Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

A SMOOTHING STOCHASTIC QUASI-NEWTON METHOD FOR NON-LIPSCHITZIAN STOCHASTIC OPTIMIZATION PROBLEMS

Farzad Yousefian

School of Industrial Engineering
and Management
Oklahoma State University
Stillwater, OK 74078, USA

Angelia Nedić

School of Electrical, Computer,
and Energy Engineering
Arizona State University
Tempe, AZ 85287, USA

Uday V. Shanbhag

Industrial & Manufacturing Engineering
Pennsylvania State University
University Park, PA 16802, USA

ABSTRACT

Motivated by big data applications, we consider unconstrained stochastic optimization problems. Stochastic quasi-Newton methods have proved successful in addressing such problems. However, in both convex and non-convex regimes, most existing convergence theory requires the gradient mapping of the objective function to be Lipschitz continuous, a requirement that might not hold. To address this gap, we consider problems with not necessarily Lipschitzian gradients. Employing a local smoothing technique, we develop a smoothing stochastic quasi-Newton (S-SQN) method. Our main contributions are three-fold: (i) under suitable assumptions, we show that the sequence generated by the S-SQN scheme converges to the unique optimal solution of the smoothed problem almost surely; (ii) we derive an error bound in terms of the smoothed objective function values; and (iii) to quantify the solution quality, we derive a bound that relates the iterate generated by the S-SQN method to the optimal solution of the original problem.

1 INTRODUCTION

The problem of interest in this paper is an unconstrained stochastic optimization problem given as follows:

$$\min_{x\in\mathbb{R}^n} f(x) := \mathsf{E}[F(x,\xi(\omega))], \tag{SO}$$

where $F:\mathbb{R}^n\times\mathbb{R}^d\to\mathbb{R}$ is a function, the random vector ξ is given as $\xi:\Omega\to\mathbb{R}^d$, $(\Omega,\mathscr{F},\mathbb{P})$ denotes the associated probability space and the expectation $\mathsf{E}[F(x,\xi)]$ is taken with respect to $\mathbb{P}$. A wide range of big data applications arising from statistical learning and signal processing can be formulated as (SO). In these applications, a training sample $\{(a_i,b_i)_{i=1}^N\}$ is given comprising of input objects a_i and output objects b_i. The problem of interest is to learn a classifier $h(x,a)$ such that the empirical risk function of the form $\frac{1}{N}\sum_{i=1}^N \ell(h(x,a_i),b_i)$ is minimized, where ℓ is a loss function. In these problems, when the sample size N is large, the implementation of deterministic first order and second order methods becomes challenging. In contrast, stochastic approximation (SA) methods, first introduced by Robbins and Monro (Robbins and Monro 1951), have been widely used in addressing stochastic optimization (Nemirovski et al. 2009; Ghadimi and Lan 2012) and variational inequality problems (Juditsky, Nemirovski, and Tauvel 2011). In the classical SA method, the update rule is given by

$$x_{k+1} := x_k - \gamma_k \nabla F(x_k,\xi_k), \tag{SA}$$

where $\gamma_k > 0$ is the stepsize parameter, $\nabla F(x_k, \xi_k)$ is the sample of the stochastic gradient at x_k, and $k = 0, 1, \ldots$ is the iteration number. In the past few decades, there have been much interest in the development of efficient variants of SA schemes and their convergence analysis in addressing stochastic optimization and variational problems. While the convergence properties and rate statements of these scheme have been established in the literature, it has been observed that the performance of SA methods can be very sensitive to the problem properties, choice of the stepsize, and dataset characteristics. Motivated by the need to address some of these shortcomings, stochastic variants of quasi-Newton methods, for solving stochastic optimization problems have been developed in the past few years. In this class of methods, x_k is updated according to the following rule:

$$x_{k+1} := x_k - \gamma_k H_k \nabla F(x_k, \xi_k), \quad \text{for } k \geq 0, \tag{SQN}$$

where $H_k \succeq 0$ is an approximation of the inverse of Hessian at iteration k that incorporates the curvature information of the objective function within the scheme. The choice of the matrix H_k and the stepsize γ_k play a key role in establishing the convergence of SQN methods. In (Schraudolph, Yu, and Gunter 2007), the performance of SQN methods was studied numerically and was compared to that of SA schemes. Mokhtari et al. (Mokhtari and Ribeiro 2014) considered stochastic optimization problems with strongly convex objectives and developed a regularized BFGS method (RES) in that the matrix H_k is updated using a modified version of the classical BFGS update rule. To address large scale applications, limited memory variants of these scheme were developed to address problems with high dimentionality of the solution space (Mokhtari and Ribeiro 2015; Byrd et al. 2016). The extensions to non-convex regimes were studied in for example (Wang, Ma, and Liu 2017). Moreover, a variance reduced SQN method with a constant stepsize was developed (Lucchi, McWilliams, and Hofmann) addressing smooth strongly convex problems. Recently, we developed a regularized SQN method addressing problem (SO) in absence of strong convexity of the objective function (Yousefian, Nedić, and Shanbhag 2016b)

Motivation and summary of contributions: One of the main assumptions required to establish the convergence of the current SQN method, is the Lipschitzian property of the gradient mapping $\nabla F(x, \xi)$. For example see (Mokhtari and Ribeiro 2014; Mokhtari and Ribeiro 2015; Byrd et al. 2016; Wang, Ma, and Liu 2017). To the best of our knowledge, in absence of this assumption, neither convergence nor rate statements of the current SQN methods have been addressed in the literature. Motivated by this gap, in this paper, we consider the case where $\nabla F(x, \xi)$ is differentiable but non-Lipschitzian. Our goal lies in establishing the asymptotic convergence and also deriving the convergence rate statements. To this end, we employ a smoothing technique introduced by Steklov (Steklov 1907) and employed in stochastic optimization problems (Bertsekas 1973, Lakshmanan and Farias 2008, Duchi, Bartlett, and Wainwright 2012). Given a function $f : \mathbb{R}^n \to \mathbb{R}$ and a random variable z associcated with a probability distribution, the function $\hat{f}(x) := \mathsf{E}[f(x+z)]$ is considered a smoothed approximation of f. While the properties of $\hat{f}$ is well-studied in the literature, direct application of this technique in solving problem (SO) is challenging. This is because in stochastic regimes, the closed form of the function f, and consequently $\hat{f}$ is either unavailable or computationally expensive to be evaluated. To contend with this challenge, in our previous work, employing this local smoothing technique, we developed smoothing SA methods for solving both stochastic optimization problems (Yousefian, Nedić, and Shanbhag 2012, Yousefian, Nedić, and Shanbhag 2016a) and stochastic variational inequality problems (Yousefian, Nedić, and Shanbhag 2013, Yousefian, Nedić, and Shanbhag 2017a) in the absence of Lipschitzian property. In a similar vein, in this paper, we develop a smoothing stochastic quasi-Newton method, referred to as S-SQN method. The convergence and rate analysis of the S-SQN scheme in this paper is different than that of our earlier work on SA methods. This is mainly because in SQN methods, the presence of the stochastic matrix H_k introduces numerous challenges in the analysis of the underlying algorithm and a direct extension of the convergence analysis in SA schemes is not straightforward.

