla_x f(x)$ over X, we have:

$$\begin{aligned}\nabla_x f(y_{k+1})^T(x^*-y_{k+1}) &= (\nabla_x f(y_{k+1}) - \nabla_x f(x^*) + \nabla_x f(x^*))^T(x^*-y_{k+1}) \\ &\leq -\eta\|y_{k+1}-x^*\|^2 + \nabla f(x^*)^T(x^*-y_{k+1}).\end{aligned}$$

From the optimality of x^*, we have that $\nabla f(x^*)^T(x^*-y_{k+1}) \leq 0$, inequality (1) in Lemma 2 may be rewritten as follows, where y is chosen as x^*:

$$\begin{aligned}\|x_{k+1}-x^*\|^2 &\leq \|x_k-x^*\|^2 - (1-3\gamma_k^2L^2)\|y_{k+1}-x_k\|^2 - 2\gamma_k\eta\|y_{k+1}-x^*\|^2 \\ &\quad + 2\gamma_k w_{k,N_k}^T(x^*-y_{k+1}) + 3\gamma_k^2(\|w_{k,N_k}\|^2 + \|w_{k,N_k}\|^2).\end{aligned} \tag{8}$$

By recalling that $(a_1+\ldots+a_m)^2 \leq m(a_1^2+\ldots+a_m^2)$ for $m \geq 2$, $2(\|y_{k+1}-x^*\|^2 + \|x_k-y_{k+1}\|^2) \geq (\|y_{k+1}-x^*\| + \|x_k-y_{k+1}\|)^2$. By the triangle inequality, $(\|y_{k+1}-x^*\| + \|x_k-y_{k+1}\|)^2 \geq (\|x^*-x_k\|)^2$, implying that $\|y_{k+1}-x^*\|^2 \geq \frac{1}{2}\|x^*-x_k\|^2 - \|y_{k+1}-x_k\|^2$ and (8) reduces to

$$\begin{aligned}\|x_{k+1}-x^*\|^2 &\leq (1-\gamma_k\eta)\|x^*-x_k\|^2 + (2\gamma_k\eta - 1 + 3\gamma_k^2L^2)\|y_{k+1}-x_k\|^2 \\ &\quad + 2\gamma_k w_{k,N_k}^T(x^*-y_{k+1}) + 3\gamma_k^2(\|w_{k,N_k}\|^2 + \|w'_{k,N_k}\|^2).\end{aligned}$$

By invoking (7), taking conditional expectations and applying Assumption 1 (b) on the conditional first and second moments, we obtain

$$\mathbb{E}[\|x_{k+1}-x^*\|^2 | \mathscr{F}_{k-1} \cup \mathscr{F}'_k] \leq q_k\|x^*-x_k\|^2 + 3\gamma_k^2\left(\frac{\nu^2+(\nu')^2}{N_k}\right).$$

The result follows by taking unconditional expectations on both sides. ∎

Under assumptions of either constant or increasing sample size, we now show that the mean-squared error diminishes at a linear rate.

Theorem 1 (Linear convergence in K**)** Suppose Assumption 1 holds and consider the **eg-VSSA** scheme. (i) Suppose K is given and $N_k = N$ where N is defined as follows:

$$N := \begin{cases} \lceil \beta_K q^{-K} \rceil, & \text{where} \quad \beta_K \triangleq (\frac{M}{K} - 1) q^K, \quad \gamma_k \triangleq \gamma \\ \lceil \beta_K q_K^{-K} \rceil. & \text{where} \quad \beta_K \triangleq (\frac{M}{K} - 1) q_K^K, \quad \gamma_k \triangleq \frac{\theta}{k} \end{cases}$$

Then the following holds for all $k \leq \bar{K}$:

$$\mathbb{E}[\|x_{k+1} - x^*\|]^2 \leq \begin{cases} q^k \left(C + 3 \frac{\min\left\{\bar{K},(1-q)^{-1}\right\} \gamma^2 (\nu^2 + (\nu')^2)}{\beta_{\bar{K}}} \right), & \gamma \leq 1/\eta, q = (1 - \gamma\eta) \\ q_{\bar{K}}^k \left(C + 3 \frac{\pi^2 \theta^2 (\nu^2 + (\nu')^2)}{6\beta_{\bar{K}}} \right), & \gamma_k \leq \theta/k, \theta \leq 1/\eta, q_k = (1 - \eta\gamma_k). \end{cases}.$$

(ii) Suppose K is given and N_k is defined as follows:

$$N_k := \begin{cases} \lceil \beta_K q^{-k} \rceil & \text{if} \quad \beta_K \triangleq \frac{M-K}{\sum_{k=1}^K q^{-k}}, \quad \gamma_k := \gamma \\ \lceil \frac{\beta_K}{\prod_{k=1}^k q_j} \rceil & \text{if} \quad \beta_K \triangleq \frac{M-K}{\sum_{k=1}^K \frac{1}{\prod_{j=1}^k q_j}}, \quad \gamma_k := \frac{\theta}{k}. \end{cases}$$

Then the following holds for all $k \leq \bar{K}$:

$$\mathbb{E}[\|x_{k+1} - x^*\|]^2 \leq \begin{cases} q^k \left(D + 3 \frac{\gamma^2 (\nu^2 + (\nu')^2) \bar{K}}{\beta_{\bar{K}}} \right), & q = (1 - \gamma\eta), \gamma < 1/\eta. \\ q_{\bar{K}}^k \left(D + 3 \frac{\pi^2 \theta^2 (\nu^2 + (\nu')^2)}{6\beta_{\bar{K}}} \right), & \gamma_k = \theta/k, \theta < 1/\eta, q = (1 - \gamma_k \eta). \end{cases}$$

Remark 1 This result demonstrates that the expected error decays at a suitably defined **linear** rate in terms of projection steps, akin to gradient schemes for deterministic strongly convex optimization. We omit this proof, given its similarity to the result provided for standard SA by Shanbhag and Blanchet (2015).

