Budget-Constrained Stochastic Approximation

Uday V. Shanbhag
Department of Industrial and Manufacturing Engineering
Pennsylvania State University
310 Leonhard Building
University Park, PA 16802, USA

José Blanchet
Department of Industrial Engineering and Operations Research and Statistics,
Columbia University,
340 W. Mudd Building, 500 West 120th Street,
New York, NY 10027-6699, USA

Abstract

Traditional stochastic approximation (SA) schemes for stochastic optimization employ a single gradient or a fixed batch of noisy gradients in computing a new iterate. We consider the development of SA schemes in which \( N_k \) gradient samples are utilized at step \( k \) and the total computational budget is \( M \), so that \( \sum_{k=1}^{K} N_k \leq M \) where \( K \) denotes the terminal step. This paper makes the following contributions: (i) We conduct an error analysis for constant batches (\( N_k = N \)) both for constant and diminishing steplength regime and show linear convergence in terms of expected optimal value; (ii) we extend the two schemes in (i), and the corresponding linear convergence rates, now in the setting of increasing sample sizes (increasing \( N_k \)), assuming constant or diminishing steplength; (iii) finally, when steplengths are constant, we obtain the optimal number of projection steps that minimize the bound on the mean-squared error.

1 Introduction

First suggested by Robbins and Monro (1951) in the context of root finding problems, stochastic approximation schemes (Borkar 2008) have proven to be enormously effective in a broad class of stochastic computational problems including convex optimization, variational inequality problems, and Markov decision processes. We provide a brief description of the method in the context of stochastic convex optimization:

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

where \( X \subseteq \mathbb{R}^n, \xi : \Omega \rightarrow \mathbb{R}^d, f : \mathbb{R}^n \times \mathbb{R}^d \rightarrow \mathbb{R} \) and \((\Omega, \mathcal{F}, \mathbb{P})\) denotes the associated probability space and \( \mathbb{E}[\bullet] \) denotes the expectation with respect to \( \mathbb{P}[\bullet] \). Our interest lies in problems where \( X \) is a convex set and \( f(x) \) is a convex function in \( x \) where \( f(x) := \mathbb{E}[f(x, \xi(\omega))] \). Vanilla implementations of stochastic approximation schemes have comprised of the following update rule: Given an \( x_0 \in X \), an SA scheme is based on the following update rule:

\[
x_{k+1} := \Pi_X (x_k - \gamma_k (\nabla_x f(x_k) + w_k)), \quad k \geq 0
\]

(SA)

where \( w_k := \nabla_x f(x_k; \omega_k) - \nabla_x f(x_k) \) and \( \nabla_x f(x, \xi(\omega)) \) is referred to as \( \nabla_x f(x, \omega) \). If \( \{\gamma_k\} \) is a square-summable but non-summable sequence, then \( \{x_k\} \rightarrow x^* \) where \( x^* \in \text{SOL}(X, \nabla f) \). Furthermore, under
strong convexity, it was shown that when $x \in \text{int}(X)$, $\mathbb{E}[f(x_k; \omega_k) - f^*] = O(1/k)$. When both strong convexity and differentiability of the function are weakened, this convergence rate diminishes to $O(1/\sqrt{k})$ which has been shown to be unimprovable by Nemirovski and Yudin (1983). Over the last few years, there has been a renewed effort to re-examine stochastic approximation schemes:

**Steplength choices:** The choice of the steplengths can prove devastating from the standpoint of algorithm performance, motivating diverse efforts to design schemes where steplengths are chosen in a sensible fashion. Nemirovski et al. (2009) presented a constant steplength scheme where the choice is contingent on the termination length and achieves the optimal rate. While these choices require an a priori specification of the termination length, such schemes naturally provide approximate solutions at best. Asymptotically exact schemes have been developed by Yousefian et al. (2012) in which the steplength sequence is selected in accordance with problem parameters (such as Lipschitz constant, convexity constant etc.) and are seen to display the optimal rate. Adaptively chosen steplengths have also been considered with a study of the associated rates (See (Ciccek, Broadie, and Zeevi 2011) and references therein).

**Constant sample size SA schemes:** Constant sample size SA schemes are generally referred to as mini-batch SA schemes and there has been significant analysis of the associated error bounds (cf. Ghadimi, Lan, and Zhang 2014)). However, much of this work has assumed that the size of the batch is taken as a parameter. In contrast, the present work additionally focuses on determining the optimal batch size so as to recover fast rates of convergence while accommodating a budget constraint on the number of samples.

