ON STOCHASTIC AND DETERMINISTIC QUASI-NEWTON METHODS FOR NON-STRONGLY CONVEX OPTIMIZATION: CONVERGENCE AND RATE ANALYSIS

FARZAD YOUSEFIAN†, ANGELIA NEDIČ‡, AND UDAY V. SHANBHAG§

Abstract. Motivated by applications arising from large scale optimization and machine learning, we consider stochastic quasi-Newton (SQN) methods for solving unconstrained convex optimization problems. The convergence analysis of the SQN methods, both full and limited-memory variants, require the objective function to be strongly convex. However, this assumption is fairly restrictive and does not hold for applications such as minimizing the logistic regression loss function. To our knowledge, in the literature, no rate statements exist in the absence of this assumption for SQN methods. Also, there is no such rate for stochastic gradient methods addressing non-strongly convex problems with unbounded gradients. Motivated by these gaps, we consider optimization problems with non-strongly convex objectives with Lipschitz but possibly unbounded gradients. The main contributions of the paper are as follows: (i) Addressing large scale stochastic optimization problems, we develop a regularized stochastic limited-memory BFGS algorithm, where the stepsize, regularization parameter, and the Hessian inverse approximate matrix are updated iteratively. We establish the convergence to an optimal solution of the original problem both in an almost-sure and mean senses. We derive the convergence rate in terms of the objective function’s values and show that it is of the order $O\left( k^{-\left(\frac{1}{3} - \epsilon\right)} \right)$, where $\epsilon$ is an arbitrary small positive scalar; (ii) In deterministic regime, we show that the regularized limited-memory BFGS algorithm displays a rate of the order $O\left( \frac{1}{k^{1+\epsilon’}} \right)$, where $\epsilon’$ is an arbitrary small positive scalar. We present our numerical experiments performed on a large scale text classification problem.

Key words. stochastic optimization, quasi-Newton, regularization, large scale optimization

AMS subject classifications. 65K05, 90C06, 90C30, 90C53

1. Introduction. Consider the following stochastic optimization problem:

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

where $F : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ is a function, the random vector $\xi$ is defined as $\xi : \Omega \to \mathbb{R}^d$, $(\Omega, \mathcal{F}, \mathbb{P})$ denotes the associated probability space and the expectation $\mathbb{E}[F(x, \xi)]$ is taken with respect to $\mathbb{P}$. Problem (SO) provides a general framework that can capture a wide range of applications in operations research, machine learning, statistics and control to name a few. In machine learning applications, the objective function $f$ is given as an average of a loss function (see for example [2, 3, 28]). In such problems, a training set containing a large number of input/output pairs \( \{(u_i, v_i)\}_{i=1}^N \) is given where $u_i$ is the vector of inputs while $v_i$ denotes the output value. The goal is to learn a classifier (e.g., a hyperplane) $h(x, u)$ where $x$ is the vector of parameters of the function $h$ and $u$ is the input data. To measure the distance of an observed output $v_i$ from the classifier function $h$, a real-valued loss function $\ell(h; v)$ is defined. The
objective function is considered as the following averaged loss over the training set

\[ f(x) := \frac{1}{N} \sum_{i=1}^{N} \ell(h(x, u_i); v_i). \]  

When the size of the training sample \( N \) is very large, the preceding objective can be seen as a stochastic optimization model of the form (SO), where \( F(x, \xi) := \ell(h(x, u); v) \) and \( \xi = (u, v) \). Although, problem (SO) may be seen as a deterministic problem, the challenges still arise when standard deterministic schemes are employed. In particular, when the expectation is over a general measure space (computing \( \nabla_x E[F(x, \xi)] \) is difficult or impossible) or when the distribution \( P \) is unavailable, standard gradient or Newton-based schemes cannot be directly applied. This has led to significant research on Monte Carlo sampling techniques. Monte Carlo simulation methods have been used widely in the literature to solve stochastic optimization problems. Of these, sample average approximation (SAA) methods [26] and stochastic approximation (SA) methods [20,24] (also referred to as stochastic gradient methods in the context of optimization) are among popular approaches. It has been discussed that when the sample size is large, the computational effort for implementing SAA schemes does not scale with the number of samples and these methods become inefficient (cf. [20,34]). The SA methods, introduced by Robbins and Monro [24], produce a sequence \( \{x_k\} \), given a randomly generated \( x_0 \in \mathbb{R}^n \):

\[ x_{k+1} := x_k - \gamma_k \nabla F(x_k, \xi_k), \quad k \geq 0, \]  

where \( \gamma_k > 0 \) denotes the stepsize and \( \nabla F(x_k, \xi_k) \) denotes the sampled gradient of the function \( f \) with respect to \( x \) at \( x_k \). Note that the sample gradient \( \nabla F(x_k, \xi_k) \) is assumed to be an unbiased estimator of the true value of the gradient \( \nabla f(x) \) at \( x_k \).

The SA schemes are characterized by several disadvantages, including the poorer rate of convergence (than their deterministic counterparts) and the detrimental impact of conditioning on their performance. In deterministic regimes, Newton schemes (when second derivatives are available) and their quasi-Newton counterparts have proved to be useful alternatives, particularly from the standpoint of displaying faster rates of convergence. Recently, an incremental Newton method has been developed addressing minimizing the sum of a large number of strongly convex functions [8]. The BFGS method, named after Broyden, Fletcher, Goldfarb, and Shanno, is known as the most celebrated quasi-Newton method [5, 7] and displays a super-linear convergence rate without requiring the second-order information. Addressing large scale deterministic problems, the limited-memory variant of the BFGS method, denoted by L-BFGS, was developed [13] that accepts an \( \mathbb{R} \)-linear convergence rate under strong convexity of the objective function (see Theorem 6.1 in [13]).

Recently, there has been a growing interest in applying stochastic variants of quasi-Newton methods, denoted by SQN methods, for solving large scale optimization and machine learning problems. In these methods, \( x_k \) is given by the following rule:

\[ x_{k+1} := x_k - \gamma_k H_k \nabla F(x_k, \xi_k), \quad k \geq 0, \]  

where \( H_k \geq 0 \) is an approximation of the inverse of the Hessian matrix at iteration \( k \) that incorporates the curvature information of the objective function within the algorithm. The convergence of this class of algorithms can be derived under a careful choice of the matrix \( H_k \) and the stepsize sequence \( \gamma_k \). In particular, boundedness of the eigenvalues of \( H_k \) is an important factor in achieving global convergence in convex
and nonconvex problems (cf. [1, 12]). While in [25] the performance of SQN methods displayed to be favorable in solving high dimensional problems, Mokhtari et al. [17] considered stochastic optimization problems with strongly convex objectives and developed a regularized BFGS method (RES) by updating the matrix $H_k$ according to a modified version of BFGS update rule to assure convergence. To address large scale applications, limited-memory variants were employed to ascertain scalability in terms of the number of variables [3, 18]. In a recent extension [31], a stochastic quasi-Newton method is presented for solving nonconvex stochastic optimization problems. Also, a variance reduced SQN method with a constant stepsize was developed [14] for smooth strongly convex problems characterized by a linear convergence rate.