3 CONVEX STOCHASTIC OPTIMIZATION

In this section we assume that the function $f(x)$ is continuously differentiable and convex but not strongly convex (which we refer to as Assumption 1(a′)). We proceed to derive error bounds in terms of number of projection steps and sample complexity when sample size is either constant or increasing with either constant or diminishing steplengths.

Lemma 4 Consider the **eg-VSSA** scheme and suppose Assumption 1(a′,b,c) hold. Assume that $\gamma_k \leq \frac{1}{\sqrt{3}L}$ and suppose $\bar{y}_K \triangleq \frac{\sum_{k=1}^K \gamma_k y_{k+1}}{\sum_{k=1}^K \gamma_k}$. Then for all $K \geq 1$, we have the following:

$$\mathbb{E}[(f(\bar{y}_K) - f(x^*))] \leq \frac{C^2}{2\sum_{k=1}^K \gamma_k} + 3 \frac{\sum_{k=1}^K \gamma_k^2 \left(\frac{\nu^2 + (\nu')^2}{N_k} \right)}{2\sum_{k=1}^K \gamma_k}. \tag{9}$$

Proof. Recall that we have the following for any $y \in X$:

$$\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - (1 - 3L^2\gamma_k^2)\|y_{k+1} - x_k\|^2 + 2\gamma_k (\nabla_x f(y_{k+1}) + w_{k,N_k})^T (y - y_{k+1})$$
$$+ 3\gamma_k^2 (\|w_{k,N_k}\|^2 + \|w'_{k,N_k}\|^2).$$

Consequently, this holds for an optimal solution x^*.

$$\begin{aligned}
\|x_{k+1}-x^*\|^2 &\leq \|x_k-x^*\|^2-(1-3L^2\gamma_k^2)\|y_{k+1}-x_k\|^2+2\gamma_k(\nabla_x f(y_{k+1})+w_{k,N_k})^T(x^*-y_{k+1}) \\
&\quad +3\gamma_k^2(\|w_{k,N_k}\|^2+\|w'_{k,N_k}\|^2) \\
&= \|x_k-x^*\|^2-(1-3L^2\gamma_k^2)\|y_{k+1}-x_k\|^2+2\gamma_k\nabla_x f(y_{k+1})^T(x^*-y_{k+1})+2\gamma_k w_{k,N_k}^T(x^*-y_{k+1}) \\
&\quad +3\gamma_k^2(\|w_{k,N_k}\|^2+\|w'_{k,N_k}\|^2).
\end{aligned}$$

By convexity, we have that

$$2\gamma_k\nabla_x f(y_{k+1})^T(x^*-y_{k+1}) \leq 2\gamma_k(f(x^*)-f(y_{k+1})),$$

implying that

$$\begin{aligned}
\|x_{k+1}-x^*\|^2 &\leq \|x_k-x^*\|^2-(1-3L^2\gamma_k^2)\|y_{k+1}-x_k\|^2+2\gamma_k(f(x^*)-f(y_{k+1}))+2\gamma_k w_{k,N_k}^T(x^*-y_{k+1}) \\
&\quad +3\gamma_k^2(\|w_{k,N_k}\|^2+\|w'_{k,N_k}\|^2).
\end{aligned}$$

But this implies that

$$\begin{aligned}
2\gamma_k(f(y_{k+1})-f(x^*)) &\leq \|x_k-x^*\|^2-\|x_{k+1}-x^*\|^2-(1-3L^2\gamma_k^2)\|y_{k+1}-x_k\|^2+2\gamma_k w_{k,N_k}^T(x^*-y_{k+1}) \\
&\quad +3\gamma_k^2(\|w_{k,N_k}\|^2+\|w'_{k,N_k}\|^2).
\end{aligned}$$

If $\gamma_k \leq 1/(\sqrt{3}L)$, we have that

$$2\gamma_k(f(y_{k+1})-f(x^*)) \leq \|x_k-x^*\|^2-\|x_{k+1}-x^*\|^2+2\gamma_k w_{k,N_k}^T(x^*-y_{k+1})+3\gamma_k^2(\|w_{k,N_k}\|^2+\|w'_{k,N_k}\|^2).$$

$$\begin{aligned}
\implies \sum_{k=1}^K 2\gamma_k(f(y_{k+1})-f(x^*)) &\leq \|x_1-x^*\|^2-\|x_{K+1}-x^*\|^2 \\
&\quad +\sum_{k=1}^K \left(2\gamma_k w_{k,N_k}^T(x^*-y_{k+1})+3\gamma_k^2(\|w_{k,N_k}\|^2+\|w'_{k,N_k}\|^2)\right),
\end{aligned}$$