**Variable sample size SA schemes:** Amongst related prior work is that by (Friedlander and Schmidt 2012) in which sample-averages of gradients are utilized within the gradient method. Akin to the question investigated here, Friedlander and Schmidt investigate how rates of deterministic methods can be achieved. Our work differs in that we assume that the sampling budget is constrained and precise schemes for updating the sampling size are provided so as to recover linear rates of convergence.

**Research question:** We consider a generalization where at step $k$, $N_k$ samples of the gradient are obtained. As a consequence, given a randomly generated $x_0 \in X$, the sequence $\{x_k\}$ is given by the following update rule:

$$ x_{k+1} := \Pi_X\left( x_k - \gamma_k \frac{1}{N_k} \sum_{j=1}^{N_k} \nabla_x f(x_k; \xi_{j,k}) \right), \quad k \geq 0 \quad (SA_k) $$

One may immediately note that when $N_k := 1$ for all $k$, this reduces to the standard SA scheme. In the context of this scheme, we consider the following question. **Given a simulation budget of $M$ samples, how should $N_k$ be selected as a function of $M$ and problem parameters so as to attain the rate of convergence seen in deterministic regimes?** To this end, this paper makes the following contributions:

(i) We consider a constant sample-size (batch) SA scheme under constant and diminishing steplength regimes. Specifically, given a budget on computational complexity, we prove that both schemes display linear convergence in a mean-squared sense;

(ii) Next, we extend the two prior SA schemes to allow for increasing sample sizes and provide a scheme for updating the sample size that allows for recovering the linear rate of convergence under a computational budget constraint when steplength sequences are either constant or diminishing;

(iii) Finally, when steplengths are constant, we resolve the question of the optimal number of projection steps that minimize the bound on the mean-squared error.
Furthermore, let $F$ implying that the no more than $\mu$ convex function with convexity constant $\mu$.

Lemma 1 defined. We begin with a simple bound on the conditional second moment of the error. In this section, we consider the setting in which a fixed sample size is utilized at every step. Our original research question then reduces to what the size of this sample should be given that a fixed simulation budget is available. In fact, based on this sample size, the overall number of computational steps can be defined. We begin with a simple bound on the conditional second moment of the error.

Lemma 1 Consider the variable sample SA scheme denoted by $(SA_N)$ and suppose $f(x)$ is a strongly convex function with convexity constant $\mu$. Furthermore suppose $f$ is continuously differentiable in $x$ with Lipschitz continuous gradients with constant $\mu$.

Then for every nonnegative $k$, the following holds

$$E[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \leq (1 - 2\eta \mu + \eta^2 L^2)\|x_k - x^*\|^2 + \frac{\gamma^2 \nu^2}{N_k}.$$

Proof. We begin by noting that $x_k$ can be expressed as follows:

$$x_{k+1} = \Pi_X \left( x_k - \gamma k \sum_{j=1}^{N_k} \nabla f(x_k, w_{k,j}) \right) = \Pi_X \left( x_k - \gamma k \frac{\sum_{j=1}^{N_k} \nabla f(x_k, w_{k,j})}{N_k} \right)$$

$$= \Pi_X \left( x_k - \gamma k \left( \nabla f(x_k) + \frac{\sum_{j=1}^{N_k} w_{k,j}}{N_k} \right) \right) = \Pi_X \left( x_k - \gamma k (\nabla f(x_k) + \tilde{w}_{k,N_k}) \right),$$

where $\tilde{w}_{k,N_k} := \frac{\sum_{j=1}^{N_k} \nabla f(x_k, w_{k,j})}{N_k}$. By leveraging the non-expansivity of the Euclidean projector, we may express $\|x_{k+1} - x^*\|^2$ as follows:

$$\|x_{k+1} - x^*\|^2 = \|\Pi_X (x_k - \gamma k (\nabla f(x_k) + \tilde{w}_{k,N_k}) - \Pi_X (x^* - \gamma k \nabla f(x^*))\|^2$$

$$\leq \|(x_k - \gamma k (\nabla f(x_k) + \tilde{w}_{k,N_k})) - (x^* - \gamma k \nabla f(x^*))\|^2$$

$$= (1 - 2\eta \gamma k + \eta^2 L^2)\|(x_k - x^*)\|^2 + \gamma^2 \nu^2 \|\tilde{w}_{k,N_k}\|^2.$$

Taking conditional expectations on $\mathcal{F}_k$, we obtain the following inequality:

$$E[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \leq (1 - 2\eta \gamma k + \eta^2 L^2)\|x_k - x^*\|^2 + \gamma^2 \nu^2 E[\|\tilde{w}_{k,N_k}\|^2 \mid \mathcal{F}_k],$$

where $E[\tilde{w}_{k,N_k} \mid \mathcal{F}_k] = 0$. Furthermore, since we have that $E[\|w_k\|^2 \mid \mathcal{F}_k] \leq \nu^2$, it follows that $E[\|w_{k,N_k}\|^2 \mid \mathcal{F}_k] \leq \frac{\nu^2}{N_k}$. The result follows.