We summarize our contributions as follows: (i) under suitable assumptions on the stepsize γ_k, the Hessian approximation H_k, the stochastic noise, and boundedness of the iterate x_k, we show that the

sequence generated by the S-SQN scheme converges to the unique optimal solution of the smoothed problem almost surely; (ii) we then derive an error bound of the order $\mathscr{O}\left(\frac{1}{k}\right)$ in terms of the smoothed objective function values, where k is the iteration number; and (iii) to quantify the solution quality, under a local boundedness assumption of the Hessian, we derive a bound that relates the iterate generated by the S-SQN method to the optimal solution of the original problem (SO).

The reminder of the paper is organized as follows: In Section 2, we present an outline of the S-SQN scheme, discuss the underlying assumptions, and introduce the smoothing technique. The convergence analysis of the S-SQN scheme is provided in Section 3 in an almost sure sense. In Section 4, we derive the convergence rate of the generated iterate by the S-SQN method to the optimal solution of the original problem. Lastly, we outline the concluding remarks in Section 5.

Notation: Throughout this paper, a vector x is assumed to be a column vector and x^T denotes its transpose. $\|x\|$ denotes the Euclidean vector norm, i.e., $\|x\| = \sqrt{x^T x}$. We write *a.s.* as the abbreviation for "almost surely", and use $\mathsf{E}[z]$ to denote the expectation of a random variable z. The mapping F is Lipschitz continuous with parameter $L > 0$ if for any $x, y \in \mathbb{R}^n$, we have $\|F(x) - F(y)\| \leq L\|x - y\|$. For a given vector $x \in \mathbb{R}^n$ and scalar $\varepsilon > 0$, we use $B(x, \varepsilon)$ to denote an n-dimensional ball centered at x with radius $\varepsilon > 0$.

2 ALGORITHM OUTLINE

To address problem (SO) in absence of Lipschitzian property of the gradient mappings, we consider the following scheme: given an $x_0 \in \mathbb{R}^n$, let x_k be generated by the following recursive update rule:

$$x_{k+1} := x_k - \gamma_k H_k \nabla F(x_k + z_k, \xi_k), \quad \text{for all } k \geq 0, \tag{S-SQN}$$

where γ_k denotes the steplength sequence, $H_k \in \mathbb{R}^{n \times n}$ represents a matrix that captures the curvature information of the objective function, and $z_k \in \mathbb{R}^n$ is a uniform random variable drawn from a ball centered at the origin with radius $\varepsilon > 0$, i.e., $z_k \in B(0, \varepsilon)$. An immediate distinction between the standard SQN scheme and (S-SQN) is the presence of random vector z_k. At iteration k, the stochastic gradient $\nabla F(., \xi_k)$ is evaluated, not at x_k, but at the perturbed vector $x_k + z_k$. We will show that under this modification, and under suitable assumptions, we can establish the convergence properties of the scheme in absence of Lipschitzian property. Throughout, we let $\mathscr{F}_k$ denote the history of the method up to time k, i.e.,

$$\mathscr{F}_k = \{x_0, \xi_0, z_0, \xi_1, z_1, \ldots, \xi_{k-1}, z_{k-1}\}, \quad \text{for } k \geq 1,$$

and $\mathscr{F}_0 = \{x_0\}$. Next, we state the main assumptions in our work. In the results in this paper, whenever needed, we may refer to all or a subset of these assumptions.

Assumption 1 (Differentiability) The function $F(x, \xi)$ is continuously differentiable for all $x \in \mathbb{R}^n$ and $\xi \in \Omega$.

The following assumption imposes boundedness of the gradient mapping over $\mathbb{R}^n$.

Assumption 2 (Boundedness of gradients) There exists a scalar C such that for all $x \in \mathbb{R}^n$, we have $\mathsf{E}\left[\|\nabla F(x, \xi)\|^2\right] \leq C^2$.

It is important to note that Assumption 2 may hold for some merely convex or even non-convex functions F, but it does not hold when F is strongly convex. This can be seen since for a strongly convex function F, we can write

$$\|\nabla F(x, \xi) - \nabla F(y, \xi)\| \geq \mu\|x - y\|, \quad \text{for all } x, y \in \mathbb{R}^n, \xi \in \Omega.$$

where $\mu > 0$ is the strong convexity parameter. By setting $y = 0$, it can be seen that $\|\nabla F(x, \xi)\|$ will become unbounded over $\mathbb{R}^n$ when $x \in \mathbb{R}^n$. In some parts of our analysis where we impose strong convexity assumption, we consider a weaker version of Assumption 2 as follows:

Assumption 2′ For any scalar $M > 0$ such that $\|x\| < M$, there exists $C > 0$ such that $\mathsf{E}\left[\|\nabla F(x,\xi)\|^2\right] \leq C^2$.

The following two assumptions regulate the standard requirements on the inherent uncertainty characterized by ξ, as well as the properties of the smoothing random variable z.

Assumption 3 (Random variable ξ) (a) Random variables ξ_k are i.i.d. for any $k \geq 0$;
(b) The stochastic gradient $\nabla F(x,\xi)$ is an unbiased estimator of $\nabla f(x)$, i.e. $\mathsf{E}[\nabla F(x,\xi)] = \nabla f(x)$;

Assumption 4 (Random variable z) (a) Random variables $z_k \in \mathbb{R}^n$ are i.i.d. and independent of random variables ξ_k. Additionally, z_k is uniformly distributed over $B(0,\varepsilon)$, a ball centered at the origin with a radius $\varepsilon > 0$.

Next we consider general assumptions on the structure and properties of the matrix H_k. These assumptions are standard requirements to establish the convergence of the SQN method. For example, see (Byrd et al. 2016).

Assumption 5 (Conditions on matrix H_k) Let the following hold for any $k \geq 0$:
(a) Matrix H_k is $\mathscr{F}_k$-measurable, i.e., $\mathsf{E}[H_k \mid \mathscr{F}_k] = H_k$.
(b) Matrix $H_k \in \mathbb{R}^{n\times n}$ is symmetric and satisfies the following condition: There exist positive scalars $\lambda_{\min}$ and $\lambda_{\max}$ such that $\lambda_{\min}\mathbf{I} \preceq H_k \preceq \lambda_{\max}\mathbf{I}$, for all $k \geq 0$.

One natural research question lies in the development of an update rule for H_k that satisfies Assumption 5. This indeed can be done for example through a modification of the existing limited memory stochastic BFGS update rules (e.g., (Mokhtari and Ribeiro 2015)) even in absence of strong convexity or Lipschitzian property. However, the design of an update rule for H_k that satisfies Assumption 5 is beyond the scope of our work in this paper and remains as a future research direction to our work. Next, we state the requirements on the stepsize sequence.