Motivation: In the developed SQN methods, both full and limited-memory variants (e.g. [3, 18]), it is uniformly assumed that the objective function is strongly convex. This assumption plays an important role in deriving the rate of convergence of the algorithm. However, in many applications, the objective function is convex, but not strongly convex such as, for example, the logistic regression function that is given by $\ell(u^T x, v) := \ln(1 + \exp(-u^T x v))$ for $u, x \in \mathbb{R}^n$ and $v \in \mathbb{R}$. While lack of strong convexity might lead to a very slow convergence, no theoretical results on the convergence rate is given in the literature of SQN methods. A simple remedy to address this challenge is to regularize the objective function with the term $\frac{1}{2} \mu \|x\|^2$ and solve the approximate problem of the form

$$\min_{x \in \mathbb{R}^n} f(x) + \frac{\mu}{2} \|x\|^2,$$

where $\mu > 0$ is the regularization parameter. Several challenges arise in applying this technique. A drawback of this technique is that the optimal solution to the approximate problem (2) is not an optimal of the original problem (SO). The main challenge is related to the choice of $\mu$. While large values of $\mu$ result in large deviations from the true optimal solutions, importantly, choosing $\mu$ to be a small number leads to a deterioration of the convergence rate of the algorithm because the constant factor is proportional to the inverse of $\mu$ (see Theorem 7 in [18], or Theorem 3.2 in [3]). This issue is resolved in SA schemes through employing averaging techniques for non-strongly convex problems displaying the rate of $O\left(\frac{1}{\sqrt{k}}\right)$ (see [19, 20]). More precisely, in [20] the following rate for the stochastic mirror descent (SMD) method is derived:

$$\mathbb{E}[f(\bar{x}_k)] - f^* = O\left(\frac{D_X M}{\sqrt{k}}\right),$$

where $\bar{x}_k$ is a window-based weighted averaging sequence generated by the SMD scheme, $D_X$ is an upper bound on the Bregman function over the constraint set $X$, and $M$ is an upper bound on the norm of the stochastic gradient $\nabla F(x, \xi)$ (see Section 2.3 in [20] for more details). Averaging SA scheme and their generalized variant, the SMD method, have been widely used to address optimization problems in absence of strong convexity and have a robust performance in presence on uncertainty. However, the main limitation to averaging schemes is that boundedness of the gradient mapping is required to achieve such a rate (e.g., boundedness of $M$ in (3)). In this paper, our main goal lies in employing an iterative regularization scheme in order to address such shortcomings and establish a rate of convergence in absence of strong convexity. Next, we review the relevant research on regularization techniques.

Related research on regularization: In solving optimization problems, in order to obtain solutions with desirable properties, it is common to regularize the problem
(SO) as follows

\[ \min_{x \in \mathbb{R}^n} f_\mu(x) := f(x) + \mu R(x), \]

where \( R : \mathbb{R}^n \to \mathbb{R} \) is a proper convex function and the scalar \( \mu > 0 \) is the regularization parameter. Examples of regularized optimization include:

**Tikhonov regularization:** The most common form of regularization is named after Andrey Tikhonov when we set \( R(x) = \frac{1}{2} \|x\|^2 \) where \( \| \cdot \| \) denotes the Euclidean norm. This technique can be used to obtain the least \( \ell_2 \)-norm optimal solution of the problem (SO). It is also employed to establish convergence rates to an approximate optimal solution, for example in minimizing the logistic regression loss function.

**\( \ell_1 \) regularization:** This is when we choose \( R(x) = \|x\|_1 \) which has been widely used to improve sparsity of the solution in large scale problems. A special case is Lasso [29] where \( f \) is a linear least-squares function of the form (1) with \( \ell(u^T x; v) := \frac{1}{2} (u^T x - v)^2 \). Applications of \( \ell_1 \) regularization have been studied by researchers in different areas such as signal processing and compressed sensing (e.g., [9]).

**\( \ell_\infty \) regularization:** This is when \( R(x) = \|x\|_\infty \) (also referred to as max-norm regularization) and has been applied in collaborative filtering problems [27].

**Frobenius norm regularization:** In matrix completion problems, for example, where the goal is estimating a low-rank matrix given a linear system of equations, the problem can be reformulated and regularized using Frobenius norm regularization [23,32].

Applications of the regularized problem and its relation with the original problem have been investigated by different researchers. Mangasarian and his colleagues appear among the first researchers who studied exact regularization of linear and nonlinear programs [15,16]. A regularization is said to be exact when an optimal solution of (4), is also optimal for problem (SO) if \( \mu \) is small enough. Paul Tseng et al. [6,30] established the necessary and sufficient conditions of exact regularization for convex programs and derived error bounds for inexact regularized convex problems. Very recently, exact regularization of variational inequities has been studied in [4].

There is a trade-off between the deviation from the original problem due to regularization by the term \( \mu R(x) \) and the resulted properties of the solution. A challenging question is concerned about the choice of the regularization parameter \( \mu \). A common approach to find an acceptable value for \( \mu \) is through a two-loop scheme where in the inner loop, problem (4) is solved for a fixed value of \( \mu \), while in the outer loop, \( \mu \) is tuned. The main drawback of this approach is that in general, there is no guidance on the tuning rule for \( \mu \). In addition, this approach is computationally inefficient. Our main goal in this paper lies in resolving this issue. The main idea is to implement an iterative single-loop algorithm where we update the regularization parameter \( \mu \) at each iteration of the scheme and reduce it iteratively to converge to zero. There are two important questions arising in this framework: (i) what is the proper update rule for \( \mu \) within the algorithm to ascertain the convergence to the optimal solution of (SO)? (ii) given that the algorithm generates \( x_k \) at iteration \( k \), what is the convergence rate of \( f(x_k) \) to the optimal objective function of the original problem? The first question has been addressed for merely monotone stochastic variational (VI) inequalities in [10]. Extensions of this work for the non-Lipschitzian VIs have been recently developed in our work [36]. However, we are unaware of any work that has addressed the second question for the class of stochastic and deterministic quasi-Newton methods.