Our first result assumes that a fixed number of samples, namely $N$, are employed at each iteration implying that the no more than $K \triangleq \left\lceil \frac{M}{N} \right\rceil$ steps are taken. We now analyze the error bounds at the $K$th iteration under both constant and diminishing stepleength regimes.

Proposition 2 Consider the scheme (VSSA) and the associated mean-squared error after $K$ steps where $N_k := N$ for all $k$. Then the following hold:
Taking unconditional expectations, we obtain the following:

\[ \mathbb{E}[\|x_{k+1} - x^*\|^2] \leq \left( D + \frac{\gamma^2 \nu^2}{\beta_k} \left( \frac{1}{1-q} \right) \right) q^k, \quad \forall k \geq 0. \]

(i) Suppose \( \gamma_k := \gamma \) for all \( k \) and \( q \triangleq (1 - 2\eta \gamma + \gamma^2 L^2) \). Then the mean-squared error may be bounded as follows:

\[ \mathbb{E}[\|x_{k+1} - x^*\|^2] \leq \left( D + \frac{\gamma^2 \nu^2}{\beta_k} \left( \frac{1}{1-q} \right) \right) q^k, \quad \forall k \geq 0. \]

Proof. We begin by considering the general case in which the steplength is denoted by \( \gamma_k \). Then after \( K \) gradient steps, the mean-squared error is given by the following:

\[ \mathbb{E}[\|x_{K+1} - x^*\|^2 \mid \mathcal{F}_K] \leq (1 - 2\eta \gamma_k + \gamma_k^2 L^2) \|(x_K - x^*)\|^2 + \frac{\gamma_k^2 \nu^2}{N}. \]

Taking unconditional expectations, we obtain the following:

\[ \mathbb{E}[\|x_{K+1} - x^*\|^2] \leq q_K \mathbb{E}[\|(x_K - x^*)\|^2] + \frac{\gamma_k^2 \nu^2}{N}, \]

where \( q_K := (1 - 2\eta \gamma_k + \gamma_k^2 L^2) \). Consequently, we have the following:

\[
\mathbb{E}[\|x_{K+1} - x^*\|^2] \leq q_K q_{K-1} \mathbb{E}[\|(x_{K-1} - x^*)\|^2] + q_K \frac{\gamma_{K-1}^2 \nu^2}{N} + \frac{\gamma_k^2 \nu^2}{N} \\
\leq D \prod_{i=1}^{K} q_i + q_K \sum_{i=2}^{K} q_i \frac{\gamma_i^2 \nu^2}{N} + \frac{\gamma_k^2 \nu^2}{N} + q_K \frac{\gamma_{K-1}^2 \nu^2}{N} + \frac{\gamma_k^2 \nu^2}{N}.
\]

(2)

(i) Suppose \( \gamma_k := \gamma \) for all \( k \), implying that \( q_k = q \) for all \( k \). Therefore, we may further bound (2) as follows:

\[ \text{RHS of (2)} = q^K D + \sum_{j=0}^{K-1} q^j \frac{\gamma^2 \nu^2}{N} = q^K D + \frac{\gamma^2 \nu^2}{N} \left( \frac{1 - q^{K-2}}{1-q} \right). \]

Suppose we choose \( N = \lceil \beta_k q^{-K} \rceil \) such that \( K \beta_k q^{-K} = M - K \). Consequently, given \( q \), \( \beta_k \) can be expressed as follows:

\[ \beta_k = \left( \frac{M}{K} - 1 \right) q^K. \]

It follows that since \( N \geq \beta_k q^{-K} \), we have that the following holds:

\[ q^K D + \frac{\gamma^2 \nu^2}{N} \left( \frac{1 - q^{K-2}}{1-q} \right) \leq q^K D + \frac{\gamma^2 \nu^2}{\beta_k q^{-K}} \left( \frac{1 - q^{K-2}}{1-q} \right) \leq q^K D + \frac{\gamma^2 \nu^2 q^K}{\beta_k} \left( \frac{1}{1-q} \right). \]

In effect, we have that the following holds for \( k \geq 0:

\[ \mathbb{E}[\|x_{k+1} - x^*\|^2] \leq q^K \left( D + \frac{\gamma^2 \nu^2}{\beta_k} \left( \frac{1}{1-q} \right) \right). \]
Theorem 3

While a cursory look might suggest a geometric rate of convergence, a closer study suggests that this is not the case given that \( \beta \)

Remark: While a cursory look might suggest a geometric rate of convergence, a closer study suggests that this is not the case given that \( \beta \) and \( q \) are both dependent on \( K \). However, if \( K \leq K^* \), a consequence of imposing a computational budget, we note that a geometric rate of convergence may be obtained.