where the second inequality follows from summing from $k=1$ to K. If $\bar{y}_K \triangleq \frac{\sum_{k=1}^K \gamma_k y_{k+1}}{\sum_{k=1}^K}$, by convexity, we have $\left(\sum_{k=1}^K \gamma_k\right)(f(\bar{y}_K)-f(x^*)) \leq \sum_{k=1}^K \gamma_k(f(y_{k+1})-f(x^*))$. It follows that

$$\begin{aligned}
\left(2\sum_{k=1}^K \gamma_k\right)(f(\bar{y}_K)-f(x^*)) &\leq \|x_1-x^*\|^2-\|x_{K+1}-x^*\|^2 \\
&\quad +\sum_{k=1}^K \left(2\gamma_k w_{k,N_k}^T(x^*-y_{k+1})+3\gamma_k^2(\|w_{k,N_k}\|^2+\|w'_{k,N_k}\|^2)\right) \\
\implies \quad \mathbb{E}[(f(\bar{y}_K)-f(x^*))] &\leq \frac{\mathbb{E}[\|x_1-x^*\|^2]-\mathbb{E}[\|x_{K+1}-x^*\|^2]}{2\sum_{k=1}^K \gamma_k}+\underbrace{\frac{\sum_{k=1}^K 2\gamma_k\mathbb{E}[w_{k,N_k}^T(x^*-y_{k+1})]}{2\sum_{k=1}^K \gamma_k}}_{=0} \\
&\quad +\frac{3\sum_{k=1}^K \gamma_k^2(\mathbb{E}[\|w_{k,N_k}\|^2+\|w'_{k,N_k}\|^2])}{2\sum_{k=1}^K \gamma_k} \\
&\leq \frac{\mathbb{E}[\|x_1-x^*\|^2]}{2\sum_{k=1}^K \gamma_k}+\frac{3\sum_{k=1}^K \gamma_k^2(\mathbb{E}[\|w_{k,N_k}\|^2+\|w'_{k,N_k}\|^2])}{2\sum_{k=1}^K \gamma_k} \\
&\leq \frac{C^2}{2\sum_{k=1}^K \gamma_k}+3\frac{\sum_{k=1}^K \gamma_k^2(\frac{\nu^2+(\nu')^2)}{N_k}}{2\sum_{k=1}^K \gamma_k}.
\end{aligned}$$

∎

We now refine this result through four corollaries, one for each specialization of our **eg-VSSA** scheme.

Corollary 2 (**Constant sample size and steplength**) Consider the **eg-VSSA** scheme. Suppose Assumption 1(a$'$,b,c) holds and for all $k \geq 1$, $N_k = N$, and $\gamma_k = \gamma \leq 1/\sqrt{3}L$. Then we have the following:

(i) If $N_k = 1$ and $\gamma_k = \gamma$, $\mathbb{E}\left[f(\bar{y}_K) - f(x^*)\right] \leq \frac{C^2L\sqrt{3}}{\sqrt{K}} = \frac{C^2L\sqrt{3}}{\sqrt{M}}$ for all $K \geq 1$.

(ii) If $N_k = N \geq \frac{\nu^2+(\nu')^2}{L^2C^2}$ and $\gamma_k = \gamma$, $\mathbb{E}\left[f(\bar{y}_K) - f(x^*)\right] \leq \frac{C^2L\sqrt{3}}{\sqrt{K}} \leq \frac{C^2L\sqrt{3N}}{\sqrt{M}}$ for all $K \geq 1$.

Proof. We omit the proof of (i) and only prove part (ii). By setting $N_k = N$ and $\gamma_k = \gamma$ in inequality (9) and assuming that β satisfies $\beta \geq \max(1, \frac{L^2C^2N}{(\nu^2+(\nu')^2)})$, we obtain:

$$\mathbb{E}\left[f(\bar{y}_K) - f(x^*)\right] \leq \frac{C^2}{2K\gamma} + \frac{3(\nu^2+(\nu')^2)\gamma}{2N} \leq \frac{C^2}{2K\gamma} + \frac{3\beta(\nu^2+(\nu')^2)\gamma}{2N}. \tag{10}$$

Minimizing the right hand side in γ, we obtain

$$\gamma^* = \frac{CN^{1/2}}{\sqrt{3\beta K(\nu^2+(\nu')^2)}}$$

where $\gamma^* \leq \frac{1}{\sqrt{3}L}$ by choice of β. Now by substituting γ^* in (10) and using the fact that $K = \lceil (M/N) \rceil \geq M/N$, we obtain the desired result. ∎

Next, we assume that N_k is an increasing sequence and the steplength is constant; specifically, let $N_k = N_0k^a$ and $\gamma_k = \gamma$ for all k.

