Rate of Convergence Analysis of Discrete Simultaneous Perturbation Stochastic Approximation Algorithm

Qi Wang and James C. Spall

Abstract—A middle point discrete version of simultaneous perturbation stochastic approximation (DSPSA) algorithm was previously introduced to solve the discrete stochastic optimization problem. We consider the rate of convergence of DSPSA in this paper. This rate will allow for objective comparisons with other discrete stochastic optimization methods.

I. INTRODUCTION

Wang and Spall (2011) introduce an algorithm, discrete simultaneous perturbation stochastic approximation (DSPSA), which is designed to solve discrete stochastic optimization problems. DSPSA is motivated by SPSA, which is discussed in Spall (1992). Wang and Spall (2011) show that under some conditions, DSPSA leads to a sequence $\{\hat{\theta}_k\}$ converging to the optimal solution $\theta^*$ almost surely, where $\theta$ represents the parameter vector being optimized. In this paper, we discuss the rate of convergence of DSPSA. We show that $E \| \hat{\theta}_k - \theta^* \|^2$ goes to 0 at the rate $O(1/k^\alpha)$, where $\alpha$ is the decaying rate of the gain (step size) sequence and $k$ is the iteration index. This result can produce the rate of convergence of $P(\hat{\theta}_k = \theta^*)$, where $[\hat{\theta}_k]$ indicates the nearest multivariate-integer point of $\hat{\theta}_k$. The convergence rate of $P(\hat{\theta}_k = \theta^*)$ can be used to compare DSPSA with other discrete stochastic algorithms (Alrefaei and Andradottir (1999), Andradottir (1995), Yan and Mukai (1992), Gong, Ho, and Zhai (2000), Hong and Nelson (2006), Sklenar and Popela (2010), Hill (2004), Spall (2003) etc).

II. ALGORITHM DESCRIPTION

First of all, we describe the algorithm of DSPSA. We consider the case when $\theta$ is $p$-dimensional, $p = 1, 2, 3, \ldots$. We have the algorithm as below for function $y = L + \varepsilon$, where $L: \mathbb{Z}^p \to \mathbb{R}$ and $\varepsilon$ is noise.

The basic algorithm is:

Step 1: Pick an initial guess $\hat{\theta}_0$.

Step 2: Generate $\Delta_k = [\Delta_{k1}, \Delta_{k2}, \ldots, \Delta_{kp}]^T$, where $\Delta_{ki}$ are independent Bernoulli random variables taking the values $\pm 1$ with probability $1/2$.

Step 3: $\pi(\hat{\theta}_k) = \left[ \hat{\theta}_k + 1_p/2 \right]$ where $1_p$ is a $p$-dimensional vector with all components being unity, and $\left[ \hat{\theta}_k \right] + \left[ \hat{\theta}_{k1}, \ldots, \hat{\theta}_{kp} \right]^T$. ($\pi(\hat{\theta}_k)$ is the middle point of unit hypercube, and $\left[ \right]$ is the operation of floor function.)

Step 4: Evaluate $y$ at $\pi(\hat{\theta}_k) + \Delta_k/2$ and $\pi(\hat{\theta}_k) - \Delta_k/2$, and form the estimate of $\hat{g}_k(\hat{\theta}_k)$.

$$\hat{g}_k(\hat{\theta}_k) = \left[ y \left( \pi(\hat{\theta}_k) + \frac{1}{2} \Delta_k \right) - y \left( \pi(\hat{\theta}_k) - \frac{1}{2} \Delta_k \right) \right] \Delta_k^{-1},$$

where $\Delta_k^{-1} = [\Delta_{k1}^{-1}, \ldots, \Delta_{kp}^{-1}]^T$.

Step 5: Update the estimate according to the recursion

$$\hat{\theta}_{k+1} = \hat{\theta}_k - d_k \hat{g}_k(\hat{\theta}_k).$$

Step 6: After $M$ iterations, set the approximated optimal solution to be $\hat{\theta}_M$, where $M$ is the maximum number of allowed iterations based on the cost limit.