Theorem 3 (Linear convergence rate under finite computational budget) Suppose \( K \) projection steps are available.

(i) Let \( \gamma_k := \theta/k \) for all \( k \). Then the following holds:

\[
\mathbb{E}[\|x_{k+1} - x^\star\|^2] \leq q_k^k \left( D + \frac{\theta^2 \nu^2}{\bar{\beta}} \left( \frac{1}{1 - q_k} \right) \right), \quad \text{for all } k \leq K.
\]

(ii) Let \( \gamma_k := \theta/k \) for all \( k \). Then the following holds:

\[
\mathbb{E}[\|x_{k+1} - x^\star\|^2] \leq q_k^k \left( D + \frac{\theta^2 \nu^2}{6\bar{\beta}} \left( \frac{1}{1 - q_k} \right) \right), \quad \text{for all } k \leq K.
\]
Proof. (i) Suppose $K \leq \bar{K}$. Then we have that
\[ \beta_K = \left( \frac{M}{K} - 1 \right) q^K \geq \left( \frac{M}{\bar{K}} - 1 \right) q^\bar{K} \geq \left( \frac{M}{\bar{K}} - 1 \right) q^\bar{K} \triangleq \beta_{\bar{K}}. \]

It follows that when computational budget is bounded and steplengths are fixed, we have that for $K \leq \bar{K}$:
\[ \mathbb{E}[\|x_{K+1} - x^*\|^2] \leq q^K \left( D + \frac{\gamma^2 v^2}{\beta_{\bar{K}}} \frac{1}{1-q} \right). \]

(ii) Consider the case where $\gamma_k = \theta/k$. Since $q_K \leq q_1$, it follows that
\[ \left( \frac{1}{1-q_K} \right) \leq \left( \frac{1}{1-q_1} \right). \]

As a result, when computational budget is bounded and steplengths are diminishing, we have that for $K \leq \bar{K}$:
\[ \mathbb{E}[\|x_{K+1} - x^*\|^2] \leq q_K^k \left( D + \frac{\pi^2 \theta^2 v^2}{6 \beta_k} \frac{1}{1-q_k} \right) \leq q_K^k \left( D + \frac{\pi^2 \theta^2 v^2}{6 \beta_{\bar{K}}} \left( \frac{1}{1-q_1} \right) \right), \quad \forall K \leq \bar{K}. \]

Remark: This result is instructive in that one can derive an error bound that reduces to that observed in strongly convex deterministic regimes when the uncertainty disappears. Specifically, if $v = 0$, we notice that $\|x^k - x^*\|^2 \leq q^k \|x_0 - x^*\|^2$, in the constant steplength regime while in the diminishing steplength regime, we obtain $\|x^k - x^*\|^2 \leq \Pi_{j=1}^k \|x_0 - x^*\|^2 \leq q_k^k \|x_0 - x^*\|^2$.

We conclude this section with an investigation of the optimal $K$ (or equivalently the optimal $N$) in the constant steplength regime, obtained by minimizing the upper bound.

**Proposition 4 (Optimal choice of $K$ and $N$)** Consider the SA scheme in which $\gamma_k = \gamma$ and $N_k = N$ for all $k$. Then the following hold:

(i) The error bound $h(K)$ is convex in $K$ for all $K$ where
\[ h(K) \triangleq q^K D + \frac{1}{\left( \frac{M}{K} - 1 \right)} \frac{\gamma^2 v^2}{1-q}. \]

(ii) Suppose $\gamma$ satisfies the following:
\[ \frac{v^2}{MD} < \left( \frac{2\eta}{\gamma} - L^2 \right) \ln(1/q(\gamma)). \]

Then the optimal $K^*$ is achieved in the interior of $[0,M]$. 

Proof. (i): Consider $h(K)$ which is defined as follows:
\[ \mathbb{E}[\|x_{K+1} - x^*\|^2] \leq q^K \left( D + \frac{\gamma^2 v^2}{\beta_k} \frac{1}{1-q} \right) = q^K D + \frac{1}{\left( \frac{M}{K} - 1 \right)} \frac{\gamma^2 v^2}{1-q} \triangleq h(K). \]
Shanbhag and Blanchet

The first and second derivatives of \( h \) can then be derived.

\[
h'(K) = Dq^K \ln(q) + \frac{1}{(M - K)^2} \frac{M \gamma^2 \nu^2}{1 - q} = Dq^K \ln(q) + \frac{M \gamma^2 \nu^2}{(M - K)^2} \frac{1}{1 - q}
\]

\[
h''(K) = Dq^K (\ln(q))^2 + \frac{2M \gamma^2 \nu^2}{(M - K)^3} \frac{1}{1 - q} > 0, \quad \text{for all } K.
\]

It follows that \( h(K) \) is convex in \( K \).