**Contributions:** Motivated by this gap, our goal lies in developing algorithms for finding an exact optimal solution of (SO) equipped with convergence rate statements. We
consider stochastic optimization problems with non-strongly convex objective functions and Lipschitz but possibly unbounded gradient mappings. Our main contributions are as follows:

(i) Establishing convergence: We develop a regularized SQN method in that the stepsize, regularization parameter, and the approximate Hessian inverse matrix are updated iteratively. We first assume that the matrix \( H_k \) satisfies a set of general assumptions on its eigenvalues, and its dependency to the uncertainty. The convergence of the algorithm is established under a careful choice of the Lyapunov function. Given that the algorithm generates the sequence \( x_k \), we obtain a set of suitable conditions on the stepsize and regularization sequences for which \( f(x_k) \) converges to the optimal objective value, i.e., \( f^* \), of (SO) in an almost-sure sense and in mean. We also derive an upper bound for \( f_k - f^* \) in term of the two sequences and problem parameters.

(ii) Establishing rates of convergence for regularized L-BFGS methods: To address large scale stochastic optimization problems, motivated by our earlier work [35] on development of SQN methods for small scale stochastic optimization problems with non-strongly convex objectives, here we consider regularized SQN methods where the matrix \( H_k \) is updated according to a stochastic limited-memory BFGS rule. The main distinction between this algorithm and the work on SQN methods in the literature (e.g., [3, 18]) is two-fold: (a) the gradient mapping is regularized in terms of a varying regularization parameter that is updated through the implementation of the algorithm; (b) the matrix \( H_k \) is regularized in terms of the varying regularization parameter and is updated iteratively. We show that under a careful choice of the step-size sequence and the regularization parameter, the developed algorithm displays a convergence rate in terms of the objective function’s values of the order \( O(k^{-(\frac{1}{2}-\epsilon)}) \), where \( \epsilon \) is an arbitrary small positive scalar. We also consider a regularized deterministic L-BFGS method and show that the convergence rate is upgraded to the order \( O \left( \frac{1}{k^{1-\epsilon'}} \right) \), where \( \epsilon' \) is an arbitrary small positive scalar. This result is new for the classical deterministic limited-memory BFGS methods where the strong convexity of the objective function is required to achieve the \( \mathbb{R} \)-linear convergence rate (see Assumption 6.1 and Theorem 6.1 in [13]).

Outline of the paper: The rest of the paper is organized as follows. Section 2 presents the general framework of the proposed SQN algorithm and the sets of main assumptions. In Section 3, we prove the convergence of the scheme in both almost sure and expected senses and derive the a general error bound. In Section 4, we develop a regularized stochastic L-BFGS method and derive the convergence rate. The rate analysis is also provided for the deterministic variant of this scheme. We present the results of our numerical experiments performed on a large scale classification problem in Section 5. The paper ends with some concluding remarks in Section 6.

Notation: A vector \( x \) is assumed to be a column vector and \( x^T \) denotes its transpose, while \( \|x\| \) denotes the Euclidean vector norm, i.e., \( \|x\| = \sqrt{x^T x} \). We write \( a.s. \) as the abbreviation for “almost surely”. For a symmetric matrix \( B \), we write \( \lambda_{\text{min}}(B) \) and \( \lambda_{\text{max}}(B) \) to denote its smallest and largest eigenvalue, respectively. We use \( \mathbb{E}[z] \) to denote the expectation of a random variable \( z \). A function \( f : X \subset \mathbb{R}^n \rightarrow \mathbb{R} \) is said to be strongly convex with parameter \( \mu > 0 \), if \( f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2} \|x-y\|^2 \), for any \( x, y \in X \). A mapping \( F : X \subset \mathbb{R}^n \rightarrow \mathbb{R} \) is Lipschitz continuous with parameter \( L > 0 \) if for any \( x, y \in X \), we have \( \|F(x) - F(y)\| \leq L\|x-y\| \). For a continuously differentiable function \( f \) with Lipschitz gradients with parameter \( L > 0 \), we have \( f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2} \|x-y\|^2 \), for any \( x, y \in X \). For a vector \( x \in \mathbb{R}^n \) and a non-empty set \( X \subset \mathbb{R}^n \), the Euclidean distance of \( x \) from \( X \) is denoted
by $\text{dist}(x, X)$. We denote the optimal objective value of problem (SO) by $f^*$ and the set of the optimal solutions by $X^*$.

### 2. Outline of the algorithm.

In this section, we first begin by describing a general SQN algorithm that captures stochastic BFGS methods discussed in subsequent sections. We then state the main assumptions for problem (SO) and note on some of the distinctions with the assumptions made in the literature.

Our algorithm is defined through an iterative regularized SQN scheme: Let $x_0 \in \mathbb{R}^n$ be an arbitrary initial point, and $x_k$ be generated by the following recursive rule

\[(R\text{-SQN}) \quad x_{k+1} := x_k - \gamma_k H_k (\nabla F(x_k, \xi_k) + \mu_k (x_k - x_0)), \quad \text{for all } k \geq 0.
\]

Here, $\gamma_k$ and $\mu_k$ are the steplength and the regularization parameter, respectively. $H_k \in \mathbb{R}^{n \times n}$ is a matrix that carries out the curvature information of the objective function. The (R-SQN) scheme can be seen as a regularized variant of classical stochastic SQN methods. Here we regularize the gradient map by the term $\mu_k x_k$ to inherit the strong monotonicity property. In absence of strong convexity of $f$, unlike the classical schemes where $\mu_k$ is maintained fixed, we let $\mu_k$ be updated and reduce to zero. There are three challenging questions that will be addressed in our analysis: (i) At what rate $\gamma_k$ and $\mu_k$ should be updated to recover convergence? (ii) What is the convergence rate of the scheme? (iii) How to update $H_k$ to incorporate the curvature information while maintaining a secant condition despite lacking the strong convexity property? These questions are addressed in the following sections depending on the type of the problem (stochastic or deterministic) and type of the underlying scheme (L-BFGS or gradient method). Throughout, we let $\mathcal{F}_k$ denote the history of the method up to time $k$, i.e.,

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

and $\mathcal{F}_0 = \{x_0\}$. Next, we state the main assumptions.

**Assumption 1.** Consider problem (SO).

(a) The function $f(x)$ is convex over $\mathbb{R}^n$.

(b) $f(x)$ is continuously differentiable with Lipschitz continuous gradients over $\mathbb{R}^n$ with parameter $L > 0$.

(c) The optimal solution set of problem (SO) is nonempty.