In the theoretical analysis below, we use the following mean gradient-like expression centered at $\pi(\theta)$:

$$\mathbb{E}(\pi(\theta)) = \frac{E}{L} \left[ L \left( \pi(\theta) + \frac{1}{2} \Delta \right) - L \left( \pi(\theta) - \frac{1}{2} \Delta \right) \right] \Delta^{-1} | \theta \right\},$$

where $\Delta$ is a $p$-dimensional vector that has the same definition as $\Delta_k$ mentioned above, and $\theta$ may be a random variable in some cases. If each direction is chosen equally, then

$$\mathbb{E}(\pi(\theta)) = \frac{1}{2^p} \sum_{\Delta} \left[ L \left( \pi(\theta) + \frac{1}{2} \Delta \right) - L \left( \pi(\theta) - \frac{1}{2} \Delta \right) \right] \Delta^{-1},$$

where $\sum_{\Delta}$ indicates the summation over all possible directions $\Delta$. Note that $\Delta_k^{-1} = \Delta_k$ and $\Delta^{-1} = \Delta$ in the Bernoulli $\pm 1$ case; we use $\Delta_k^{-1}$ to accommodate future extensions to perturbation distributions other than Bernoulli $\pm 1$.  

Qi Wang is with the Department of Applied Mathematics and Statistics of the Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: qwang29@jhu.edu).

James C. Spall is with the Johns Hopkins University, Applied Physics Laboratory, Laurel, MD 20723-6099 USA.
III. RATE OF CONVERGENCE ANALYSIS

Now let us consider the rate of convergence of DSPSA. Generally, for convergent discrete stochastic algorithm, if the optimal solution is unique and all the points in the sequence \(\{\hat{\theta}_k\}\) are multivariate integer points, then it is natural to use \(P(\hat{\theta}_k \neq \theta^*)\) as the measure of rate of convergence. However, for DSPSA, \(\{\hat{\theta}_k\}\) are non-multivariate integer points, then we can use \(E\|\hat{\theta}_k - \theta^*\|^2\) as the measurement. Due to the relationship: \(E\|\hat{\theta}_k - \theta^*\|^2 \geq 0^2 P([\hat{\theta}_k] = \theta^*) + 0.5^2 P([\hat{\theta}_k] \neq \theta^*) = 0.25 P([\hat{\theta}_k] \neq \theta^*)\), where \([\hat{\theta}_k]\) indicates the nearest multivariate-integer point of \(\hat{\theta}_k\), we can get an upper bound of \(P([\hat{\theta}_k] \neq \theta^*)\) to compare DSPSA with other algorithms in the large \(O\). Suppose \(\mathcal{X}_k = \{\theta_0, \theta_1, \ldots, \theta_k\}\).

Theorem 1. Assume \(L\) is a function on \(\mathbb{Z}^p\), and \(L\) has an unique minimal point \(\theta^*\). Assume also (i) \(a_k = a/(1 + A + k)^\alpha\), \(0.5 < \alpha \leq 1\), \(a\) and \(A\) are positive scalars; (ii) the components of \(\Delta_k\) are independently Bernoulli \(\pm 1\) distributed; (iii) For all \(k\), \(E[\epsilon_k^2 - \epsilon_k^2] \cdot \mathcal{X}_k, \Delta_k] = 0\) a.s., and \(\text{Var}(\epsilon_k^2)\) is uniformly bounded in \(k\); (iv) \(E\left[L\left(\pi(\theta_k) + \frac{1}{2} \Delta_k\right) - L\left(\pi(\theta_k) - \frac{1}{2} \Delta_k\right)\right]^2\) is uniformly bounded for all \(k\); (v) there exists \(\mu > 0\) such that \(E\left[\left(\hat{\theta}_k - \theta^*\right)^T g(\pi(\theta_k)) - \mu (\hat{\theta}_k - \theta^*)^T (\hat{\theta}_k - \theta^*)\right] \geq 0\) for all \(k\); (vi) \(2a/(1 + A + k)^\alpha < 1\) for all \(k\). Then

\[
E\|\hat{\theta}_k - \theta^*\|^2 \leq \exp\left(\frac{2\mu a \left(\left(1 + A\right)^{1-\alpha} - \left(1 + A + k\right)^{1-\alpha}\right)}{1 - \alpha}\right) E\|\hat{\theta}_0 - \theta^*\|^2 + \exp\left(-\frac{2\mu a \left(1 + A + k\right)^{1-\alpha}}{1 - \alpha}\right) pba^2 C(\alpha)
\]

\[
\leq k \times \left(1 + A + k\right)^{-2\alpha} \exp\left(\frac{2\mu a \left(1 + A + k\right)^{1-\alpha}}{1 - \alpha}\right) dx, 0.5 < \alpha < 1,
\]