Corollary 3 (**Increasing sample size and constant steplength**) Consider the **eg-VSSA** scheme. Suppose Assumption 1(a$'$,b,c) holds and for all k, $N_k = N_0k^a$ and $\gamma_k = \gamma \leq 1/\sqrt{3}L$, where $a \in [0,1)$. If $N_0 \geq \frac{\nu^2+(\nu')^2}{C^2L^2(1-a)}$, then we have the following for all $K \geq 1$:

$$\mathbb{E}\left[f(\bar{y}_K) - f(x^*)\right] \ \leq \frac{C^2L\sqrt{3}}{K^{(\frac{1+a}{2})}} \leq \frac{C^2L\sqrt{3N_0}}{\sqrt{M(1+a)}}. \tag{11}$$

Proof. By letting $N_k = N_0k^a$ and $\gamma_k = \gamma$ in (9) we obtain:

$$\mathbb{E}\left[f(\bar{y}_K) - f(x^*)\right] \leq \frac{C^2}{2K\gamma} + \frac{3\gamma(\nu^2+(\nu')^2)\sum_{k=1}^{K}k^{-a}}{2KN_0} \leq \frac{C^2}{2K\gamma} + \frac{3\beta\gamma(\nu^2+(\nu')^2)K^{-a}}{2N_0(1-a)}, \tag{12}$$

where the second inequality follows by noting that $\sum_{k=1}^{K}k^{-a} \leq \int_0^K k^{-a}dk$ and by choosing $\beta = \frac{L^2C^2N_0(1-a)}{(\nu^2+(\nu')^2)}$. Note that $\beta \geq 1$ based on our assumption on N_0. The minimal value of the right-hand side is at

$$\gamma^* = \sqrt{\frac{C^2N_0(1-a)}{3\beta K^{(1-a)}(\nu^2+(\nu')^2)}},$$

where $\gamma^* \leq 1/(\sqrt{3}L)$ by choice of β and N_0. Now by substituting γ^* and β into (12) we obtain the desired result in terms of K. Since K is the largest integer such that $\sum_{k=1}^{K}N_0k^a \leq M$ we conclude that $M \leq \sum_{k=1}^{K+1}N_0k^a$. By using the inequality $\sum_{k=1}^{K+1}N_0k^a \leq \int_{k=0}^{K+2}N_0x^adx$, we obtain that

$$K \geq \frac{(M(a+1))^{\frac{1}{a+1}}}{N_0^{\frac{1}{a+1}}}.$$

By substituting the aforementioned inequality in (12), we obtain the bound in terms of M. ∎

Next, we consider a setting where $\gamma_k = \gamma_0 k^{-b}$, $b \in [0,1/2)$ and the sample-size is either constant (Cor. 4) or increasing (Cor. 5).

Corollary 4 (**Constant sample size and diminishing steplength**) Consider the **eg-VSSA** scheme. Suppose Assumption 1(a$'$,b,c) holds and for $k \geq 1$, $N_k = N$, and $\gamma_k = \gamma_0 k^{-b}$ where $\gamma_0 \leq 1/\sqrt{3}L$ and $b \in [0,1/2)$. If $N \geq \frac{\nu^2+(\nu')^2}{C^2L^2(1-2b)}$, then the following holds for all $K \geq 1$:

$$\mathbb{E}\left[f\left(\bar{y}_K\right) - f(x^*)\right] \leq \frac{C^2 L\sqrt{3}(1-b)K^{\frac{1}{2}-b}}{((1+K)^{(1-b)}-1)} \leq \frac{C^2 L\sqrt{3}(1-b)}{\left((\frac{M}{N}+1)^{\frac{1}{2}} - (\frac{M}{N}+1)^{b-\frac{1}{2}}\right)}. \tag{13}$$

Proof. In inequality (9) let $N_k = N$ and $\gamma_k = \gamma_0 k^{-b}$. Consequently we have that

$$\begin{aligned}
\mathbb{E}\left[f\left(\bar{y}_K\right) - f(x^*)\right] &\leq \frac{C^2}{2\gamma_0 \sum_{k=1}^K k^{-b}} + 3\frac{(\nu^2+(\nu')^2)\gamma_0 \sum_{k=1}^K k^{-2b}}{2N\sum_{k=1}^K k^{-b}} \\
&\leq \frac{C^2(1-b)}{2\gamma_0((1+K)^{(1-b)}-1)} + \frac{3\beta\gamma_0(\nu^2+(\nu')^2)(1-b)K^{(1-2b)}}{2(1-2b)N((1+K)^{(1-b)}-1)},
\end{aligned} \tag{14}$$

where $\sum_{k=1}^K k^{-b} \geq \int_1^{K+1} x^{-b}dx$, $\sum_{k=1}^K k^{-2b} \leq \int_0^K x^{-2b}dx$ and $\beta = C^2L^2(1-2b)N/(\nu^2+(\nu')^2) \geq 1$ by choice of N. The optimal value of γ_0 follows by minimizing the right hand side in γ_0, leading to

$$\gamma_0^* = \sqrt{\frac{C^2(1-2b)N}{3\beta(\nu^2+(\nu')^2)K^{(1-2b)}}},$$

where $\gamma^* \leq \frac{1}{\sqrt{3}L}$ by choice of β and $N_k = N \geq \frac{\nu^2+(\nu')^2}{C^2L^2(1-2b)}$. By substituting γ_0^* and β, the right hand side of (14) can be optimized as follows:

$$\mathbb{E}\left[f\left(\bar{y}_K\right) - f(x^*)\right] \leq \frac{C^2 L\sqrt{3}(1-b)K^{\frac{1}{2}-b}}{((1+K)^{(1-b)}-1)}.$$