(ii) An unconstrained minimizer \( K^* \) is given by a solution to

\[
(M - K)^2 q^K = \frac{\gamma^2 \nu^2 M}{\ln(1/q)(1 - q)D}.
\]

We now proceed to show that \( K^* \) is indeed feasible with respect to the bound constraint \( K \in [0, M] \):

- \( K \leq M \): Since \( q < 1 \), we have that

\[
\frac{\gamma^2 \nu^2 M}{\ln(1/q)(1 - q)D} = (M - K)^2 q^K \leq (M - K)^2
\]

\[
\implies (M - K) \geq \sqrt{\frac{\gamma^2 \nu^2 M}{\ln(1/q)(1 - q)D}} \quad \text{or} \quad K \leq M - \sqrt{\frac{\gamma^2 \nu^2 M}{\ln(1/q)(1 - q)D}} \leq M.
\]

- \( K \geq 0 \): This follows by showing that under the stated assumption, we have that \( h'(0) < 0 \). Specifically, we have that

\[
D \ln(q(\gamma)) + \frac{M \gamma^2 \nu^2}{M^2 (1 - q(\gamma))} = D \ln(q(\gamma)) + \frac{\gamma^2 \nu^2}{M(1 - q(\gamma))} = D \ln(q(\gamma)) + \frac{\gamma \nu^2}{M(2\eta - \gamma L^2)}.
\]

Through a rearrangement of terms, we have that \( h'(0) < 0 \) if

\[
\frac{\nu^2}{MD} < \left( \frac{2\eta}{\gamma} - L^2 \right) \ln(1/q(\gamma)),
\]

which holds by assumption.

\[\blacksquare\]

Remark: The reader may note that the optimal choice of \( K \) satisfies the following bound:

\[
K \leq M - d(\gamma) \sqrt{\frac{MV^2}{D}},
\]

where \( d(\gamma) = \sqrt{\frac{\gamma}{\ln(1/q(\gamma))(2\eta - \gamma L^2)}} \). In fact, this optimal choice of \( K \) contrasts with the standard value of \( K = M \) which is a result of using a single sample at every stage. Given the size restrictions of the current paper, we do not investigate the condition on \( \gamma \) that is essential to guarantee that \( h'(0) < 0 \) and leave it for future consideration.
3 Stochastic approximation with increasing sample sizes

In the prior section, we considered a setting where the sample size $N_k$ was fixed for every step at $N$. In this section, we consider an alternate approach in which the sample size is raised at every step through a prescribed update rule, with the overall goal of obtaining linear convergence rates over a finite horizon. Suppose $N_k \geq \left\lceil \frac{\beta_k}{\prod_{j=1}^{i} q_j} \right\rceil$. Then if the overall sampling budget is denoted by $M$, we have that

$$M = \sum_{k=1}^{K} N_k \geq \sum_{k=1}^{K} \frac{\beta_k}{\prod_{j=1}^{i} q_j} = M - K \implies \beta_k \leq \frac{M - K}{\sum_{k=1}^{K} \frac{1}{\prod_{j=1}^{i} q_j}}.$$

**Proposition 5** Consider the scheme (VSSA) and the associated mean-squared error after $K$ steps where $N_k \geq \left\lceil \frac{\beta_k}{\prod_{j=1}^{i} q_j} \right\rceil$ for all $k$ where $\beta_k = \frac{M - K}{\sum_{k=1}^{K} \frac{1}{\prod_{j=1}^{i} q_j}}$. Then the following hold:

(i) Suppose $\gamma_k := \gamma$ for all $k$. Then the mean-squared error may be bounded as follows:

$$\mathbb{E}[\|x_{K+1} - x^*\|^2] \leq q_k^2 D + \frac{\gamma_k^2 \nu^2 K}{M} \left( \frac{1 - q_k^{K-2}}{1 - q_k} \right).$$

(ii) Suppose $\gamma_k := \theta / k$ for all $k$. Then the mean-squared error may be bounded as follows:

$$\mathbb{E}[\|x_{K+1} - x^*\|^2] \leq q_k^2 D + \frac{\pi^2 \theta^2 \nu^2 K}{6M} \left( \frac{1 - q_k^{K-2}}{1 - q_k} \right).$$