Assumption 6 The stepsize is such that $\gamma_k > 0$ for all k, $\sum_{k=0}^{\infty}\gamma_k = \infty$, and $\sum_{k=0}^{\infty}\gamma_k^2 < \infty$.

As mentioned in Section 1, the intuition behind the (S-SQN) method is employment of a local smoothing technique within the standard SQN scheme. To this end, we introduce the smoothing technique by defining a smoothed function in the following:

Definition 1 (**Smoothed function**) Consider function $f : \mathbb{R}^n \to \mathbb{R}^n$. Let $z \in \mathbb{R}^n$ be uniformly distributed in $B(0,\varepsilon)$. The smoothed (approximate) function $f_\varepsilon : \mathbb{R}^n \to \mathbb{R}$ is defined by $f_\varepsilon(x) = \mathsf{E}[f(x+z)]$.

The function f^ε is characterized by the random variable z and the parameter ε. Throughout, we refer to ε as the smoothing parameter. Note that the probability density function of the uniform random variable z is given as $p_u(z) = \frac{1}{c_n\varepsilon^n}$, for $z \in B(0,\varepsilon)$ and 0 otherwise, where c_n is the volume of the unit ball in $\mathbb{R}^n$, i.e., $c_n = \int_{B(0,1)} dy = \dfrac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)}$, and Γ is the gamma function. Next, we present the main properties of the smoothing technique used in the convergence analsis of the (S-SQN) scheme including the strong convexity of f_ε and the Lipschitz continuity of $\nabla_x f_\varepsilon$. The following result is an extension of Lemma 8 in (Yousefian, Nedić, and Shanbhag 2012).

Lemma 1 (**Properties of smoothed function**) Consider the smoothed function f_ε prescribed in Definition 1. Let Assumption 1 hold. Then, we may claim the following:

(a) The function $f_\varepsilon : \mathbb{R}^n \to \mathbb{R}$ is differentiable with gradients $\nabla f_\varepsilon(x) = \mathsf{E}[\nabla f(x+z)] = \mathsf{E}[\nabla F(x+z,\xi)]$.
(b) For all $x, y \in \mathbb{R}^n$ we have

$$\|\nabla f_\varepsilon(x) - \nabla f_\varepsilon(y)\| \leq \left(\sup_{v\in B(x,\varepsilon)\cup B(y,\varepsilon)} \|\nabla f(v)\|\right)\frac{\kappa n!!}{(n-1)!!\varepsilon}\|x-y\|, \tag{1}$$

where $\kappa = 1$ if n is odd and $\kappa = \frac{2}{\pi}$ otherwise.

(c) Let Assumption 2 hold. Then, the gradient mapping ∇f_ε is Lipschitz continuous over $\mathbb{R}^n$ with the parameter $\kappa \frac{n!!}{(n-1)!!}\frac{C}{\varepsilon}$, i.e.,

$$\|\nabla f_\varepsilon(x) - \nabla f_\varepsilon(y)\| \leq \frac{\kappa n!! C}{(n-1)!!\varepsilon}\|x-y\|, \quad \text{for all } x,y \in \mathbb{R}^n. \tag{2}$$

(d) Let function f be strongly convex with parameter $\mu > 0$ over $\mathbb{R}^n$. Then, f_ε is also strongly convex with parameter $\mu > 0$ over $\mathbb{R}^n$.

Proof. (a) First we show that under Assumption 1, $\nabla f(x)$ exists and $\nabla f(x) = \mathsf{E}[\nabla F(x,\xi)]$. Note that since F is differentiable for all $x \in \mathbb{R}^n$ and $\xi \in \Omega$, we have $\frac{\partial F(x,\xi)}{\partial x_j} = \lim_{h\to 0}\frac{F(x+e_j h,\xi)-F(x,\xi)}{h}$ for all $j = 1,\ldots,n$, where e_j is a column vector in $\mathbb{R}^n$ with the jth element equal to 1, and all other elements equal to 0. Taking expectations on both sides of the preceding relation, we obtain

$$\mathsf{E}\left[\frac{\partial F(x,\xi)}{\partial x_j}\right] = \mathsf{E}\left[\lim_{h\to 0}\frac{F(x+e_j h,\xi)-F(x,\xi)}{h}\right] = \lim_{h\to 0}\frac{\mathsf{E}[F(x+e_j h,\xi)]-\mathsf{E}[F(x,\xi)]}{h} = \mathsf{E}\left[\frac{\partial f(x)}{\partial x_j}\right],$$

where in the second equation, since F is assumed to be differentiable implying that $\lim_{h\to 0}\frac{F(x+e_j h,\xi)-F(x,\xi)}{h}$ exists and is bounded, we may apply the Lebesgue's dominated convergence theorem. Therefore, we have $\nabla f(x) = \mathsf{E}[\nabla F(x,\xi)]$. In a similar fashion, using the definition of f_ε, it can be shown that $\nabla f_\varepsilon(x) = \mathsf{E}[\nabla f(x+z)]$. Combining this with the previous result, we conclude that the statement of part (a) holds.
(b) From part (a), we may express $\|\nabla f_\varepsilon(x) - \nabla f_\varepsilon(y)\|$ as follows

$$\|\nabla f_\varepsilon(x) - \nabla f_\varepsilon(y)\| = \left\| \int_{\mathbb{R}^n} \nabla f(x+z) p_u(z) dz - \int_{\mathbb{R}^n} \nabla f(y+z) p_u(z) dz \right\|.$$

By a change of the integral variable in the preceding relation, it follows that

$$\begin{aligned}
&\|\nabla f_\varepsilon(x) - \nabla f_\varepsilon(y)\| = \left\| \int_{\mathbb{R}^n} (p_u(v-x) - p_u(v-y))\nabla f(v) dv \right\| \\
&= \left\| \int_{B(x,\varepsilon)\cup B(y,\varepsilon)} (p_u(v-x) - p_u(v-y))\nabla f(v) dv \right\| \leq \int_{B(x,\varepsilon)\cup B(y,\varepsilon)} |p_u(v-x) - p_u(v-y)| \|\nabla f(v)\| dv \\
&\leq \sup_{v\in B(x,\varepsilon)\cup B(y,\varepsilon)} \|\nabla f(v)\| \int_{B(x,\varepsilon)\cup B(y,\varepsilon)} |p_u(v-x) - p_u(v-y)| dv,
\end{aligned} \tag{3}$$

where the first inequality follows from Jensen's inequality and the second inequality is an implication of the boundedness of the mapping ∇f over $\mathbb{R}^n$. The remainder of the proof is similar to that of (Yousefian, Nedić, and Shanbhag 2012, Lemma 8).
(c) The proof this part follows from the result of part (b) and Assumption 2.
(d) The proof of this statement follows directly by the definition of strong convexity and the definition of function f_ε. □

3 CONVERGENCE ANALYSIS

In this section, we establish the asymptotic convergence of the (S-SQN) method. To this end, in Lemma 4, we derive a recursive relation on the error bound of the scheme. Then, in Theorem 1, we present the convergence properties of the scheme in an almost sure sense. In our analysis, we use the following definition for referring to the stochastic errors of the gradient mapping ∇F:

$$w_k := \nabla F(x_k + z_k, \xi_k) - \nabla f(x_k + z_k), \quad \text{for all } k \geq 0. \tag{4}$$

The following result, is used in the analysis of the scheme. It states the conditional expectation of the stochastic error w_k is zero. This indeed is a consequence of the Assumptions 3 and 4.