**Remark 1.** A key distinction between our work and prior research papers (e.g. [3, 18]) is that we weaken the strong convexity condition and replace it by the convexity property in Assumption 1a. The convexity assumption of $f$ is essential in the convergence analysis and deriving the convergence rate. In the subsequent sections, when required, we substitute this assumption by convexity of $F(\cdot, \xi)$ for all $\xi \in \mathbb{R}^d$.

Next, we state the assumptions on the random variable $\xi$ and the properties of the stochastic estimator of the gradient mapping, i.e. $\nabla F$.

**Assumption 2 (Random variable $\xi$).**

(a) Random variables $\xi_k$ are i.i.d. for any $k \geq 0$;

(b) The stochastic gradient mapping $\nabla F(x, \xi)$ is an unbiased estimator of $\nabla f(x)$, i.e. $\mathbb{E}[\nabla F(x, \xi)] = \nabla f(x)$ for all $x$, and has a bounded variance, i.e., there exists a scalar $\nu > 0$ such that $\mathbb{E}[\|\nabla F(x, \xi) - \nabla f(x)\|^2] \leq \nu^2$, for all $x \in \mathbb{R}^n$.

**Remark 2.** By Assumption 2(b), we study settings where the stochastic gradient mapping $\nabla F(x, \xi)$ is unbiased with a bounded second moment. Our results can be extended to the case of a biased estimator. The boundedness of the second moment
is a classical assumption in stochastic optimization that cannot be relaxed in general unless for example, a suitable variance reduction scheme is employed. Here we want to state a key distinction between our assumption with that of [3,18]. In both of these papers, it is assumed that \( F(x,\xi) \) is strongly convex, smooth and has bounded gradients over \( \mathbb{R}^n \) for all \( \xi \). However, it is not hard to show that there does not exist such a function. Let us assume \( F(x,\xi) \) is differentiable and strongly convex with parameter \( \mu > 0 \). Therefore, we can write
\[
\mu \|x - y\|^2 \leq (x - y)^T(\nabla F(x,\xi) - \nabla F(y,\xi)) \leq \|x - y\|\|\nabla F(x,\xi) - \nabla F(y,\xi)\|,
\]
for some \( \xi \) and any \( x, y \in \mathbb{R}^n \). Consider the sequence \( x_k = k1_n \) and let \( y = 0 \), where \( 1_n \) denotes a vector in \( \mathbb{R}^n \) with all unit elements. From the preceding inequality, replacing \( x \) and \( y \) by their values and canceling the term \( \|x - y\| \), we obtain
\[
\mu \|k1_n - 0\| \leq \|\nabla F(k1_n,\xi) - \nabla F(0,\xi)\| \leq \|\nabla F(k1_n,\xi)\| + \|\nabla F(0,\xi)\|,
\]
where the last relation is implied by the triangle inequality. Taking limit from the preceding relation, we get
\[
\lim_{k \to \infty} \|\nabla F(k1_n,\xi)\| \geq \sqrt{n} \mu \lim_{k \to \infty} k - \|\nabla F(0,\xi)\| = \infty,
\]
where without loss of generality, we assume that \( \|\nabla F(0,\xi)\| \) is finite. This contradicts the assumption that \( \nabla F \) is bounded. In this work, we relax the boundedness assumption and replace it by boundedness of the second moment of \( \nabla F \).

The next assumption is about the properties of \( H_k \). We assume that \( H_k \) is independent of \( \xi_k \) and its eigenvalues are bounded in terms of the regularization parameter. 

**Assumption 3 (Conditions on matrix \( H_k \)).** Let the following hold for all \( k \geq 0 \):
(a) The matrix \( H_k \in \mathbb{R}^{n \times n} \) is \( \mathcal{F}_k \)-measurable, i.e., \( \mathbb{E}[H_k \mid \mathcal{F}_k] = H_k \).
(b) Matrix \( H_k \) is symmetric and positive definite and satisfies the following condition: There exist positive scalars \( \lambda_{\text{min}}, \lambda \) and scalar \( \alpha \leq 0 \) such that
\[
\lambda_{\text{min}} I \leq H_k \leq \lambda \mu_k^\alpha I, \quad \text{for all} \quad k \geq 0.
\]
Assumption 3 holds for the stochastic gradient method where \( H_k \) is the identity matrix, \( \lambda_{\text{min}} = \lambda = 1 \) and \( \alpha = 0 \). In the case of employing an appropriate L-BFGS update rule that will be discussed in Section 4, the maximum eigenvalue is obtained in terms of the regularization parameter.

**3. Convergence analysis.** In this section we present the convergence analysis of the (R-SQN) method. Our discussion starts by some preliminary definitions and properties. After obtaining a recursive error bound for the method in Lemma 3, we show a.s. convergence in Theorem 4, and establish convergence in mean and derive an error bound in Theorem 5. The specific convergence rate statements will be provided for stochastic BFGS schemes in subsequent sections.

**Definition 1 (Regularized function and gradient mapping).** Consider the sequence \( \{\mu_k\} \) of positive scalars and the starting point of the algorithm (R-SQN), i.e., \( x_0 \). The regularized function \( f_k : \mathbb{R}^n \to \mathbb{R} \) and its gradient are defined as follows:
\[
f_k(x) \triangleq f(x) + \frac{\mu_k}{2}\|x - x_0\|^2, \quad \text{for any} \quad k \geq 0,
\]
\[
\nabla f_k(x) \triangleq \nabla f(x) + \mu_k(x - x_0), \quad \text{for any} \quad k \geq 0.
\]
In a similar way, we denote the regularized stochastic function \( F(x,\xi) \) and its gradient with \( F_k \) and \( \nabla F_k \) for any \( \xi \), respectively.
Assumptions 1, 2, and 3 hold. Also, assume $E_0$ and $\mu$. Note that under Assumption 2, from the definition of (5)

$$2\mu_k(f_k(x) - f_k(x^*_k)) \leq \|\nabla f_k(x)\|^2 \leq 2(L + \mu_k)(f_k(x) - f_k(x^*_k)).$$

The existence and uniqueness of $x^*_k$ in Property 1(c) is due to the strong convexity of the function $f_k$ (see, for example, Section 1.3.2 in [22]), while the relation for the gradient is known to hold for a strongly convex function with a parameter $\mu$ that also has Lipschitz gradients with a parameter $L$ (see Lemma 1 on page 23 in [22]). In the convergence analysis, we make use of the following result, which can be found in [22] (see Lemma 11 on page 50).