\[
\frac{(1 + A)^2a}{(1 + A + k)2\mu a} E\|\hat{\theta}_0 - \theta^*\|^2 + \frac{pba^2 C(\alpha)}{(1 + A + k)2\mu a} \int_0^k (1 + A + x)^{2\mu a - 2} dx, \quad \alpha = 1,
\]

where for \(0.5 < \alpha < 1\)

\[
C(\alpha) = \exp\left(\frac{2\mu a \left(\left(1 + A + 1\right)^{1-\alpha} - \left(1 + A + A\right)^{1-\alpha}\right)}{1 - \alpha}\right) \left(1 + \frac{1}{1 + A}\right)^{2\alpha};
\]

for \(\alpha = 1\) we have

\[
C(1) = \left(1 + \frac{1}{1 + A}\right)^{2\mu a + 2};
\]

and \(b\) is an uniform upper bound for \(E\left[\epsilon_k^2 - \epsilon_k^2\right]\)

\[
+ E\left[L\left(\pi(\hat{\theta}_k) + \frac{1}{2} \Delta_k\right) - L\left(\pi(\hat{\theta}_k) - \frac{1}{2} \Delta_k\right)\right]^2.
\]

Remarks:

1. Condition (v) can be written as

\[
E\left[(\hat{\theta}_k - \theta^*)^T g(\pi(\hat{\theta}_k)) \right] \geq \mu E\left[(\hat{\theta}_k - \theta^*)^T (\hat{\theta}_k - \theta^*)\right],
\]

which is analog to the definition of strongly convexity. Thus \(\mu\) is determined by the structure (curvature) of the loss function.

2. Condition (iv) requires the sequence of \(\{\hat{\theta}_k\}\) does not jump to the “far away” area so often.

3. When \(\alpha \rightarrow 1\), the form of the upper bound for \(0.5 < \alpha < 1\) converges to the upper bound for the case of \(\alpha = 1\). We further discuss it in Lemma 1 below.

4. Condition (vi) is used to keep the stability of the algorithm and this condition is automatically satisfied when \(k\) is large enough. When \(k\) is small, we can choose large \(A\) to make condition (vi) to be satisfied.

Lemma 1. The upper bound for the case \(0.5 < \alpha < 1\) converges to the bound for the case of \(\alpha = 1\), when \(\alpha \rightarrow 1\).

Proof of Lemma 1. By L’Hopital’s rule, we have

\[
\lim_{\alpha \rightarrow 1} \frac{2\mu a \left(\left(1 + A\right)^{1-\alpha} - \left(1 + A + x\right)^{1-\alpha}\right)}{1 - \alpha} = \ln(1 + A)^{2\mu a} - \ln(1 + A + x)^{2\mu a},
\]

which indicates that

\[
\lim_{\alpha \rightarrow 1} \exp\left(\frac{2\mu a \left(\left(1 + A\right)^{1-\alpha} - \left(1 + A + x\right)^{1-\alpha}\right)}{1 - \alpha}\right) = \left(1 + A\right)^{2\mu a}.
\]

Then we have

\[
\lim_{\alpha \rightarrow 1} C(\alpha) = \lim_{\alpha \rightarrow 1} \exp\left(\frac{2\mu a \left(\left(1 + A + 1\right)^{1-\alpha} - \left(1 + A + A\right)^{1-\alpha}\right)}{1 - \alpha}\right) \left(1 + \frac{1}{1 + A}\right)^{2\alpha} = \left(1 + \frac{1}{1 + A}\right)^{2\mu a + 2} = C(1).
\]
In addition,
\[
\lim_{\alpha \to 1} \left( \frac{2\mu a (1 + A + k) - 1}{1 - \alpha} \right)
\times k \int_0^k (1 + A + x)^{2\alpha} \exp\left( \frac{2\mu a (1 + A + x + k + 1)^{1-\alpha}}{1 - \alpha} \right) dx
= \lim_{\alpha \to 1} k \int_0^k (1 + A + x)^{2\alpha} \exp\left( \frac{2\mu a (1 + A + x)^{1-\alpha} - (1 + A + k + 1)^{1-\alpha}}{1 - \alpha} \right) dx
= \frac{1}{(1 + A + k)^{2\mu a - 2}} \int_0^k (1 + A + x)^{2\mu a - 2} dx.
\]

Thus the result in Lemma 1 is true. Q.E.D.