**Proof.** After $K$ gradient steps, the mean-squared error is given by the following:

$$\mathbb{E}[\|x_{K+1} - x^*\|^2 | \mathcal{F}_K] \leq (1 - 2\eta \gamma_k + \gamma_k^2 L^2) \|(x_k - x^*)\|^2 + \frac{\gamma_k^2 \nu^2}{N_k}.$$

Taking unconditional expectations, we obtain the following:

$$\mathbb{E}[\|x_{K+1} - x^*\|^2] \leq q_k \mathbb{E}[\|(x_k - x^*)\|^2] + \frac{\gamma_k^2 \nu^2}{N_k},$$

where $q_k := (1 - 2\eta \gamma_k + \gamma_k^2 L^2)$. Consequently, we have the following:

$$\mathbb{E}[\|x_{K+1} - x^*\|^2] \leq q_k q_k \mathbb{E}[\|(x_{K-1} - x^*)\|^2] + q_k \gamma_{K-1}^2 \nu^2 + \gamma_k^2 \nu^2$$

$$\leq D \prod_{i=1}^{K} q_i + \left( \prod_{i=2}^{K} q_i \right) \frac{\gamma_1^2 \nu^2}{N_1} + \ldots + q_k q_k \gamma_{K-2}^2 \nu^2 + q_k \gamma_{K-1}^2 \nu^2 + \frac{\gamma_k^2 \nu^2}{N_{K-1}} + \gamma_k^2 \nu^2.$$

(i) Suppose $\gamma_k := \gamma$ for all $k$, implying that $q_k = q$ for all $k$. Therefore, we may further bound the above expression as follows:

$$D \prod_{i=1}^{K} q_i + \prod_{i=2}^{K} q_i \frac{\gamma_1^2 \nu^2}{N_1} + \ldots + q_k q_k \gamma_{K-2}^2 \nu^2 + q_k \gamma_{K-1}^2 \nu^2 + \frac{\gamma_k^2 \nu^2}{N_K}$$

$$\leq Dq^K + q^K \frac{\gamma_1^2 \nu^2}{\beta_K} + \ldots + q_k \gamma_{K-2}^2 \nu^2 + q_k \gamma_{K-1}^2 \nu^2 + q_k \gamma_k^2 \nu^2 = Dq^K + q^K \frac{\gamma^2 \nu^2}{\beta_K},$$

where $\beta_K = \frac{M - K}{\sum_{k=1}^{K} \frac{1}{\prod_{j=1}^{i} q_j}}$, implying that

$$\mathbb{E}[\|x_{K+1} - x^*\|^2] \leq q^K \left( D + \frac{\gamma^2 \nu^2 K}{\beta_K} \right).$$
Suppose $\gamma_k := \theta/k$ for all $k$ and by recalling that $q$ is a decreasing function in $\gamma$, we may further bound the above expression as follows:

\[
D \prod_{i=1}^{K} q_i + \prod_{i=2}^{K} q_i \frac{\gamma_i^2 v^2}{N_1} + \cdots + q_K q_{K-1} \frac{\gamma_{K-2}^2 v^2}{N_{K-2}} + q_K \frac{\gamma_{K-1}^2 v^2}{N_{K-1}} + \frac{\gamma_K^2 v^2}{N_K}
\]

\[
= D \prod_{i=1}^{K} q_i + \prod_{i=2}^{K} q_i \frac{\gamma_i^2 v^2}{\beta_k q_i^{-1}} + \cdots + q_K q_{K-1} \frac{\gamma_{K-2}^2 v^2}{\beta_k \prod_{i=0}^{K-2} q_i^{-1}} + q_K \frac{\gamma_{K-1}^2 v^2}{\beta_k \prod_{i=0}^{K-1} q_i^{-1}} + \frac{\gamma_K^2 v^2}{\beta_k \prod_{i=0}^{K} q_i^{-1}}
\]

\[
\leq D \left( \prod_{i=1}^{K} q_i \right) + \left( \prod_{i=1}^{K} q_i \right) \sum_{j=1}^{K-1} \frac{\theta j^2 v^2}{\beta_k j^2} \leq \left( \prod_{i=1}^{K} q_i \right) \left( D + \frac{\theta j^2 v^2}{6\beta_k} \sum_{j=1}^{K-1} \frac{\gamma_j^2 v^2}{\beta_k j^2} \right),
\]

where the second inequality is a consequence of Hölder’s inequality. By recalling that $\sum_{j=1}^{K-1} 1/j^2 \leq 1/6$, we can further bound this expression as follows:

\[
\left( \prod_{i=1}^{K} q_i \right) \left( D + \frac{\theta j^2 v^2}{6\beta_k} \sum_{j=1}^{K-1} \frac{\gamma_j^2 v^2}{\beta_k j^2} \right) \leq \left( \prod_{i=1}^{K} q_i \right) \left( D + \frac{\theta j^2 v^2}{6\beta_k} \right) \Rightarrow \mathbb{E}[\|x_{k+1} - x^*\|^2] \leq q_k^K \left( D + \frac{\theta j^2 v^2}{6\beta_k} \right).
\]

We now examine the convergence rate when constrained by a finite computational budget.