**Lemma 2.** Let $\{v_k\}$ be a sequence of nonnegative random variables, where $E[v_0] < \infty$, and let $\{\alpha_k\}$ and $\{\beta_k\}$ be deterministic scalar sequences such that:

$$E[v_{k+1}|v_0, \ldots, v_k] \leq (1 - \alpha_k)v_k + \beta_k \quad a.s. \text{ for all } k \geq 0,$$

$$0 \leq \alpha_k \leq 1, \quad \beta_k \geq 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \beta_k < \infty, \quad \lim_{k \to \infty} \frac{\beta_k}{\alpha_k} = 0.$$

Then, $v_k \to 0$ almost surely.

Throughout, we denote the stochastic error of the gradient estimator by

$$w_k \triangleq \nabla F(x_k, \xi_k) - \nabla f(x_k), \quad \text{for all } k \geq 0.$$

Note that under Assumption 2, from the definition of $w_k$ in (5), we obtain $E[w_k \mid F_k] = 0$ and $E[\|w_k\|^2 \mid F_k] < \nu^2$.

**Lemma 3** (A recursive error bound). Consider the (R-SQN) method and suppose Assumptions 1, 2, and 3 hold. Also, assume $\mu_k$ is a non-increasing sequence and let $\gamma_k$ and $\mu_k$ satisfy

$$\gamma_k \mu_k^2 \leq \frac{\lambda_{\min}}{\lambda^2(L + \mu_0)}, \quad \text{for all } k \geq 0.$$

Then, the following inequality holds:

$$E[f_{k+1}(x_{k+1}) \mid F_k] - f^* \leq (1 - \lambda_{\min}\mu_k \gamma_k)(f_k(x_k) - f^*) + \frac{\lambda_{\min} \text{dist}^2(x_0, X^*)}{2} \mu_k^2 \gamma_k$$

$$+ \frac{(L + \mu_k)\lambda^2 \nu^2}{2} \mu_k^{2\alpha} \gamma_k^2.$$

**Proof.** The Lipschitzian property of $\nabla f_k$ and the update rule (R-SQN) imply that

$$f_k(x_{k+1}) \leq f_k(x_k) + \nabla f_k(x_k)^T(x_{k+1} - x_k) + \frac{(L + \mu_k)}{2} \|x_{k+1} - x_k\|^2$$

$$= f_k(x_k) - \gamma_k \nabla f_k(x_k)^T H_k (\nabla F(x_k, \xi_k) + \mu_k(x_k - x_0))$$

$$+ \frac{(L + \mu_k)}{2} \gamma_k^2 \|H_k (\nabla F(x_k, \xi_k) + \mu_k(x_k - x_0))\|^2.$$
Next, taking conditional expectations in (8), and using (9) and (10), we obtain

\[ f_k(x_{k+1}) \leq f_k(x_k) - \gamma_k \nabla f_k(x_k)^T H_k (\nabla f(x_k) + w_k + \mu_k(x_k - x_0)) + \frac{(L + \mu_k)}{2} \gamma_k^2 \| H_k (\nabla f(x_k) + w_k) \|^2 \]

(8)

where in the last equality we used the definition of \( f_k \). Next, we estimate the conditional expectation of Term 1 and 2. From Assumption 3, we have

\[ \text{Term 1} = \nabla f_k(x_k)^T H_k \nabla f_k(x_k) + \nabla f_k(x_k)^T H_k w_k \geq \lambda_{\min} \| \nabla f_k(x_k) \|^2 + \nabla f_k(x_k)^T H_k w_k. \]

Taking expectations conditioned on \( F_k \), from the preceding inequality, we obtain

(9) \[ \mathbb{E}[\text{Term 1} \mid F_k] \geq \lambda_{\min} \| \nabla f_k(x_k) \|^2 + \mathbb{E}[\nabla f_k(x_k)^T H_k w_k \mid F_k] = \lambda_{\min} \| \nabla f_k(x_k) \|^2, \]

where we recall that \( \mathbb{E}[w_k \mid F_k] = 0 \) and \( \mathbb{E}[H_k \mid F_k] = H_k \). Similarly, in Term 2, invoking Assumption 3(b), we may write

\[ \text{Term 2} = (\nabla f_k(x_k) + w_k)^T H_k^2 (\nabla f_k(x_k) + w_k) \leq (\lambda \mu_k)^2 \| \nabla f_k(x_k) + w_k \|^2 = \lambda^2 \mu_k^2 \alpha (\| \nabla f_k(x_k) \|^2 + \| w_k \|^2 + 2 \nabla f_k(x_k)^T w_k). \]

Taking conditional expectations from the preceding inequality, and using Assumption 2, we obtain

\[ \mathbb{E}[\text{Term 2} \mid F_k] \leq \lambda^2 \mu_k^2 \alpha (\| \nabla f_k(x_k) \|^2 + \nu^2). \]

Next, taking conditional expectations in (8), and using (9) and (10), we obtain

\[ \mathbb{E}[f_k(x_{k+1}) \mid F_k] \leq f_k(x_k) - \gamma_k \lambda_{\min} \| \nabla f_k(x_k) \|^2 + \lambda^2 \mu_k^2 \alpha \left( \frac{(L + \mu_k)}{2} \right) \gamma_k^2 \| \nabla f_k(x_k) \|^2 + 2 \gamma_k \lambda_{\min} \| \nabla f_k(x_k) \|^2 \left( 2 - \frac{\lambda^2 \mu_k^2 \alpha \gamma_k (L + \mu_k)}{\lambda_{\min}} \right) + \lambda^2 \mu_k^2 \alpha \left( \frac{(L + \mu_k)}{2} \right) \gamma_k^2 \nu^2. \]

From the assumption that \( \gamma_k \) and \( \mu_k \) satisfy \( \gamma_k \mu_k^2 \alpha \leq \frac{\lambda_{\min}}{\lambda^2 (L + \mu_k)} \) for any \( k \geq 0 \) and that \( \mu_k \) is non-increasing, we have \( \gamma_k \mu_k^2 \alpha \leq \frac{\lambda_{\min}}{\lambda^2 (L + \mu_k)} \). As a consequence, we get

\[ 2 - \frac{\lambda^2 \mu_k^2 \alpha \gamma_k (L + \mu_k)}{\lambda_{\min}} \geq 1. \]

Therefore, from preceding inequality, we obtain

\[ \mathbb{E}[f_k(x_{k+1}) \mid F_k] \leq f_k(x_k) - \frac{\lambda_{\min} \mu_k \gamma_k}{2} (f_k(x_k) - f_k(x_k^*)) + \lambda^2 \mu_k^2 \alpha \left( \frac{(L + \mu_k)}{2} \right) \gamma_k^2 \nu^2. \]