In Lemma 2, we set up a lower bound for \( C(\alpha) \), and the result is used in the mathematics induction of Theorem 1.

\[ C(\alpha) \geq \frac{(1 + A + k + 1)^{2\mu a}}{(1 + A + k)^2} \int_k^{k+1} (1 + A + x)^{2\mu a - 2} dx \]

Proof of Lemma 2. For the case of \( 0.5 < \alpha < 1 \), by mean value theorem for integration, there exists \( x \in (k, k+1) \) such that
\[
\exp \left( \frac{2\mu a (1 + A + k + 1)^{1-\alpha}}{1 - \alpha} - \frac{2\mu a (1 + A + x)^{1-\alpha}}{1 - \alpha} \right)
\times \left( 1 + \frac{x - k}{1 + A + k} \right)^{2\alpha}.
\]

Furthermore, by Taylor expansion we have
\[
(1 + A + k + 1)^{1-\alpha} = (1 + A + x)^{1-\alpha} + (1 - \alpha)(k + 1 - x)(1 + A + x)^{\alpha},
\]
where \( \bar{x} \in (\bar{x}, k + 1) \), and
\[
(1 + A + 1)^{1-\alpha} = (1 + A)^{1-\alpha} + \frac{1 - \alpha}{1 + A + \bar{x})^{1-\alpha}},
\]
where \( \bar{x} \in (0, 1) \). Due to the value of \( \bar{x} \), \( x \) and \( \bar{x} \), we know that
(1 + A + 1)^{1-\alpha} - (1 + A)^{1-\alpha} \geq (1 + A + k + 1)^{1-\alpha} - (1 + A + \bar{x})^{1-\alpha}.

It follows that
\[
\exp \left( 2\mu a (1 + A + k + 1)^{1-\alpha} - (1 + A + \bar{x})^{1-\alpha} \right) \left( 1 + \frac{\bar{x} - k}{1 + A + k} \right)^{2\alpha}
\leq \exp \left( 2\mu a - (1 + A + k + 1)^{1-\alpha} - (1 + A)^{1-\alpha} \right) \left( 1 + \frac{1}{1 + A} \right)^{2\alpha} = C(\alpha).
\]

For the case of \( \alpha = 1 \), by mean value theorem for integration, we have
\[
\frac{(1 + A + k + 1)^{2\mu a}}{(1 + A + k)^2} \int_k^{k+1} (1 + A + x)^{2\mu a - 2} dx
= \frac{(1 + A + k + 1)^{2\mu a}}{(1 + A + \bar{y})^{2\mu a}} \int_k^{k+1} (1 + A + x)^{2\mu a - 2} dx
\leq \left( \frac{1}{1 + A} \right)^{2\mu a} = C(1).
\]

Q.E.D.

Proof of Theorem 1. The key formula for DSPSA is
\[
\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \left[ y \left( \pi(\hat{\theta}_k) + \frac{1}{2} \Lambda_k \right) - y \left( \pi(\hat{\theta}_k) - \frac{1}{2} \Lambda_k \right) \right] \Lambda_k^{-1}.
\]

Then we have
\[
E \left[ \left( \hat{\theta}_{k+1} - \theta^* \right)^T \left( \hat{\theta}_{k+1} - \theta^* \right) \right] = E \left[ \left| \hat{\theta}_k - \theta^* \right|^2 \right] - 2a_k
\times E \left[ \left( \hat{\theta}_k - \theta^* \right)^T \left[ y \left( \pi(\hat{\theta}_k) + \frac{1}{2} \Lambda_k \right) - y \left( \pi(\hat{\theta}_k) - \frac{1}{2} \Lambda_k \right) \right] \Lambda_k^{-1} - g \left( \pi(\hat{\theta}_k) \right) \right]
- 2a_k E \left[ \left( \hat{\theta}_k - \theta^* \right)^T g \left( \pi(\hat{\theta}_k) \right) \right]
+ a_k^2 E \left[ \left| y \left( \pi(\hat{\theta}_k) + \frac{1}{2} \Lambda_k \right) - y \left( \pi(\hat{\theta}_k) - \frac{1}{2} \Lambda_k \right) \right| \Lambda_k^{-1} \right].
\]