Lemma 2 (Conditional first moment of w_k) Consider the (S-SQN) scheme and suppose Assumptions 1, 3, and 4 hold. Then, for any $k \geq 0$ we have $\mathsf{E}[w_k \mid \mathscr{F}_k \cup \{z_k\}] = 0$.

Proof. Let $k \geq 0$ be a a fixed integer. The definition of w_k in (4) and Assumption 3(b) imply that

$$\mathsf{E}[w_k \mid \mathscr{F}_k \cup \{z_k\}] = \mathsf{E}[\nabla F(x_k + z_k, \xi_k) \mid \mathscr{F}_k \cup \{z_k\}] - \nabla f(x_k + z_k) = \nabla f(x_k + z_k) - \nabla f(x_k + z_k) = 0,$$

where we employ the independence of z_k between ξ_k and by recalling that z_k and ξ_k are both i.i.d. random variables. □

We use the following Lemma in establishing the convergence of (S-SQN) method (see (Polyak 1987), page 50).

Lemma 3 (Robbins-Siegmund) Let v_k, u_k, α_k, and β_k be nonnegative random variables, and let the following relations hold almost surely:

$$\mathsf{E}\left[v_{k+1} \mid \tilde{\mathscr{F}}_k\right] \leq (1+\alpha_k)v_k - u_k + \beta_k \quad \text{for all } k, \qquad \sum_{k=0}^{\infty} \alpha_k < \infty, \qquad \sum_{k=0}^{\infty} \beta_k < \infty,$$

where $\tilde{\mathscr{F}}_k$ denotes the collection $v_0, \ldots, v_k$, $u_0, \ldots, u_k$, $\alpha_0, \ldots, \alpha_k$, $\beta_0, \ldots, \beta_k$. Then, the following holds

$$\lim_{k\to\infty} v_k = v, \qquad \sum_{k=0}^{\infty} u_k < \infty \qquad \text{a.s.},$$

where $v \geq 0$ is a random variable.

Next, we derive a recursive relation for the smoothed objective function value f_ε. This relation is a key in establishing the convergence and rate analysis of the developed (S-SQN) method.

Lemma 4 [**A recursive error bound**] Consider the (S-SQN) scheme. Let Assumptions 1, 3, 4, and 5 hold.

(a) Let us define $\theta_{\varepsilon,k} := \sup_{v \in [x_k, x_{k+1}] + B(0,\varepsilon)} \|\nabla f(v)\|$. The following inequality holds:

$$\mathsf{E}[f_\varepsilon(x_{k+1}) \mid \mathscr{F}_k] \leq f_\varepsilon(x_k) - \lambda_{\min}\gamma_k \|\nabla f_\varepsilon(x_k)\|^2 + \frac{\kappa \lambda_{\max}^2\, n!!}{2\varepsilon(n-1)!!}\gamma_k^2 \mathsf{E}\left[\theta_{\varepsilon,k}\|\nabla F(x_k + z_k, \xi_k)\|^2 \mid \mathscr{F}_k\right]. \tag{5}$$

(b) Suppose Assumption 2 holds in addition. Then,

$$\mathsf{E}[f_\varepsilon(x_{k+1}) \mid \mathscr{F}_k] \leq f_\varepsilon(x_k) - \lambda_{\min}\gamma_k \|\nabla f_\varepsilon(x_k)\|^2 + \frac{\kappa C^3 \lambda_{\max}^2\, n!!}{2\varepsilon(n-1)!!}\gamma_k^2. \tag{6}$$

(c) Let $f(x)$ be a strongly convex function with parameter μ and $f_\varepsilon^* \triangleq \min_{x\in\mathbb{R}^n} f_\varepsilon(x)$. We have

$$\begin{aligned}\mathsf{E}[f_\varepsilon(x_{k+1}) - f_\varepsilon^* \mid \mathscr{F}_k] \leq{}& (1 - 2\mu\lambda_{\min}\gamma_k)\left(f_\varepsilon(x_k) - f_\varepsilon{}^*\right) \\ &+ \frac{\kappa \lambda_{\max}^2\, n!!}{2\varepsilon(n-1)!!}\gamma_k^2 \sup_{v \in B(x_k,\varepsilon)} \mathsf{E}\left[\theta_{\varepsilon,k}\|\nabla F(v, \xi_k)\|^2 \mid \mathscr{F}_k\right].\end{aligned} \tag{7}$$

Proof. (a) From Lemma 1(b), we have

$$\|\nabla f_\varepsilon(x) - \nabla f_\varepsilon(y)\| \leq \left(\sup_{v \in B(x,\varepsilon)\cup B(y,\varepsilon)} \|\nabla f(v)\|\right) \frac{\kappa n!!}{(n-1)!!\varepsilon}\|x - y\|, \quad \text{for all } x, y \in \mathbb{R}^n. \tag{8}$$

Let us define function $g:[0,1]\to\mathbb{R}$ as $g(t)\triangleq f_\varepsilon(y+t(x-y))$. Note that $g(0)=f_\varepsilon(y)$, $g(1)=f_\varepsilon(x)$, and $\nabla g(0)=\nabla f_\varepsilon(y+t(x-y))^T(x-y)$. Note that $\int_0^1\nabla g(t)dt=g(1)-g(0)$. This implies that

$$f_\varepsilon(x)-f_\varepsilon(y)-\nabla f_\varepsilon(y)^T(x-y)=\int_0^1\nabla f_\varepsilon(y+t(x-y))^T(x-y)dt-\nabla f_\varepsilon(y)^T(x-y)$$
$$=\int_0^1\Big(\nabla f_\varepsilon(y+t(x-y))-\nabla f_\varepsilon(y)\Big)^T(x-y)dt\leq\|x-y\|\int_0^1\Big\|\nabla f_\varepsilon(y+t(x-y))-\nabla f_\varepsilon(y)(x-y)\Big\|dt,$$

where the last inequality follows by the Cauchy-Schwarz inequality. Applying relation (8), we obtain

$$f_\varepsilon(x)-f_\varepsilon(y)-\nabla f_\varepsilon(y)^T(x-y)\leq\frac{\kappa n!!}{(n-1)!!\varepsilon}\|x-y\|^2\int_0^1\sup_{v\in B(y+t(x-y),\varepsilon)\cup B(y,\varepsilon)}\|\nabla f(v)\|tdt$$
$$\leq\frac{\kappa n!!}{(n-1)!!\varepsilon}\|x-y\|^2\int_0^1\sup_{v\in\cup_{\alpha\in[0,1]}B(y+\alpha(x-y),\varepsilon)}\|\nabla f(v)\|tdt,$$

where the last inequality follows since $B(y+t(x-y),\varepsilon)\cup B(y,\varepsilon)\subset\cup_{\alpha\in[0,1]}B(y+\alpha(x-y),\varepsilon)$ holds for any $t\in[0,1]$. Note that the set $\cup_{\alpha\in[0,1]}B(y+\alpha(x-y),\varepsilon)$ can be written as $[x,y]+B(0,\varepsilon)=[x,y]+B(0,\varepsilon)$, where the addition is in the Minkowski sense. Therefore, we have that

$$f_\varepsilon(x)-f_\varepsilon(y)-\nabla f_\varepsilon(y)^T(x-y)\leq\left(\sup_{v\in[x,y]+B(0,\varepsilon)}\|\nabla f(v)\|\right)\frac{\kappa n!!}{2(n-1)!!\varepsilon}\|x-y\|^2,\ \text{for all } x,y\in\mathbb{R}^n.$$