**Theorem 6 (Linear convergence rate under finite computational budget)** Suppose $K$ projection steps are available.

(i) Let $\gamma_k = \gamma$ for all $k$. Then the following holds:

\[
\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq q_k \left( D + \frac{\gamma_j^2 v^2 K}{\beta_k} \right), \quad \forall k \leq K.
\]

(ii) Let $\gamma_k = \theta/k$ for all $k$. Then the following holds:

\[
\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq q_k^K \left( D + \frac{\theta j^2 v^2 \gamma_j^2 v^2}{6\beta_k} \right), \quad \forall k \leq K.
\]

**Proof.** (i): Consider the error given by $q_i^K \left( D + \frac{\gamma_j^2 v^2 K}{\beta_k} \right)$. If $K \leq \bar{K}$, then we have that

\[
\beta_K = \frac{M}{\sum_{k=1}^{K} q_k} \geq \frac{M}{\sum_{k=1}^{K} \frac{1}{q_k}} \triangleq \bar{\beta}_K.
\]

It follows that when computational budget is bounded and steplengths are fixed, we have that for $K \leq \bar{K}$:

\[
\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq q_k \left( D + \frac{\gamma_j^2 v^2 k}{\beta_k} \right) \leq q_k \left( D + \frac{\gamma_j^2 v^2 \bar{K}}{\bar{\beta}_k} \right), \quad \forall k \leq \bar{K}.
\]

(ii): When computational budget is bounded and steplengths are diminishing, we have that for $K \leq \bar{K}$:

\[
\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq q_k^K \left( D + \frac{\theta j^2 v^2 \gamma_j^2 v^2}{6\beta_k} \right) \leq q_k^K \left( D + \frac{\theta j^2 v^2 \gamma_j^2 v^2}{6\beta_k} \right), \quad \forall K \leq \bar{K}.
\]
We conclude this section with a discussion of how to optimally choose $K$ when steplength sequences are constant. We discuss the convexity of a bound on the mean-squared error which allows for deriving an optimal choice of $K$.

**Lemma 7** Consider the SA scheme in which $N_k$ is increasing as per the prescribed rule and $\gamma_k = \gamma$ for all $k$. Then the function $Dq^K + \frac{\gamma^2v^2}{K}$ is convex in $K$.

**Proof.** Since $\beta_k = \frac{M-K}{K^2}$, it follows that $\beta_k \geq \frac{M-K}{K^2}$ from noting that $1/q^j \geq 1/q^K$. Consequently $Dq^K + \frac{\gamma^2v^2}{K} \leq Dq + \frac{\gamma^2v^2}{K}$ and

$$
\begin{align*}
&h'(K) = DqK \ln(q) - \frac{\gamma^2v^2}{(M-K)^2} \left( -\frac{2M}{K^3} + \frac{1}{K} \right) = DqK \ln(q) + \frac{\gamma^2v^2}{(M-K)^2} \left( \frac{2M}{K} - 1 \right) \\
&h''(K) = DqK \ln^2(q) + \frac{M}{K^2} \frac{\gamma^2v^2}{(M-K)^2} \left( \frac{2M}{K} - 1 \right) + \frac{\gamma^2v^2}{(M-K)^2} \left( -\frac{M}{K^2} \right) \\
&\quad = DqK \ln^2(q) + \frac{M\gamma^2v^2}{(M-K)^2} \left( \frac{3(M/K) - 1}{((M/K) - 1)^2} \right) \geq 0, \text{ when } K \leq M.
\end{align*}
$$

**Remark:** Note that this prior lemma does not suffice in ensuring that the minimizer lies in $[0, M]$. This requires analysis akin to that provided in Prop. 4.

### 4 Numerical results

We provide some preliminary numerical investigations on a stochastic quadratic program defined as follows:

$$
\min_{x \in X} \mathbb{E} \left[ \frac{1}{2} x^T Q(\xi)x - d^T x \right],
$$

where $X \triangleq [0, 1]^n$. Furthermore, we assume that $\mathbb{E}[Q(\xi)] = 2I + R^T R$ where $R$ is a randomly generated $n$-dimensional square matrix (using a uniform distribution over $[0, 1]$) and $I$ denotes the identity matrix. Furthermore $d = -2e$ where $e$ denotes the column of ones in $n$-space. The optimal solution is in the interior of the set and is given by $Q^{-1}d$ where $Q = \mathbb{E}[Q(\xi)]$. We examine the performance of four SA schemes.