Let us now discuss the terms on the right hand side of equation (2). After dropping the \(-2a_k\) multiplier, the second term of the right hand side is
The fourth term on the right hand side of equation (2) is
\[ E\left[ y\left( \pi(\hat{\theta}_k) + \frac{1}{2} \Delta_k \right) - y\left( \pi(\hat{\theta}_k) - \frac{1}{2} \Delta_k \right) \right] \Lambda_k^{-1} \]
\[ = E\left[ \left( L\left( \pi(\hat{\theta}_k) + \frac{1}{2} \Delta_k \right) - L\left( \pi(\hat{\theta}_k) - \frac{1}{2} \Delta_k \right) \right)^2 \Lambda_k^{-1} \right] + 2E\left[ \left( L\left( \pi(\hat{\theta}_k) + \frac{1}{2} \Delta_k \right) - L\left( \pi(\hat{\theta}_k) - \frac{1}{2} \Delta_k \right) \right)(\epsilon^*_k - \epsilon_k) \right] \Lambda_k^{-1} \Lambda_k \]
\[ + E\left[ (\epsilon^*_k - \epsilon_k)^2 \right] \Lambda_k^{-1} \Lambda_k. \]

By condition (ii) and condition (iii), we have
\[ E\left[ \left( L\left( \pi(\hat{\theta}_k) + \frac{1}{2} \Delta_k \right) - L\left( \pi(\hat{\theta}_k) - \frac{1}{2} \Delta_k \right) \right)(\epsilon^*_k - \epsilon_k) \right] = 0. \]

Then the forth term on the right hand side of equation (2) can be written as
\[ E\left[ y\left( \pi(\hat{\theta}_k) + \frac{1}{2} \Delta_k \right) - y\left( \pi(\hat{\theta}_k) - \frac{1}{2} \Delta_k \right) \right] \Lambda_k^{-1} \]
\[ = E\left[ \left( L\left( \pi(\hat{\theta}_k) + \frac{1}{2} \Delta_k \right) - L\left( \pi(\hat{\theta}_k) - \frac{1}{2} \Delta_k \right) \right)^2 \Lambda_k^{-1} \right] + 2E\left[ \left( L\left( \pi(\hat{\theta}_k) + \frac{1}{2} \Delta_k \right) - L\left( \pi(\hat{\theta}_k) - \frac{1}{2} \Delta_k \right) \right)(\epsilon^*_k - \epsilon_k) \right] \Lambda_k^{-1} \Lambda_k \]
\[ + E\left[ (\epsilon^*_k - \epsilon_k)^2 \right] \Lambda_k^{-1} \Lambda_k. \]

Due to condition (iii) and (iv), we get \( E(\epsilon^*_k - \epsilon_k)^2 \) and
\[ E\left[ \left( L\left( \pi(\hat{\theta}_k) + \frac{1}{2} \Delta_k \right) - L\left( \pi(\hat{\theta}_k) - \frac{1}{2} \Delta_k \right) \right)^2 \right] \] are uniformly bounded respectively. Thus, there exists a positive scalar \( b \) such that
\[ E\left[ \left( L\left( \pi(\hat{\theta}_k) + \frac{1}{2} \Delta_k \right) - L\left( \pi(\hat{\theta}_k) - \frac{1}{2} \Delta_k \right) \right)^2 \right] + E(\epsilon^*_k - \epsilon_k)^2 \leq b, \] which indicates that the value of equation (3) is smaller than or equal to \( pb \). It follows
\[ E\left[ \hat{\theta}_{k+1} - \theta^* \right]^2 \leq E\left[ \hat{\theta}_k - \theta^* \right]^2 - 2a_k E\left[ \left( \hat{\theta}_k - \theta^* \right)^T \tilde{g}(\pi(\hat{\theta}_k)) \right] + a_k^2 pb. \]

Furthermore, by condition (v), we have
\[ E\left[ \hat{\theta}_k - \theta^* \right]^2 \leq (1 - 2a_k \mu) E\left[ \hat{\theta}_k - \theta^* \right]^2 + a_k^2 pb \tag{4} \]
for all \( k \).