Then by noting that $K = \lceil M/N \rceil \geq M/N$, our result follows with a little algebra as seen next.

$$\frac{C^2 L\sqrt{3}(1-b)K^{\frac{1}{2}-b}}{((1+K)^{(1-b)}-1)} \leq \frac{C^2 L\sqrt{3}(1-b)(\frac{M}{N}+1)^{\frac{1}{2}-b}}{((\frac{M}{N}+1)^{1-b}-1))} = \frac{C^2 L\sqrt{3}(1-b)}{\left((\frac{M}{N}+1)^{\frac{1}{2}} - (\frac{M}{N}+1)^{b-\frac{1}{2}}\right)}.$$

∎

Corollary 5 (**Increasing sample size and diminishing steplength**) Consider the **eg-VSSA** scheme. Suppose Assumption 1(a$'$,b,c) holds and for $k \geq 1$, $N_k = N_0 k^a$ and $\gamma_k = \gamma_0 k^{-b} \leq 1/\sqrt{3}L$ where $a \in [0,1)$ and $b \in [0,1/2)$. Then if $2b+a < 1$ and $N_0 \geq \frac{\nu^2+(\nu')^2}{C^2L^2(1-2b-a)}$, then we have the following for $K \geq 1$:

$$\begin{aligned}
\mathbb{E}\left[f\left(\bar{y}_K\right) - f(x^*)\right] &\leq \frac{C^2 L(1-b)\sqrt{3}}{((K+1)^{(\frac{1}{2}+\frac{a}{2})} - (K+1)^{(b+\frac{a}{2}-\frac{1}{2})})} && (15) \\
&\leq \frac{C^2 L(1-b)\sqrt{3}}{\left((\frac{M(a+1)}{N_0})^{\frac{1}{a+1}} - 1\right)^{(\frac{1}{2}+\frac{a}{2})} - \left((\frac{M(a+1)}{N_0})^{\frac{1}{a+1}} + 1\right)^{(b+\frac{a}{2}-\frac{1}{2})}}. && (16)
\end{aligned}$$

Proof. First, we let $N_k = N_0 k^a$ and $\gamma_k = \gamma_0 k^{-b}$ in inequality (9) and obtain the following:

$$\mathbb{E}\left[f\left(\bar{y}_K\right) - f(x^*)\right] \leq \frac{C^2}{2\gamma_0 \sum_{k=1}^K k^{-b}} + \frac{3\gamma_0(\nu^2 + (\nu')^2)\sum_{k=1}^K k^{-2b-a}}{2N_0 \sum_{k=1}^K k^{-b}}.$$

We know that $\sum_{k=1}^K k^{-b} \geq \int_1^{K+1} x^{-b}dx$ and $\sum_{k=1}^K k^{(-2b-a)} \leq \int_0^K x^{(-2b-a)}dx$. Therefore,

$$\mathbb{E}\left[f\left(\bar{y}_K\right) - f(x^*)\right] \leq \frac{C^2(1-b)}{2\gamma_0((K+1)^{(1-b)}-1)} + \frac{3\beta\gamma_0(\nu^2 + (\nu')^2)(1-b)(K+1)^{(1-2b-a)}}{2N_0((K+1)^{(1-b)}-1)(1-2b-a)}. \tag{17}$$

By optimizing the right hand side over γ_0, we obtain $\gamma_0^* = \sqrt{\frac{1}{3L^2(K+1)^{(1-2b-a)}}} \leq \frac{1}{\sqrt{3}L}$ by choosing $\beta = \frac{C^2L^2N_0(1-2b-a)}{\nu^2+(\nu')^2}$. By substituting γ_0^* in (17), we obtain the following:

$$\mathbb{E}\left[f\left(\bar{y}_K\right) - f(x^*)\right] \leq \frac{C^2L(1-b)\sqrt{3(K+1)^{(1-2b-a)}}}{((K+1)^{(1-b)}-1)} = \frac{C^2L(1-b)\sqrt{3}}{((K+1)^{(\frac{1}{2}+\frac{a}{2})} - (K+1)^{(b+\frac{a}{2}-\frac{1}{2})})}. \tag{18}$$

To obtain an error bound in terms of M, we derive lower and upper bounds for K:

$$\begin{aligned} &\int_{k=0}^{K} N_0 k^a dk \leq \sum_{k=1}^{K} N_0 k^a \leq M \leq \sum_{k=1}^{K+1} N_0 k^a \leq \int_{k=0}^{K+2} N_0 k^a dk \\ \implies \quad & \frac{(M(a+1))^{\frac{1}{a+1}}}{N_0^{\frac{1}{a+1}}} - 2 \leq K \leq \frac{(M(a+1))^{\frac{1}{a+1}}}{N_0^{\frac{1}{a+1}}}. \end{aligned} \tag{19}$$

Now by substituting (19) in (18), the required bound in terms of M can be seen to be (16). ∎

4 INSIGHTS AND TRADE-OFFS

In this section, we provide some insights regarding the rate statements in the previous section and conclude with a trade-off analysis between computational and sample complexity.

(i) Constant sample size and steplength: As captured by Corollary 2, we observe that the theoretical bound deteriorates by $\sqrt{N}$ in comparison with case for standard stochastic approximation with $N = 1$ (See Nemirovski et al. (2009)) but provides a solution in $\lceil (M/N) \rceil$ projection steps.