Employing Property 1(c), we have

\[ \mathbb{E}[f_k(x_{k+1}) \mid F_k] \leq f_k(x_k) - \lambda_{\min} \mu_k \gamma_k (f_k(x_k) - f_k(x_k^*)) + \lambda^2 \mu_k^2 \alpha \left( \frac{(L + \mu_k)}{2} \right) \gamma_k^2 \nu^2. \]
Note that, since $\mu_k$ is a non-increasing sequence, Definition 1 implies that
\[
E[f_{k+1}(x_{k+1}) \mid F_k] \leq E[f_k(x_{k+1}) \mid F_k].
\]

Therefore, we obtain
\[
E[f_{k+1}(x_{k+1}) \mid F_k] \leq f_k(x_k) - \lambda \min \mu_k \gamma_k (f_k(x_k) - f_k(x^*_k)) + \frac{\lambda^2 \mu_k^2}{2} \frac{(L + \mu_k)}{\gamma_k^2} \nu^2. 
\]

Next, we derive a lower bound for Term 3. Since $x^*_k$ is the unique minimizer of $f_k$, we have $f_k(x^*_k) \leq f_k(x_k)$. Therefore, invoking Definition 1, for an arbitrary optimal solution $x^* \in X^*$, we have
\[
f_k(x_k) - f_k(x^*_k) \geq f_k(x_k) - f_k(x^*) = f_k(x_k) - f^* - \frac{\mu_k}{2} \|x^* - x_0\|^2.
\]

From the preceding relation and (11), we have
\[
E[f_{k+1}(x_{k+1}) \mid F_k] \leq f_k(x_k) - \lambda \min \mu_k \gamma_k (f_k(x_k) - f^*) + \frac{\lambda \min \|x^* - x_0\|^2}{2} \mu_k^2 \gamma_k \\
+ \frac{(L + \mu_k)\lambda^2 \nu^2}{2} \mu_k^2 \gamma_k.
\]

Since $x^*$ is an arbitrary optimal solution, taking minimum from the right-hand side of the preceding inequality over $X^*$, we can replace $\|x^*-x_0\|$ by dist$(x_0, X^*)$. Then, subtracting $f^*$ from both sides of the resulting relation yields the desired inequality.$\square$

Next, we show the convergence of the scheme. In order to apply Lemma 2 to inequality (7) and prove the almost sure convergence, we use the following definitions:
\[
\begin{align*}
v_k & := f_k(x_k) - f^*, \quad \alpha_k := \lambda \min \gamma_k \mu_k, \\
\beta_k & := \frac{\lambda \min \text{dist}^2(x_0, X^*)}{2} \mu_k^2 \gamma_k + \frac{(L + \mu_k)\lambda^2 \nu^2}{2} \mu_k^2 \gamma_k.
\end{align*}
\]

To satisfy the conditions of Lemma 2, we identify a set of sufficient conditions on the sequences $\{\gamma_k\}$ and $\{\mu_k\}$ in the following assumption. Later in the subsequent sections, for each class of algorithms, we provide a set of sequences that meet these assumptions.

**Assumption 4. [Sufficient conditions on sequences for a.s. convergence]** Let the sequences $\{\gamma_k\}$ and $\{\mu_k\}$ be positive and satisfy the following conditions:

(a) $\lim_{k \to \infty} \gamma_k \mu_k^2 = 0$; \quad (b) $\{\mu_k\}$ is non-increasing and $\mu_k \to 0$;

(c) $\lambda \min \gamma_k \mu_k \leq 1$ for $k \geq 0$; \quad (d) $\sum_{k=0}^\infty \gamma_k \mu_k = \infty$;

(e) $\sum_{k=0}^\infty \mu_k^2 \gamma_k < \infty$; \quad (f) $\sum_{k=0}^\infty \gamma_k^2 \mu_k^2 < \infty$;

**Theorem 4. [Almost sure convergence]** Consider the (R-SQN) scheme. Suppose Assumptions 1, 2, 3 and 4 hold. Then, $\lim_{k \to \infty} f(x_k) = f^*$ almost surely.

**Proof.** First, note that from Assumption 4(a), there exists $K \geq 1$ such that for any $k \geq K$, we have $\gamma_k \mu_k^2 \leq \frac{\lambda \min}{L + \mu_k}$ implying that condition (6) of Lemma 3 holds.
for all $k \geq K$. Hence, relation (7) holds for any $k \geq K$. Next, we apply Lemma 2 to prove a.s. convergence of the (R-SQN) scheme. Consider the definitions in (12) for any $k \geq K$. The non-negativity of $\alpha_k$ and $\beta_k$ is implied by the definition and that $\lambda_{\text{min}}, \gamma_k$ and $\mu_k$ are positive. From (7), we have
\[
E[v_{k+1} | F_k] \leq (1 - \alpha_k)v_k + \beta_k \quad \text{for all } k \geq K.
\]
Since $f^* \leq f(x)$ for any arbitrary $x \in \mathbb{R}^n$, we can write
\[
v_k = f_k(x_k) - f^* = (f(x_k) - f^*) + \frac{\mu_k}{2}\|x_k - x_0\|^2 \geq 0.
\]
From Assumption 4(c), we obtain $\alpha_k \leq 1$. Also, from Assumption 4(d), we get $\sum_{k=1}^{\infty} \alpha_k = \infty$. Using Assumption 4(b,c,f) and the definition of $\beta_k$ in (12), for an arbitrary solution $x^*$, we can write
\[
\sum_{k=1}^{\infty} \beta_k \leq \frac{\lambda_{\text{min}} \text{dist}^2(x_0, X^*)}{2} + \sum_{k=1}^{\infty} \mu_k^2 \gamma_k + \frac{(L + \mu_0)\lambda^2 \nu^2}{2} \sum_{k=1}^{\infty} \mu_k^{2\alpha - 1} \gamma_k^2 < \infty.
\]
Similarly, we can write
\[
\lim_{k \to \infty} \beta_k \leq \frac{\text{dist}^2(x_0, X^*)}{2} \lim_{k \to \infty} \mu_k + \frac{(L + \mu_0)\lambda^2 \nu^2}{2} \lim_{k \to \infty} \mu_k^{2\alpha - 1} \gamma_k = 0,
\]
where the last equation is implied by Assumption 4(a,b). Therefore, all conditions of Lemma 2 hold and we conclude that $v_k := f_k(x_k) - f^*$ converges to 0 a.s. Let us define $v_k' := f_k(x_k) - f^*$ and $v_k'' := \frac{\mu_k}{2}\|x_k - x_0\|^2$, so that $v_k = v_k' + v_k''$. Since $v_k'$ and $v_k''$ are non-negative, and $v_k \to 0$ a.s., it follows that $v_k' \to 0$ and $v_k'' \to 0$ a.s., implying that $\lim_{k \to \infty} f(x_k) = f^*$ a.s. 