Let x_k be generated by the (S-SQN) recursion. Substituting $x=x_{k+1}$ and $y=x_k$ in the preceding relation, we have

$$f_\varepsilon(x_{k+1})\leq f_\varepsilon(x_k)-\gamma_k\nabla f_\varepsilon(x_k)^TH_k\nabla F(x_k+z_k,\xi_k)+\frac{\kappa\theta_{\varepsilon,k}\ n!!}{2\varepsilon(n-1)!!}\Big\|-\gamma_kH_k\nabla F(x_k+z_k,\xi_k)\Big\|^2,$$

where $\theta_{\varepsilon,k}\triangleq\sup_{v\in[x_k,x_{k+1}]+B(0,\varepsilon)}\|\nabla f(v)\|$. Next, from the definition of w_k in (4), and that H_k is symmetric, we can write

$$f_\varepsilon(x_{k+1})\leq f_\varepsilon(x_k)-\gamma_k\nabla f_\varepsilon(x_k)^TH_k\left(\nabla f(x_k+z_k)+w_k\right)+\frac{\kappa\theta_{\varepsilon,k}\ n!!}{2\varepsilon(n-1)!!}\gamma_k^2\Big\|H_k\nabla F(x_k+z_k,\xi_k)\Big\|^2$$
$$=f_\varepsilon(x_k)-\gamma_k\nabla f_\varepsilon(x_k)^TH_k\nabla f(x_k+z_k)-\gamma_k\nabla f_\varepsilon(x_k)^TH_kw_k$$
$$+\frac{\kappa\theta_{\varepsilon,k}\ n!!}{2\varepsilon(n-1)!!}\gamma_k^2\nabla F(x_k+z_k,\xi_k)^TH_k^2\nabla F(x_k+z_k,\xi_k).\tag{9}$$

Recall that for any positive definite matrix A with bounded eigenvalues, we have $\lambda_{\min}(A)\|x\|^2\leq x^TAx\leq\lambda_{\max}(A)\|x\|^2$ for all $x\in\mathbb{R}^n$. This implies that

$$\nabla F(x_k+z_k,\xi_k)^TH_k^2\nabla F(x_k+z_k,\xi_k)\leq\lambda_{\max}^2\|\nabla F(x_k+z_k,\xi_k)\|^2.\tag{10}$$

Therefore, by taking expectations conditioned on $\mathscr{F}_k\cup z_k$ from the relation (9), and taking into account that f_ε, x_k, and H_k are $\mathscr{F}_k\cup z_k$-measurable, we obtain

$$\mathsf{E}[f_\varepsilon(x_{k+1})\mid\mathscr{F}_k\cup z_k]\leq f_\varepsilon(x_k)-\gamma_k\nabla f_\varepsilon(x_k)^TH_k\nabla f(x_k+z_k)-\gamma_k\nabla f_\varepsilon(x_k)^TH_k\mathsf{E}[w_k\mid\mathscr{F}_k\cup z_k]$$
$$+\frac{\kappa\ n!!}{2\varepsilon(n-1)!!}\lambda_{\max}^2\gamma_k^2\mathsf{E}\left[\theta_{\varepsilon,k}\|\nabla F(x_k+z_k,\xi_k)\|^2\mid\mathscr{F}_k\cup z_k\right].$$

Invoking the result of Lemma 2, we have $\mathsf{E}[w_k \mid \mathscr{F}_k \cup z_k] = 0$. Taking this into account, by taking expectations with respect to z_k on the preceding inequality, we have

$$\mathsf{E}[f_\varepsilon(x_{k+1}) \mid \mathscr{F}_k] \leq f_\varepsilon(x_k) - \gamma_k \nabla f_\varepsilon(x_k)^T H_k f_\varepsilon(x_k) + \frac{\kappa\, n!!}{2\varepsilon(n-1)!!}\lambda_{\max}^2\gamma_k^2 \mathsf{E}\left[\theta_{\varepsilon,k}\|\nabla F(x_k+z_k,\xi_k)\|^2 \mid \mathscr{F}_k\right],$$

where we invoke the definition of the smoothed function, i.e., $\mathsf{E}[f(x_k+z_k)|\mathscr{F}_k] = f_\varepsilon(x_k)$. Using the inequality $\nabla f_\varepsilon(x_k)^T H_k \nabla f_\varepsilon(x_k) \geq \lambda_{\min}\|\nabla f_\varepsilon(x_k)\|^2$ we obtain the (5).
(b) To show relation (6), note that using Jensen's inequality and Assumption 2, for $v \in \mathbb{R}^n$ we have

$$\|\nabla f(v)\| = \sqrt{\|\nabla f(v)\|^2} = \sqrt{\|\mathsf{E}[\nabla F(v,\xi)]\,\|^2} \leq \sqrt{\|\mathsf{E}[\nabla F(v,\xi)\|^2]} \leq \sqrt{C^2} = C.$$

This implies that $\theta_{\varepsilon,k} \leq C$ for all $k \geq 0$. Moreover, we have

$$\mathsf{E}\left[\|\nabla F(x_k+z_k,\xi_k)\|^2 \mid \mathscr{F}_k\right] \leq \sup_{v\in B(x_k,\varepsilon)} \mathsf{E}\left[\|\nabla F(v,\xi_k)\|^2 \mid \mathscr{F}_k\right] \leq \sup_{v\in B(x_k,\varepsilon)} C^2 = C^2.$$

From the preceding two inequalities and relation (5), the inequality (6) follows.
(c) Note that from Lemma 1(d), function f_ε is strongly convex with parameter μ. Therefore, from Theorem 2.3.3 of (Facchinei and Pang 2003), problem $\min_{x\in\mathbb{R}^n} f_\varepsilon(x)$ has a unique optimal solution x_ε^*. Recall that as a property of strongly convex functions, we have $\|\nabla f_\varepsilon(x)\|^2 \geq 2\mu(f_\varepsilon(x) - f_\varepsilon^*)$ for any $x \in \mathbb{R}^n$. Therefore, from (5) we obtain

$$\mathsf{E}[f_\varepsilon(x_{k+1}) \mid \mathscr{F}_k] \leq f_\varepsilon(x_k) - 2\lambda_{\min}\gamma_k\mu(f_\varepsilon(x) - f_\varepsilon^*) + \frac{\kappa\lambda_{\max}^2\, n!!}{2\varepsilon(n-1)!!}\gamma_k^2 \sup_{\|z\|\leq\varepsilon} \mathsf{E}\left[\theta_{\varepsilon,k}\|\nabla F(x_k+z,\xi_k)\|^2 \mid \mathscr{F}_k\right],$$

where we used the definition of random variable z_k in the last term of the preceding inequality. Subtracting f_ε^* from both sides and taking to account that x_k is $\mathscr{F}_k$-measurable, we obtain the desired relation. □

The next result establishes convergence of the proposed scheme in an almost sure sense.