(ii) Increasing sample size and constant steplength: As seen in Corollary 3, while the rate in terms of sample-complexity stays the same, if $N_0 \leq (1+a)$, the theoretical bound actually **improves** by a constant factor of $\sqrt{N_0/(1+a)}$ in comparison with the standard case with $N = 1$ and an optimal steplength where the sample-size sequence is $N_k = N_0 k^a$ and $N_0/(1+a) \leq 1$. Specifically if $N_0 = 1, a = 1-\varepsilon$, and $(\nu^2 + (\nu')^2)/(L^2C^2\varepsilon) \leq 1$, the improvement will be $\sqrt{1/(2-\varepsilon)}$ but requires no more than $\mathscr{O}(1/K^{1+\varepsilon/2})$ projection steps. In fact, we see that the rate tends to the optimal non-accelerated rate for standard gradient methods in terms of projection steps as $\varepsilon \to 0$. It is also worth emphasizing that standard stochastic approximation schemes utilize M projection steps while the proposed scheme requires only $((2-\varepsilon)M)^{1/(2-\varepsilon)}$ projection steps.

(iii) Constant sample size and diminishing steplength: This scheme is captured by Corollary 4 and when $\gamma_k = \gamma_0 k^{-b}$ if $b = \frac{1}{2}$ and $N_k = N$, the expected sub-optimality can be bounded as follows:

$$\frac{C^2L\sqrt{3}(1-b)}{\left((\frac{M}{N}+1)^{\frac{1}{2}} - (\frac{M}{N}+1)^{b-\frac{1}{2}}\right)} = \frac{C^2L\sqrt{3}}{2\left((\frac{M}{N}+1)^{\frac{1}{2}} - 1\right)} \leq \frac{C^2L\sqrt{3}\sqrt{N}}{2\left(M^{\frac{1}{2}} - \sqrt{N}\right)}.$$

In effect, similar to (i), we obtain an approximate degradation of $\sqrt{N}$ in the bound but the rate appears to be close to the canonical rate.

(iv) Increasing sample size and diminishing steplength: Finally, by Corollary 5, when $N_k = N_0 k^a$ and $\gamma_k = \gamma_0 k^{-b}$, if $2b + a < 1$, $a = 1$ and $b = \varepsilon$, then one attains close to the optimal rate:

$$\frac{C^2 L(1-\varepsilon)\sqrt{3}}{(\sqrt{2M}-1)-(\sqrt{2M}+1)^{\varepsilon}},$$

but at a computational cost of $\sqrt{M}$ steps rather than M steps for standard SA schemes.

Trade-off analysis between sample and computational complexity: Traditional implementations of stochastic approximation with $N_k = 1$ lead to a worst-case error of $\mathcal{O}(M^{-1/2})$ and require M projection steps. Consider a setting where $\gamma_k = \gamma$ and $N_k = N_0 k^a$:

$\mathbf{N_0 = 1, a \in [0, (1-\varepsilon)]}$: Based on Corollary 3, when $N_0 = 1$, we note that the empirical error decays at $\mathcal{O}(M^{-1/2})$ when a varies from 0 to $1-\varepsilon$ while the computational effort changes from M to $((2-\varepsilon)M)^{1/(2-\varepsilon)}$. In effect, an optimal empirical error is guaranteed with far less computational effort.

$\mathbf{N_0 \in \{1, \ldots, \lfloor M^{1-\delta} \rfloor\}}$**, a constant:** From Table 1, we see a degradation in worst-case error when N_0 is raised for constant a. For instance, when $N_0 = 1$, we recover the canical result for stochastic approximation, leading to accuracy of $M^{-0.5}$ with a computational complexity of M projection steps. If N_0 is increased to $\sqrt{M}$ and $a = 1-\varepsilon$, the resulting computational complexity reduces to $(\sqrt{M})^{1/(2-\varepsilon)}(2-\varepsilon)^{1/(2-\varepsilon)}$ (approx. $M^{1/4}$) with an error of $M^{-1/4}(2-\varepsilon)^{-0.5}$. For instance, if $M =$ `1e6` and $\varepsilon = 0.01$, $N_0 = \sqrt{M}$ reduces the computational complexity from 1e6 to approximately 46 projection steps while the accuracy worsens from `1e-3` to `1.4e-(3/2)`.

Table 1: Constant sample complexity with $N_k = N_0 k^a$.

N_0	Samp-complex	Comp-complex	ε
1	M	M	$M^{-0.5}$
$M^{0.125}$	M	$(M^{0.875})^{\frac{1}{a+1}}(a+1)^{\frac{1}{a+1}}$	$M^{-0.4375}(1+a)^{-0.5}$
$M^{0.25}$	M	$(M^{0.75})^{\frac{1}{a+1}}(a+1)^{\frac{1}{a+1}}$	$M^{-0.375}(1+a)^{-0.5}$
$M^{0.5}$	M	$(M^{0.5})^{\frac{1}{a+1}}(a+1)^{\frac{1}{a+1}}$	$M^{-0.25}(1+a)^{-0.5}$
$M^{1-\delta}$	M	$(M^{\delta})^{\frac{1}{a+1}}(a+1)^{\frac{1}{a+1}}$	$M^{-\delta/2}(1+a)^{-0.5}$