In the following, our goal is to state the assumptions on the sequences $\{\gamma_k\}$ and $\{\mu_k\}$ under which we can show the convergence in mean.

\textbf{Assumption 5.} [Sufficient conditions on sequences for convergence in mean] Let the sequences $\{\gamma_k\}$ and $\{\mu_k\}$ be positive and satisfy the following conditions:
\begin{itemize}
  \item[(a)] $\lim_{k \to \infty} \gamma_k \mu_k^{2\alpha - 1} = 0$;
  \item[(b)] $\{\mu_k\}$ is non-increasing and $\mu_k \to 0$;
  \item[(c)] $\lambda_{\text{min}} \gamma_k \mu_k \leq 1$ for $k \geq 1$;
  \item[(d)] There exist $K_0$ and $0 < \beta < 1$ such that
    \[\gamma_k \mu_k^{2\alpha - 1} \leq \gamma_k \mu_k^{2\alpha - 1} (1 + \beta \lambda_{\text{min}} \gamma_k \mu_k), \quad \text{for all } k \geq K_0;\]
  \item[(e)] There exists a scalar $\rho > 0$ such that $\mu_k^{2 - 2\alpha} \leq \rho \gamma_k$ for all $k \geq 0$;
\end{itemize}

Next, we use Assumption 5 to establish the convergence in mean.

\textbf{Theorem 5.} [Convergence in mean] Consider the (R-SQN) scheme. Suppose Assumptions 1, 2, 3 and 5 hold. Then, there exists $K \geq 1$ such that the following inequality holds:
\[
E[f(x_{k+1})] - f^* \leq \theta \gamma_k \mu_k^{2\alpha - 1}, \quad \text{for all } k \geq K,
\]
where $f^*$ is the optimal value of problem (SO),
\[
\theta := \max \left\{ \frac{E[f_k(x_{k+1})] - f^*}{\gamma_k \mu_k^{2\alpha - 1}}, \frac{\rho \lambda_{\text{min}} \text{dist}^2(x_0, X^*) + (L + \mu_0)\lambda^2 \nu^2}{2(1 - \beta)} \right\}.
\]
Moreover, $\lim_{k \to \infty} E[f(x_k)] = f^*$. 

11
Proof. Note that Assumption 5(a,b) imply that (7) holds for a large enough \( k \), say after \( \hat{K} \). Then, since the conditions of Lemma 3 are met, taking expectations on both sides of (7), we obtain for any \( k \geq \hat{K} \):

\[
E[f_{k+1}(x_{k+1}) - f^*] \leq (1 - \lambda_{\min}(\mu_k \gamma_k))E[f_k(x_k) - f^*] + \frac{\lambda_{\min}\text{dist}^2(x_0, X^*)}{2} \mu_k^2 \gamma_k + \frac{(L + \mu_0)\lambda^2 \nu^2}{2} \mu_k^{2\alpha - 2}.
\]

Using Assumption 5(e), we have \( \mu_k^2 \gamma_k \leq \rho \gamma_k^2 \mu_k^{2\alpha} \). Thus, we obtain

\[
E[f_{k+1}(x_{k+1}) - f^*] \leq (1 - \lambda_{\min}(\mu_k \gamma_k))E[f_k(x_k) - f^*] + \frac{\rho \lambda_{\min}\text{dist}^2(x_0, X^*) + (L + \mu_0)\lambda^2 \nu^2}{2} \mu_k^{2\alpha - 2}.
\]

(15)

Let us define \( K := \max\{\hat{K}, K_0\} \). Using the preceding relation and by induction on \( k \), we show the desired result. To show (13), we show the following relation first:

\[
E[f_{k+1}(x_{k+1}) - f^*] \leq \theta \gamma_k \mu_k^{2\alpha - 1}, \quad \text{for all} \quad k \geq K,
\]

Note that (16) is stronger than the relation (13) since we have \( E[f(x_{k+1})] \leq E[f_{k+1}(x_{k+1})] \).

First, we show that (16) holds for \( k = 0 \). Consider the term \( E[f_{K+1}(x_{K+1})] - f^* \).

Multiplying and dividing by \( \gamma_k \mu_k^{2\alpha - 1} \), we obtain

\[
E[f_{K+1}(x_{K+1})] - f^* = \left( \frac{E[f_{K+1}(x_{K+1})] - f^*}{\gamma_k \mu_k^{2\alpha - 1}} \right) \gamma_k \mu_k^{2\alpha - 1} \leq \theta \gamma_k \mu_k^{2\alpha - 1},
\]

where the last inequality is obtained by invoking the definition of \( \theta \) in (14). This implies that (16) holds for \( k = K \). Now assume that (16) holds for some \( k \geq K \). We show that it also holds for \( k + 1 \). From the induction hypothesis and (15) we have

\[
E[f_{k+1}(x_{k+1}) - f^*] \leq (1 - \lambda_{\min}(\mu_k \gamma_k))\theta \gamma_k - 1 \mu_k^{2\alpha - 1} + \frac{\rho \lambda_{\min}\text{dist}^2(x_0, X^*) + (L + \mu_0)\lambda^2 \nu^2}{2} \mu_k^{2\alpha - 2}.
\]

Using Assumption 5(d) we obtain

\[
E[f_{k+1}(x_{k+1}) - f^*] \leq \theta \gamma_k \mu_k^{2\alpha - 1} (1 - \lambda_{\min}(\mu_k \gamma_k))(1 + \beta \lambda_{\min}(\mu_k \gamma_k)) + \frac{\rho \lambda_{\min}\text{dist}^2(x_0, X^*) + (L + \mu_0)\lambda^2 \nu^2}{2} \mu_k^{2\alpha - 2}.
\]

(17)

Next we find an upper bound for the term \( (1 - \lambda_{\min}(\mu_k \gamma_k))(1 + \beta \lambda_{\min}(\mu_k \gamma_k)) \) as follows

\[
(1 - \lambda_{\min}(\mu_k \gamma_k))(1 + \beta \lambda_{\min}(\mu_k \gamma_k)) = 1 - \lambda_{\min}(\mu_k \gamma_k) + \beta \lambda_{\min}(\mu_k \gamma_k) - \beta \lambda_{\min}(\mu_k \gamma_k)^2 \\
\leq 1 - (1 - \beta) \lambda_{\min}(\mu_k \gamma_k).
\]