Theorem 1 [**Almost sure convergence**] Consider the sequence x_k generated by the (S-SQN) scheme. Let Assumptions 1, 3, 4, 5, and 6 hold. Then, we have the following results:

(a) If Assumption 2 holds, then almost surely $\liminf_{k\to\infty} \nabla f_\varepsilon(x_k) = 0$, where $\nabla f_\varepsilon(x)$ is the smoothed gradient mapping.
(b) Let Assumption **2′** hold and f be strongly convex with parameter $\mu > 0$. Then, problem $\min_{x\in\mathbb{R}^n} f_\varepsilon(x)$ has a unique optimal solution denoted by $x_\varepsilon{}^*$, and the following statements are equivalent:
 (i) The sequence $\{x_k\}$ is bounded almost surely.
 (ii) The sequence $\{x_k\}$ converges to the unique optimal solution $x_\varepsilon{}^*$ almost surely.

Proof. (a) Note that Lemma 4(b) holds. To show (a), we apply Lemma 3. Let us define

$$v_k := f_\varepsilon(x_k), \quad \alpha_k = 0, \quad u_k = \lambda_{\min}\gamma_k\|\nabla f_\varepsilon(x_k)\|^2, \quad \beta_k = \frac{\kappa C\, n!!}{2\varepsilon(n-1)!!}\lambda_{\max}^2 C^2\gamma_k^2.$$

Note that the preceding defined sequences are nonnegative and $\sum_{k=0}^\infty \alpha_k = 0 < \infty$. Assumption 6 implies that $\sum_{k=0}^\infty \beta_k < \infty$. Therefore, from Lemma 3 and Lemma 4(b), we conclude that almost surely $\lim_{k\to\infty} f_\varepsilon(x_k)$ exists, and that $\sum_{k=0}^\infty \gamma_k\|\nabla f_\varepsilon(x_k)\|^2 < \infty$. As a consequence of the latter statement, and that $\sum_{k=0}^\infty \gamma_k = \infty$, we have $\liminf_{k\to\infty} \|\nabla f_\varepsilon(x_k)\| = 0$.
(b) The uniqueness of x_ε^* follows by strong convexity property of f_ε and Theorem 2.3.3 of (Facchinei and Pang 2003). Note that Lemma 4(c) holds. Let us also assume (i) holds. Since x_k is bounded, from

Assumption **2′**, there exists a constant $C > 0$ such that $\theta_{\varepsilon,k} \leq C$ and $\sup_{v \in B(x_k,\varepsilon)} \mathsf{E}\left[\|\nabla F(v,\xi_k)\|^2 \mid \mathscr{F}_k\right] \leq C^2$ for all k. In a similar fashion to the proof of part (a), invoking Lemma 3 and Lemma 4(c), we can conclude that x_k goes to x_ε^* almost surely. Now suppose statement (ii) holds. Therefore, x_k is a convergent sequence a.s., implying that $\{x_k\}$ is bounded a.s., i.e., statement (i) holds. Therefore, statements (i) and (ii) are equivalent. □

4 RATE ANALYSIS

The result of Theorem 1 provides asymptotic convergence properties of the (S-SQN) method. A natural question is how fast does the iterate x_k converge to the approximate optimal solution x_ε^* in some probabilistic sense. Moreover, can we derive a bound on the expected error between x_k and the optimal solution of the original problem (SO)? In this section, our goal lies in addressing these two questions. First, in the following result, we provide a bound on the error $\|x_\varepsilon^* - x^*\|$ under a strong convexity assumption of the objective function f.

Proposition 1 [**Solution quality of the smoothed problem**] Let Assumption 1 hold and f be strongly convex on $\mathbb{R}^n$ with a constant $\mu > 0$.

(a) Then, we have $\|x_\varepsilon^* - x^*\| \leq \dfrac{\sup_{\|z\| \leq \varepsilon} \|\nabla f(x^* + z)\|}{\mu}$, where x^* and x_ε^* denote the unique optimal solutions to problem (SO) and the smoothed problem $\min_{x \in \mathbb{R}^n} f_\varepsilon(x)$, respectively.

(b) Let f be twice continuously differentiable over $\mathbb{R}^n$. Suppose there exists a neighborhood of x^* in which f has a bounded Hessian. Let b_H denote a bound on the maximum eigenvalue of the Hessian matrix in that neighborhood. Then, for a sufficiently small ε we have: $\|x_\varepsilon{}^* - x^*\| \leq \frac{b_H \varepsilon}{\mu}$.

Proof. (a) The existence and uniqueness of the optimal solution to $\min_{x \in \mathbb{R}^n} f(x)$, as well as $\min_{x \in \mathbb{R}^n} f_\varepsilon(x)$, is guaranteed by Theorem 2.3.3 of (Facchinei and Pang 2003). Note that by the optimality conditions, we have $\nabla f(x^*) = \nabla f_\varepsilon(x_\varepsilon^*) = \mathbf{0}$. Using strong convexity of f_ε implied by Lemma 1(d), we have

$$(\nabla f_\varepsilon(x^*) - \mathbf{0})^T (x^* - x_\varepsilon^*) \geq \mu \|x^* - x_\varepsilon^*\|^2.$$

By invoking the Cauchy-Schwarz inequality, we obtain

$$\mu \|x^* - x_\varepsilon^*\| \leq \|\nabla f_\varepsilon(x^*)\|. \tag{11}$$

It suffices to show that, $\|\nabla f_\varepsilon(x^*)\| \leq \sup_{\|z\| \leq \varepsilon} \|\nabla f(x^* + z)\|$. From Lemma 1(a) we can write

$$\|\nabla f_\varepsilon(x^*)\| = \left\| \int_{\|z\| \leq \varepsilon} \nabla f(x^* + z) p_u(z) dz \right\| \leq \int_{\|z\| \leq \varepsilon} \sup_{\|z\| \leq \varepsilon} \|\nabla f(x^* + z)\| p_u(z) dz \leq \sup_{\|z\| \leq \varepsilon} \|\nabla f(x^* + z)\|, \tag{12}$$

where the first inequality is implied using the Jensen's inequality and the convexity of the norm. Therefore, relations (11) and (12) imply the desired result.
(b) By assumption, there exists a $\rho > 0$ where $\|\nabla^2 f(x)\| \leq b_H$ for any $x \in B(x^*, \rho)$, where $B(x^*, \rho)$ denotes an n-dimensional ball centered at x^* with radius ρ. Let $\delta \in \mathbb{R}^n$. Using the mean value theorem,

$$\nabla f(x^* + \delta) - \nabla f(x^*) = \left(\int_0^1 \nabla^2 f(x + t\delta) dt \right) \delta, \quad \text{for all } \delta \in B(0, \rho). \tag{13}$$