5 NUMERICAL EXAMPLES

Next, we apply the proposed schemes on an example where empirical and theoretical error as well as CPU time are compared for different choices of a and b. Consider a stochastic network utility problem with a network with N users, having r resources with finite capacity where c_j is capacity of j^{th} resource and matrix $A \in \{0,1\}^{r \times N}$ is the network adjacent matrix. We assume that each user can utilize a subset of resources and the resulting optimization problem is as follows: $\min_{Ax \le c} \mathbb{E}\left[-\sum_{i=1}^{N} k_i(\xi_i)\log(1+x_i) + \|Ax\|^2\right]$ where $k_i(\xi_i)$ is an uncertain parameter. Here, $c = (0.1, 0.15, 0.2, 0.1, 0.15, 0.2, 0.2, 0.15, 0.25), x \ge 0, n = 5, k_i(\xi_i)$ has an uniform distribution $U(0.2, 1)$ for all i, and $M = 1000$.

Table 2: Constant sample size and steplengths.

N	$\|f(x) - f(x^*)\|$	# iteration	Theor. bound	CPU Time
100	1.230e-01	11	4.490e+01	2.611e-01
10	1.149e-02	101	1.420e+01	2.613e+00
1	5.785e-03	1000	4.490e+00	2.592e+01

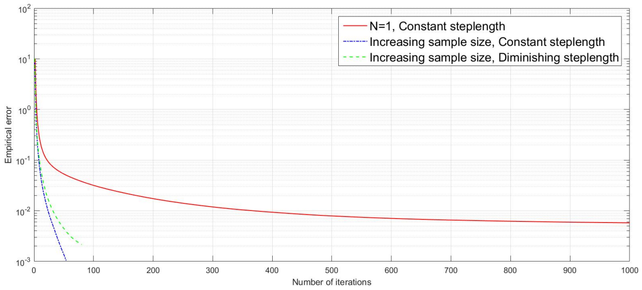

Figure 1: Comparison across schemes: Number of iterations vs empirical error.

Constant sample-size schemes with constant steplength: Table 2 compares **eg-VSSA** schemes when $N = 1$ with settings of $N = 10$ and 100 with optimally chosen steplengths. Note that when $N = 1$, the scheme requires M projection steps. It can be seen that schemes with $N = 1$ take about 100 times more than schemes with $N = 100$ while $N = 1$ provides an expected sub-optimality error of about 0.0058 in comparison with about 0.123 for $N = 100$. In short, naive batching schemes perform poorly in terms of empirical error, corresponding well with the degradation suggested by theory.

Increasing sample-size schemes: In Table 3, with $\gamma_k = \gamma$ (optimally chosen) and $N_k = k^a$, we notice that the scheme performs well for diverse choices of a compared with standard SA schemes. As seen in Table 2, when $N_k = 1$, we obtain an empirical accuracy of 0.0058 and require 1000 projection steps. In comparison, if $a = 0.5$, the empirical accuracy improves to 0.00336 (improvement by 42%) and requires approximately 132 projection steps (7.7 times less effort). When $a = 0.9$, we observe that empirical accuracy drops to 0.00105 (an improvement of 82%) and requires 54 steps (an improvement by a factor of 19) as seen in Table 3. Similarly, as seen in Table 4, when γ_k is a diminishing sequence, we notice for the same level of a, we see slight improvements in the empirical accuracy. Figure 1 provides a graphical comparison of how the three implementations compare in terms of trajectories.

Table 3: $N_k = N_0 k^a, \gamma_k = \gamma$.

a	b	$\|f(x) - f(x^*)\|$	# iteration	Theor. bound	CPU Time
1.0e-01	0	5.474e-03	529	2.473e+00	1.357e+01
2.0e-01	0	5.036e-03	369	2.369e+00	9.435e+00
3.0e-01	0	4.520e-03	249	2.278e+00	6.381e+00
4.0e-01	0	4.046e-03	177	2.198e+00	4.534e+00
5.0e-01	0	3.360e-03	132	2.128e+00	3.366e+00
6.0e-01	0	2.834e-03	102	2.065e+00	2.587e+00
7.0e-01	0	2.319e-03	80	2.009e+00	2.028e+00
8.0e-01	0	1.774e-03	65	1.958e+00	1.641e+00
9.0e-01	0	1.046e-03	54	1.914e+00	1.362e+00

Table 4: $N_k = N_0 k^a, \gamma_k = \gamma_0 k^{-b}$.

a	b	$\|f(x) - f(x^*)\|$	# iteration	Theor. bound	CPU Time
1.0e-01	4.000e-01	4.473e-03	529	2.624e+00	1.404e+01
2.0e-01	3.500e-01	4.203e-03	369	2.721e+00	9.481e+00
3.0e-01	3.000e-01	3.841e-03	249	2.814e+00	6.388e+00
4.0e-01	2.500e-01	3.515e-03	177	2.903e+00	4.514e+00
5.0e-01	2.000e-01	3.021e-03	132	2.991e+00	3.362e+00
6.0e-01	1.500e-01	2.577e-03	102	3.075e+00	2.588e+00
7.0e-01	1.000e-01	2.152e-03	80	3.157e+00	2.020e+00
8.0e-01	5.000e-02	1.682e-03	65	3.237e+00	1.633e+00
9.0e-01	0.000e+00	1.046e-03	54	3.315e+00	1.347e+00