Now let us prove Theorem 1 by the method of induction. First let us consider the case of \( 0.5 < \alpha < 1 \). For the base case: \( k = 0 \), we have
\[ \exp\left( \frac{2\mu a (1 + A)^{1-\alpha}}{1 - \alpha} - \frac{2\mu a (1 + A + k)^{1-\alpha}}{1 - \alpha} \right) E\left[ \hat{\theta}_0 - \theta^* \right]^2 \]
\[ + \exp\left( \frac{2\mu a (1 + A + k)^{1-\alpha}}{1 - \alpha} \right) p b a^2 C(\alpha) \]
\[ \times \int_0^1 (1 + A + x)^{-2\alpha} \exp\left( \frac{2\mu a (1 + A + x)^{1-\alpha}}{1 - \alpha} \right) dx \]
\[ = E\left[ \hat{\theta}_0 - \theta^* \right]^2. \tag{5} \]

The equation (5) indicates the inequality (1) is true for \( k = 0 \). Suppose the inequality (1) is true for \( k \), then for the case of \( k+1 \), by recursive relationship (4), we have
\[ E\left[ \hat{\theta}_{k+1} - \theta^* \right]^2 \]
\[ \leq \left( 1 - 2a_k \mu \right) E\left[ \hat{\theta}_k - \theta^* \right]^2 + a_k^2 pb \]
\[ = \left( 1 - \frac{2\mu a}{1 + A + k} \right) E\left[ \hat{\theta}_k - \theta^* \right]^2 + \frac{a_k^2 pb}{(1 + A + k)^{2\alpha}}. \tag{6} \]

By substituting the upper bound of \( E\left[ \hat{\theta}_k - \theta^* \right]^2 \) into inequality (6), we have...
Due to Lemma 2, we have

\[
\frac{pba}{(1 + A + k)^2}.
\]

Now let us start to discuss the terms on the right hand side of inequality (7). By second order Taylor expansion, we have

\[
\exp\left(-\frac{2\mu a(1 + A + k + 1)^{1-\alpha}}{1-\alpha}\right)
\geq 1 - \frac{2\mu a(1 + A + k)^{1-\alpha}}{1-\alpha},
\]

where \(\Xi \in (k, k+1)\). And this bound is tighter for larger \(k\). Due to Lemma 2, we have

\[
\int_0^k (1 + A + x)^{-2\alpha} \exp\left(\frac{2\mu a(1 + A + x)^{1-\alpha}}{1-\alpha}\right) dx
\]

\[
+(1 + A + k)^{2-2\alpha} C(\alpha)
\]

\[
\leq \int_0^k (1 + A + x)^{-2\alpha} \exp\left(\frac{2\mu a(1 + A + x)^{1-\alpha}}{1-\alpha}\right) dx
\]

\[
+k \int_k^{k+1} (1 + A + x)^{-2\alpha} \exp\left(\frac{2\mu a(1 + A + x)^{1-\alpha}}{1-\alpha}\right) dx
\]

\[
+\frac{pba^2 C(\alpha)}{(1 + A + k)^2}.
\]

which indicates that inequality (1) is true for \(k+1\) with regard to \(1/2 < \alpha < 1\).

For the case of \(\alpha = 1\), by using the similar arguments we also could show inequality (1) is true. Q.E.D.

Now we list some properties of the upper bound; details are available upon request:

1) The weight of the first term (initial guess term) in the upper bound decreases with \(k\); while the weight of the second term (integration term) increases with \(k\). Thus the first term describes the finite sample performance for early iterations, and the second term provides the long-run performance.

2) The coefficient in the first term (initial guess term)

\[
\exp\left(\frac{2\mu a((1 + A)^{1-\alpha} - (1 + A + k)^{1-\alpha})}{(1-\alpha)}\right)
\]

is an increasing function on \(\alpha\), a decreasing function on \(a\), and an increasing function on \(A\), which indicates we prefer smaller \(\alpha\), larger \(a\), and smaller \(A\) for better performance in early iterations. But, at the same time, we still need to concern about the stability of the algorithm, which can be achieved by larger \(A\) and smaller \(a\).

3) The values of \(b\) and \(\mu\) are based on the structure of loss function. In the upper bound, \(b\) only exists in the second term and \(\mu\) exists in both terms. For early iterations, if we pick \(\mu\) really close to the largest feasible value, then the upper bound is very tight in early iterations. But the value of \(b\) needs to be large enough for the whole process. On the contrary, if we pick \(\mu\) smaller, then the tightness of the upper bound would not be good for early iterations. But at the same time, we can pick smaller \(b\) for possible better tightness in later iterations.

In Corollary 1, we show that inequality (1) in Theorem 1 can be written in a simpler form by solving the integral, and the asymptotic performance of the algorithm can be seen more clearly.