Assume that ε is small enough such that $\varepsilon < \rho$. From (13) we obtain $\|\nabla f(x^* + z) - \nabla f(x^*)\| \leq b_H \|z\| \leq b_H \varepsilon$. The desired result follows from the preceding relation and the inequality of part (a). □

In deriving the convergence rate result, we make use of the following Lemma. The proof can be found in (Yousefian, Nedić, and Shanbhag 2017b).

Lemma 5 (**Convergence rate of a recursive sequence**) Let $\{e_k\}$ be a non-negative sequence such that for an arbitrary non-negative sequence $\{\gamma_k\}$, we have $e_{k+1} \leq (1-\alpha\gamma_k)e_k + \beta\gamma_k^2$, for all $k \geq 0$, where α and β are positive scalars. Suppose $\gamma_0 = \frac{2}{\alpha}$, $\gamma_k = \frac{\gamma_0}{k}$ for any $k \geq 1$, where $\gamma > \frac{1}{\alpha}$. Then, for all $k \geq 2$ we have $e_k \leq \frac{8\beta}{\alpha^2 k}$.

Next, we provide the rate statements of the developed (S-SQN) scheme.

Theorem 2 (**Rate statements**) Consider the sequence x_k generated by the (S-SQN) scheme. Let Assumptions 1, **2′**, 3, 4, and 5 hold. Let function f be strongly convex with parameter $\mu > 0$, and the stepsize γ_k be given by $\gamma_0 = \frac{1}{\mu\lambda_{\min}}$, and $\gamma_k = \frac{\gamma_0}{k}$ for all $k \geq 1$. Let the sequence $\{x_k\}$ be bounded almost surely. Then, there exist $C > 0$ and $\theta > 0$ such that

$$\mathsf{E}[f_\varepsilon(x_k) - f_\varepsilon^*] \leq \left(\frac{\kappa C^3 \lambda_{\max}\ n!!}{\lambda_{\min}^2 \mu^2 \varepsilon\ (n-1)!!}\right)\frac{1}{k}, \quad \text{for all } k \geq 2. \tag{14}$$

Moreover, suppose there exists a neighborhood of x^* in which f has a bounded Hessian. Let b_H denote a bound on the maximum eigenvalue of the Hessian matrix in that neighborhood. Then, for a sufficiently small ε we have:

$$\mathsf{E}\left[\|x_k - x^*\|^2\right] \leq 4\frac{\kappa C^3 \lambda_{\max}\ n!!}{\lambda_{\min}^2 \mu^3 \varepsilon\ (n-1)!!k} + \frac{2b_H^2\varepsilon^2}{\mu^2}, \quad \text{for all } k \geq 2. \tag{15}$$

Proof. Since x_k is assumed to be bounded almost surely, from Assumption **2′**, there exists a constant $C > 0$ such that $\theta_{k,\varepsilon} \leq C$ and $\sup_{v \in B(x_k,\varepsilon)} \mathsf{E}\left[\|\nabla F(v,\xi_k)\|^2 \mid \mathscr{F}_k\right] \leq C^2$. Therefore, from Lemma 4(c) we have

$$\mathsf{E}[f_\varepsilon(x_{k+1}) - f_\varepsilon^*] \leq (1 - 2\mu\lambda_{\min}\gamma_k)\,\mathsf{E}[f_\varepsilon(x_k) - f_\varepsilon{}^*] + \frac{\kappa C \lambda_{\max}^2\ n!!C^2}{2\varepsilon(n-1)!!}\gamma_k^2.$$

Let us define the terms $e_k := \mathsf{E}[f_\varepsilon(x_k) - f_\varepsilon{}^*]$, $\alpha := 2\mu\lambda_{\min}$, and $\beta := \frac{\kappa\lambda_{\max}^2\ n!!C^3}{2\varepsilon(n-1)!!}$. Applying Lemma 5, we conclude that relation (14) holds. To show (15), note that from Proposition 1, we have $\|x_\varepsilon^* - x^*\|^2 \leq \frac{b_H^2\varepsilon^2}{\mu^2}$. Moreover, strong convexity of f_ε implies that $f_\varepsilon(x_k) - f_\varepsilon^* \geq \frac{\mu}{2}\|x_k - x_\varepsilon^*\|^2$. Therefore, we can write

$$\|x_k - x^*\|^2 \leq 2\|x_k - x_\varepsilon^*\|^2 + 2\|x_\varepsilon^* - x^*\|^2 \leq \frac{4}{\mu}\left(f_\varepsilon(x_k) - f_\varepsilon^*\right) + \frac{2b_H^2\varepsilon^2}{\mu^2}.$$

Invoking relation (14), we obtain the inequality (15). □

5 CONCLUDING REMARKS

In this paper, we consider unconstrained stochastic optimization problems where the objective function is differentiable and strongly convex. To address this class of problems , we consider stochastic quasi-Newton methods. The convergence analysis and rate statements of the classical SQN methods presented in the literature require the objective function to have Lipschitzian gradients. Our goal in this paper is to weaken this assumption. To this end, employing a local smoothing technique, we develop a smoothing SQN method. Under standard assumptions on the stepsize and the approximate Hessian, we derive convergence properties of the scheme in both an almost sure and a mean sense. Importantly, we derive rate statements in terms of the expected error between the generated iterate and the optimal solution to the original problem under an additional assumption of twice continuous differentiability.. The development of efficient approximate Hessian update rules remains a future research direction.

REFERENCES

Bertsekas, D. P. 1973. "Stochastic Optimization Problems with Nondifferentiable Cost Functionals". *Journal of Optimization Theory and Applications* 12 (2): 218–231.

Byrd, R. H., S. L. Hansen, J. Nocedal, and Y. Singer. 2016. "A Stochastic Quasi-Newton Method for Large-Scale Optimization". *SIAM Journal on Optimization* 26 (2): 1008–1031.

Duchi, J. C., P. L. Bartlett, and M. J. Wainwright. 2012. "Randomized Smoothing for Stochastic Optimization". *SIAM Journal on Optimization (SIOPT)* 22 (2): 674–701.

Facchinei, F., and J.-S. Pang. 2003. *Finite-Dimensional Variational Inequalities and Complementarity Problems. Vols. I,II*. Springer Series in Operations Research. New York: Springer-Verlag.