6 CONCLUDING REMARKS

In this paper, we present an extragradient variable sample-size stochastic approximation scheme (**eg-VSSA**) and make the following contributions. First, the scheme admits similar linear rates akin to the scheme presented by Shanbhag and Blanchet (2015) for strongly convex regimes. Second, for convex programs, we derive error bounds for the expected sub-optimality in terms of sampling budget M for constant and increasing sample sizes with either constant or diminishing steplength sequences. Specifically, if $N_k = N_0 k^a$ and $\gamma_k = \gamma$ for all $k \geq 1$ and $a \in [0,1)$ and γ is optimally chosen, the expected sub-optimality displays a rate of decay of $\mathcal{O}(1/K^{(a+1)/2})$ where K denotes the number of steps while this rate is $\mathcal{O}(\sqrt{N_0}/\sqrt{(2+a)M})$ in terms of M. In effect, these schemes display the canonical rate but require approximately $(M(a+1)/N_0)^{1/(a+1)}$ steps rather than M steps. Additionally, naive naive batching schemes with $N_k = N$ lead to a degradation in the worst-case error by $\sqrt{N}$. Preliminary numerics suggest that such avenues hold much promise and provide solutions of comparable accuracy with a fraction of the effort.

REFERENCES

Bertsekas, D. P. 2003. *Convex analysis and optimization*. Athena Scientific, Belmont, MA. With Angelia Nedić and Asuman E. Ozdaglar.

Byrd, R. H., G. M. Chin, J. Nocedal, and Y. Wu. 2012. "Sample size selection in optimization methods for machine learning". *Math. Program.* 134 (1): 127–155.

Friedlander, M. P., and M. Schmidt. 2012. "Hybrid deterministic-stochastic methods for data fitting". *SIAM J. Scientific Computing* 34 (3): 1380–1405.

Ghadimi, S., G. Lan, and H. Zhang. 2016. "Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization". *Math. Program.* 155 (1-2, Ser. A): 267–305.

Juditsky, A., A. Nemirovski, and C. Tauvel. 2011. "Solving variational inequalities with stochastic mirror-prox algorithm". *Stochastic Systems* 1 (1): 17–58.

Korpelevich 1976. "The extragradient method for finding saddle points and other problems". *Ekonomika i Matematicheskie Metody* 12:747–756.

Korpelevich, G. M. 1983. "Extrapolation gradient methods and their relation to modified Lagrange functions". *Èkonom. i Mat. Metody* 19 (4): 694–703.

Koshal, J., A. Nedić, and U. V. Shanbhag. 2013. "Regularized iterative stochastic approximation methods for variational inequality problems". *IEEE Transactions on Automatic Control* 58(3):594–609.

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro. 2009. "Robust stochastic approximation approach to stochastic programming". *SIAM Journal on Optimization* 19 (4): 1574–1609.

Pasupathy, R., P. Glynn, S. Ghosh, and F. Hashemi. 2014. "How much to sample in simulation-based stochastic recursions?". *Under review at SIAM Journal of Optimization*.

Polyak, B., and A. Juditsky. 1992. "Acceleration of stochastic approximation by averaging". *SIAM J. Control Optim.* 30 (4): 838–855.

Robbins, H., and S. Monro. 1951. "A stochastic approximation method". *Ann. Math. Statistics* 22:400–407.

Shanbhag, U. V., and J. H. Blanchet. 2015. "Budget-constrained stochastic approximation". In *Proceedings of the 2015 Winter Simulation Conference, Huntington Beach, CA, USA, December 6-9, 2015,* edited by L. Yilmaz, W. K V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 368–379. Piscataway, New Jersey, USA: Institute of Electrical and Electronics Engineers, Inc.

So, A. M., and Z. Zhou. 2013. "Non-asymptotic convergence analysis of inexact gradient methods for machine learning without strong convexity". *CoRR,abs* (1309.0113,).

Spall, J. C. 2003. *Introduction to stochastic search and optimization*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Hoboken, NJ: Wiley-Interscience, John Wiley & Sons. Estimation, simulation, and control.

Yousefian, F., A. Nedić, and U. V. Shanbhag. 2012. "On stochastic gradient and subgradient methods with adaptive steplength sequences". *Automatica* 48 (1): 56–67. An extended version of the paper available at: http://arxiv.org/abs/1105.4549.

Yousefian, F., A. Nedić, and U. V. Shanbhag. 2014, Dec. "Optimal robust smoothing extragradient algorithms for stochastic variational inequality problems". In *53rd IEEE Conference on Decision and Control*, 5831–5836.

AUTHOR BIOGRAPHIES

AFROOZ JALILZADEH is a Ph.D. student in the Harold and Inge Marcus Department of Industrial and Manufacturing Engineering at Pennsylvania State University. She received her B.S. in Mathematics and Applications from University of Tehran in 2015. Her interests include stochastic optimization, variational inequality problems, convex optimization, and machine learning. Her email address is azj5286@psu.edu.

UDAY V. SHANBHAG is a professor in the Harold and Inge Marcus Department of Industrial and Manufacturing Engineering at Pennsylvania State University and can be reached at udaybag@psu.edu.