**Corollary 1:** Inequality (1) can be written as

\[
E\|\hat{\theta}_k - \theta^*\|^2
\]

\[
\leq \exp\left(\frac{2\mu a(1 + A)^{1-\alpha} - 2\mu a(1 + A + k)^{1-\alpha}}{1-\alpha}\right) E\|\hat{\theta}_0 - \theta^*\|^2
\]

\[
+\exp\left(-\frac{2\mu a(1 + A + k + 1)^{1-\alpha}}{1-\alpha}\right) pba^2 C(\alpha)
\]

\[
\times \int_0^{k+1} (1 + A + x)^{-2\alpha} \exp\left(\frac{2\mu a(1 + A + x)^{1-\alpha}}{1-\alpha}\right) dx,
\]

where

\[
\frac{1}{1 + A + k}, \quad \alpha = 1,
\]

and

\[
0.5 < \alpha < 1,
\]

\[
\frac{(1 + A)^{2\mu a}}{(1 + A + k)^{2\mu a}} \left(E\|\hat{\theta}_0 - \theta^*\|^2 - T(k, \alpha)\right) + \frac{T(k, \alpha)}{(1 + A)^{2\mu a}}.
\]
\[
T(k, \alpha) = \frac{p ba^2 C(\alpha)}{2 \mu - \alpha / (1 + A + f(k))^{1-\alpha}},
\]
and \(f(k)\) is a function determined by the mean value theorem.

Remarks:

1. In \(T(k, \alpha)\), \(f(k)\to\infty\) as \(k\to\infty\). When \(0.5 < \alpha < 1\), we have \(T(k, \alpha)\to p ba^2 / 2 \mu a\) as \(k\to\infty\), which indicates the rate of convergence for the MSE is \(O(1/k^\alpha)\) in this case. When \(\alpha = 1\), if \(2 \mu a > 1\), then \(T(k, 1) > 0\), implying the rate of convergence is \(O(1/k)\); on the other hand if \(2 \mu a < 1\) then \(T(k, 1) < 0\), implying the rate of convergence is \(O(1/k^{2\mu a})\). Generally, we can pick \(a\) large enough to satisfy \(2 \mu a > 1\). In addition, when \(k\) is large, the effect of the value of \(A\) is disappearing in \(T(k, \alpha)\); while the effect of \(a\) is always there. The overall reasons that we introduce \(f(k)\) here are that we can rewrite the integration in terms of \(O(1/k^\alpha)\) and when \(k\to\infty\) the effect of \(f(k)\) disappears in the coefficient \(T(k, \alpha)\).

IV. NUMERICAL EXPERIMENTS

In this section, we do the numerical experiment to check the upper bound derived in the last section. The loss function we consider is skewed quartic function \(L(\theta) = \theta^T B^T f(\mathbf{B}) + 0.1 \sum_{i=1}^{p} (\mathbf{B})_i^2 + 0.01 \sum_{i=1}^{p} (\mathbf{B})_i^4\), where \((\cdot)_i\) represents the \(i\)th component of vector \(\mathbf{B}\), and \(p \mathbf{B}\) is an upper triangular matrix of 1’s. Here we consider \(p = 10\). The coefficients we pick are: \(a = a/(1 + A + k)^\alpha\), \(\alpha = 0.501\), \(A = 100\), \(a = 0.5\) and the value related to the assumptions (vi) and (vii) are approximated as \(\mu = 0.5\), \(b = 10\). We do the experiment for 50 replications with 2000 iterations in each replication. Figure 1 shows the numerical result with solid line representing \(E[\theta_k - \theta^*]^2\) and dash line representing the upper bound. We can see the mean square error is bounded by the finite sample upper bound. The significant decrease of the MSE error in early iterations can be partially captured by the first part of the upper bound by exponential rate, and the later performance of MSE error can be captured by the second part of the upper bound with polynomial rate.

Figure 1. Numerical result in comparison of the finite sample upper bound and the MSE error \(E[\theta_k - \theta^*]^2\)

V. CONCLUSION

The rate of convergence for DSPSA is considered in this paper, and the finite sample upper bounds are provided for both cases of \(0.5 < \alpha < 1\) and \(\alpha = 1\). The upper bounds are composed of two terms, the first term is corresponding to the performance for early iterations and the second term is corresponding to the long run performance. From one sample of numerical experiment, we can see the finite sample upper bound is quite reasonable. And in the future work, we will compare this result with the rate of convergence of other algorithms to check the overall performance of DSPSA.

REFERENCES