#### 4.1 Performance of SA schemes

In Table 2, we consider constant sample size SA schemes with constant and diminishing steplength sequences (with $\theta = 1$). It can be seen that while standard projected gradient schemes require 20 projection steps to reach an error of $2.3e-5$, the constant sample-size schemes provide solutions with an error of the order $1e-2$ in the same number of steps. Minimizing the theoretical error bound in the constant steplength regime leads to a $K^* = 48$ which corresponds well with what is seen from the profile of the theoretical error. When one compares the diminishing steplength scheme, it can be seen that there is a significant improvement in terms of empirical performance. This is not surprising since the diminishing nature of steplength helps mute the stochastic error. It is also observed that the theoretical error tends to have a minimizer around $K^* = 40$ and then slowly edges upward, a characteristic that is supported by the empirical behavior.

When one examines increasing sample-size schemes with constant and diminishing steplength sequences, the results tend to be somewhat better. For instance, with SA schemes with constant steplength and increasing
Table 1: Constant sample size SA schemes with constant and diminishing steplengths

| #  | $|\|x_d-K||^2$ | $K^{\text{inc}}$ | $|x-x^*||^2$ | $x^{\text{const}}$ | Theor. bound |
|----|----------------|----------------|----------------|----------------|--------------|
| 10 | 2.288e-05      | 20             | 1.411e-02      | 12             | 9.356e-01    |
| 20 | 2.288e-05      | 20             | 9.058e-03      | 22             | 2.421e-01    |
| 30 | 2.288e-05      | 20             | 1.198e-02      | 33             | 6.313e-02    |
| 40 | 2.288e-05      | 20             | 1.319e-02      | 42             | 2.398e-02    |
| 50 | 2.288e-05      | 20             | 1.424e-02      | 52             | 2.979e-02    |
| 60 | 2.288e-05      | 20             | 1.607e-02      | 62             | 2.238e-02    |
| 70 | 2.288e-05      | 20             | 1.757e-02      | 72             | 2.415e-02    |
| 80 | 2.288e-05      | 20             | 1.888e-02      | 82             | 2.582e-02    |
| 90 | 2.288e-05      | 20             | 1.967e-02      | 91             | 2.066e-02    |

Table 2: Increasing sample size SA schemes with constant and diminishing steplengths

| #  | $|\|x_d-K||^2$ | $K^{\text{inc}}$ | $|x-x^*||^2$ | $x^{\text{const}}$ | Theor. bound |
|----|----------------|----------------|----------------|----------------|--------------|
| 10 | 2.288e-05      | 20             | 1.411e-02      | 12             | 9.356e-01    |
| 20 | 2.288e-05      | 20             | 3.385e-03      | 22             | 2.421e-01    |
| 30 | 2.288e-05      | 20             | 3.385e-03      | 32             | 6.313e-02    |
| 40 | 2.288e-05      | 20             | 3.385e-03      | 42             | 2.398e-02    |
| 50 | 2.288e-05      | 20             | 3.389e-03      | 52             | 2.979e-02    |
| 60 | 2.288e-05      | 20             | 3.389e-03      | 62             | 2.238e-02    |
| 70 | 2.288e-05      | 20             | 3.389e-03      | 72             | 2.415e-02    |
| 80 | 2.288e-05      | 20             | 3.389e-03      | 82             | 2.582e-02    |
| 90 | 2.288e-05      | 20             | 3.384e-03      | 91             | 2.066e-02    |

4.2 Optimal choices of $K$ and $N$

In the prior subsections, we have examined the question of minimizing the theoretical bound in $K$ to ascertain the optimal number of projection steps (and therefore the optimal sample size). In Figure 1, we plot the theoretical and empirical error for constant steplength SA schemes with constant and increasing sample sizes. From this figure, we see that the theoretical error is minimized in both regimes and this behavior is matched to some degree by the empirical error.

4.3 Linear convergence rates with finite computational budget

A key question that has motivated this work is whether one can develop methods that provide faster convergence rates which are valid in a finite (rather than non-asymptotic) regime. We note that by imposing a budget on the computational complexity, we may derive precisely such rates and these rates. Naturally these rates are weaker than the empirical behavior but Figure 2 demonstrates that the linear nature of these rates.
5 Concluding remarks

We present a set of stochastic approximation schemes that can contend with bounds on the simulation budget under either constant or increasing sample sizes at every iteration. In particular, we show that by suitable allocation policies of the simulation budget, the schemes display linear convergence, albeit with a poorer constant. Furthermore, optimal allocation policies can be derived under some conditions.

REFERENCES