Ghadimi, S., and G. Lan. 2012. "Optimal Stochastic Approximation Algorithms for Strongly Convex Stochastic Composite Optimization, part I: a generic algorithmic framework". *SIAM Journal on Optimization* 22 (4): 1469–1492.

Juditsky, A., A. Nemirovski, and C. Tauvel. 2011. "Solving Variational Inequalities with Stochastic Mirror-prox Algorithm". *Stochastic Systems* 1 (1): 17–58.

Lakshmanan, H., and D. Farias. 2008. "Decentralized Resource Allocation In Dynamic Networks of Agents". *SIAM Journal on Optimization* 19 (2): 911–940.

Lucchi, A., B. McWilliams, and T. Hofmann. "A Variance Reduced Stochastic Newton Method". *arXiv arXiv Preprint:1503.08316 (2015)*.

Mokhtari, A., and A. Ribeiro. 2014. "RES: Regularized Stochastic BFGS Algorithm". *IEEE Transactions on Signal Processing* 62 (23): 6089–6104.

Mokhtari, A., and A. Ribeiro. 2015. "Global Convergence of Online Limited Memory BFGS". *Journal of Machine Learning Research* 16:3151–3181.

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.

Polyak, B. 1987. *Introduction to optimization*. New York: Optimization Software, Inc.

Robbins, H., and S. Monro. 1951. "A Stochastic Approximation Method". *Annals of Mathematical Statistics* 22:400–407.

Schraudolph, N. N., J. Yu, and S. Gunter. 2007. "A Stochastic Quasi-Newton Method for Online Convex Optimization". *The 11th International Conference on Artificial Intelligence and Statistics (AISTATS)*:433–440.

Steklov, V. A. 1907. "Sur Les Expressions Asymptotiques Decertaines Fonctions Dfinies Par Les Equations Diffrentielles Du Second Ordre Et Leers Applications Au Problme Du Dvelopement D'une Fonction Arbitraire En Sries Procdant Suivant Les Diverses Fonctions". *Communications of the Kharkov Mathematical Society* 2 (10): 97–199.

Wang, X., S. Ma, and W. Liu. 2017. "Stochastic Quasi-Newton Methods for Nonconvex Stochastic Optimization". *SIAM Journal on Optimization* 27 (2): 927–956.

Yousefian, F., A. Nedić, and U. V. Shanbhag. 2012. "On Stochastic Gradient and Subgradient Methods with adaptive steplength sequences". *Automatica* 48 (1): 56–67. arXiv Preprint: http://arxiv.org/abs/1105.4549.

Yousefian, F., A. Nedić, and U. V. Shanbhag. 2013. "A Regularized Smoothing Stochastic Approximation (RSSA) Algorithm for Stochastic Variational Inequality Problems". In *Proceedings of the 2013 Winter Simulation Conference*, edited by A. T. R. H. R. Pasupathy, S.-H. Kim and M. E. Kuhl, 933–944. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Yousefian, F., A. Nedić, and U. V. Shanbhag. 2016a. "Self-tuned Stochastic Approximation Schemes for Non-Lipschitzian Stochastic Multi-user Optimization and Nash Games". *IEEE Transactions on Automatic Control* 61 (7): 1753–1766.

Yousefian, F., A. Nedić, and U. V. Shanbhag. 2016b. "Stochastic Quasi-Newton Methods for Non-strongly Convex Problems: Convergence and Rate Analysis". In *IEEE 55th Conference on Decision and Control (CDC), DOI: 10.1109/CDC.2016.7798953*.

Yousefian, F., A. Nedić, and U. V. Shanbhag. 2017a. "On Smoothing, Regularization, and Averaging in Stochastic Approximation Methods for Stochastic Variational Inequalities". *http://arxiv.org/pdf/1411.0209v2.pdf*. to appear in Mathematical Programming (Series B.).

Yousefian, F., A. Nedić, and U. V. Shanbhag. 2017b. "On Stochastic Mirror-prox Algorithms for Stochastic Cartesian Variational Inequalities: Randomized Block Coordinate, and Optimal Averaging Schemes". *Set-Valued and Variational Analysis, Under Review*. arXiv Preprint: https://arxiv.org/pdf/1610.08195v2.pdf.

AUTHOR BIOGRAPHIES

Farzad Yousefian is currently an assistant professor in the school of Industrial Engineering and Management at Oklahoma State University. Before joining OSU, he was a postdoctoral researcher in the Department of Industrial and Manufacturing Engineering at Penn State. He obtained his Ph.D. in industrial engineering from the University of Illinois at Urbana-Champaign in 2013. His thesis is focused on the design, analysis, and implementation of stochastic approximation methods for solving optimization and variational problems in nonsmooth and uncertain regimes. His current research interests lie in the development of efficient algorithms to address ill-posed stochastic optimization and equilibrium problems arising from machine learning and multi-agent systems. He is the recipient of the best theoretical paper award in the 2013 Winter Simulation Conference. His email addresses is farzad.yousefian@okstate.edu and his web page is https://sites.google.com/site/farzad1yousefian.

Angelia Nedić holds a Ph.D. from Moscow State University, Moscow, Russia, in Computational Mathematics and Mathematical Physics (1994), and a Ph.D. from Massachusetts Institute of Technology, Cambridge, USA in Electrical and Computer Science Engineering (2002). She has worked as a senior engineer in BAE Systems North America, Advanced Information Technology Division at Burlington, MA. She is the recipient of an NSF CAREER Award 2007 in Operations Research for her work in distributed multi-agent optimization. She is a recipient (jointly with her co-authors) of the Best Paper Award at the Winter Simulation Conference 2013 and the Best Paper Award at the International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt) 2015 (with co-authors). Her current interest is in large-scale optimization, games, control and information processing in networks. Her email address is Angelia.Nedich@asu.edu and her web page is https://ecee.engineering.asu.edu/people/angelia-nedich/.

Uday V. Shanbhag received his Ph.D. degree from the Department of Management Science and Engineering (specialization in operations research), Stanford University, Stanford, CA, in 2006. His research interests lie in the analysis and solution of stochastic optimization, game-theoretic and variational inequality problems with domain interests in power systems and markets and machine learning. He has held the Gary and Sheila Bello Chair in Industrial and Manufacturing Engineering at Pennsylvania State University since November, 2016. From 2006 to 2012, he was first an Assistant professor and subsequently a tenured Associate Professor (effective Summer, 2012) at the Industrial and Enterprise Systems Engineering (ISE) at the University of Illinois at Urbana-Champaign. He received the triennial A. W. Tucker Prize for his dissertation from the mathematical programming society (MPS) in 2006, the Computational Optimization and Applications (COAP) Best Paper Award (joint with Walter Murray) in 2007, the best paper award at the Winter Simulation Conference in 2013 (with F. Yousefian and A. Nedic), an NSF Faculty Early Career Development (CAREER) Award in Operations Research in 2012. Finally, since November 2016, he has been an Associate Editor for IEEE Transactions on Automatic Control. His email address is udaybag@psu.edu and his web page is http://personal.psu.edu/vvs3/.