Combining this relation with (17), it follows

\[
E[f_{k+1}(x_{k+1}) - f^*] \leq \theta \gamma_k \mu_k^{2\alpha - 1} - (\theta(1 - \beta) - \frac{\rho \lambda_{\min}\text{dist}^2(x_0, X^*) + (L + \mu_0)\lambda^2 \nu^2}{2}) \mu_k^{2\alpha - 2}.
\]

\[
= \theta \gamma_k \mu_k^{2\alpha - 1} - \left( \theta(1 - \beta) - \frac{\rho \lambda_{\min}\text{dist}^2(x_0, X^*) + (L + \mu_0)\lambda^2 \nu^2}{2} \right) \mu_k^{2\alpha - 2}.
\]
Note that the definition of $\theta$ in (14) implies that Term 1 is non-negative. Therefore,

$$
E[f_{k+1}(x_{k+1}) - f^*] \leq \theta \gamma_k \mu k^{2\alpha - 1}.
$$

Hence, the induction statement holds for $k + 1$. We conclude that (16) holds for all $k \geq K$. As a consequence, (13) holds for all $k \geq K$ as well. To complete the proof, we need to show $\lim_{k \to \infty} E[f(x_k)] = f^*$. This is an immediate result of (13) and Assumption 5(a).

4. Regularized stochastic and deterministic L-BFGS methods. Note that the (R-SQN) method proposes a general framework that may capture different schemes depending on the choice of the matrix $H_k$. When $H_k$ is the identity matrix for all $k$, (R-SQN) represents the regularized stochastic gradient method. Stochastic gradient methods are known to be sensitive to the choice of stepsizes. In our work [33, 34] we attempted to address this challenge by developing self-tuned stepsizes. Another avenue to enhance the robustness of the scheme is to incorporate the curvature information of the objective function. A well-known updating rule for the matrix $H_k$ that uses the curvature estimates is known as the BFGS update rule. The deterministic BFGS update rule, achieves a superlinear convergence rate (cf. Theorem 8.6 [21]) outperforming the deterministic gradient/subgradient method. In the classical deterministic BFGS scheme, the curvature information is incorporated within the algorithm using two terms: the first term is the displacement factor $s_k = x_{k+1} - x_k$, while the other is the change in the gradient mapping, $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$, where $\nabla f$ denotes the gradient mapping of the deterministic objective function. To have a well-defined update rule, it is essential that at each iteration, the curvature condition (also referred to as the secant condition) $s_k^T y_k > 0$ is satisfied. By maintaining this condition at each iteration, the positive definiteness of the Hessian approximate matrix, denoted by $B_k$, is preserved. The BFGS update rule in deterministic regime also ensures that $B_k$ satisfies a secant equation given by $B_{k+1} s_k = y_k$. This equation implies that the Hessian approximate matrix maps $s_k$ into $y_k$.

To address optimization problems in stochastic settings, a regularized BFGS update rule, namely (RES), is developed for strongly convex optimization problems [17]. In that work, the problem (SO) is considered where it is assumed that the function $f$ is strongly convex and has Lipschitzian gradients. To solve this problem, a regularized stochastic BFGS update rule is proposed as follows:

$$
B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{s_k^T y_k} + \hat{\mu} I,
$$

where the constant $\hat{\mu}$ is a positive scalar that is assumed to be smaller than the strong convexity parameter of the objective function, say $\mu$. Note that, in the RES scheme, the condition $\hat{\mu} < \mu$ is an essential assumption to ascertain the secant condition. RES requires computation of inverse of $B_k$. One of the challenges associated with implementation of this scheme is when the dimension of decision vector $x$, i.e., $n$, is large. This is the case in many machine learning applications such as speech recognition, game classification of videos and Amazon Commerce reviews (see [3]). In these examples, the number of attributes is of the order of thousands, millions, or even more. As a consequence, the computation of the matrix $B_k$ and its inverse become very costly and inefficient. Moreover, at each iteration, a matrix of size $n \times n$ needs to be stored. To address these issues in large-scale optimization problems, limited-memory variants of stochastic BFGS scheme, denoted by stochastic L-BFGS, have been developed [3, 18]. The key idea in L-BFGS update rule is that instead of storing the full
$n \times n$ matrix at each iteration, a fixed number of vectors of size $n$ are stored and used to update the approximate Hessian inverse matrix. This forms a two-loop scheme where in the inner loop, a predetermined number of gradient mappings are evaluated and used to compute $H_k$, and in the outer loop, the solution iterate is updated.

It is worth mentioning that the strong convexity property assumed in [3,18] plays a key role in developing the L-BFGS update rules and establishing the convergence. Throughout this section, our goal is to refine and construct the update rule for matrix $H_k$ in the (R-SQN) scheme, so that not only we are able to establish the convergence of the scheme, but also we ascertain the rate statement based on our analysis in Section 3. Note that in absence of strong convexity, the secant condition does not hold. To address this issue, a standard approach is to employ a damped variant of the BFGS update rule [21]. A drawback of this class of update rules is that there is no guarantee on the rate statements under such rules. Here, we resolve this issue through employing the properties of the regularized gradient map. This is carried out in (R-SQN) by adding the the regularization term $\mu_k(x_k - x_0)$ to the stochastic gradient mapping $\nabla F(x_k, \xi_k)$. If $\mu_k$ is assumed to be constant throughout the scheme, it can be simply seen that the secant condition holds. However, by maintaining $\mu_k$ constant, e.g., setting $\mu_k$ to some scalar $\mu > 0$, the sequence $x_k$ generated by (R-SQN) will converge to an approximate optimal solution, but not a true solution. Another challenge is related to the choice of $\mu$. Choosing a small $\mu$ deteriorates the convergence rate, while a large value creates undesirable deviation from the optimal solution. Our goal is to update the regularization parameter within the scheme and make it decrease to zero in a prescribed way that $x_k$ converges to an optimal solution.

To be able to maintain the secant condition, the main idea is to update the matrix $H_k$ and the parameter $\mu_k$ one at a time. Keeping the regularization parameter constant in one iteration enables us to keep the secant condition satisfied. After we update the matrix $H_k$, in the following iteration, we keep this matrix fixed and drop the value of the regularized parameter. Note that throughout the scheme, we allow for updating the stepsize sequence at each iteration. We update the regularization parameter $\mu_k$ using the following general procedure:

$$
\begin{align*}
\mu_k &= \mu_{k-1}, & \text{if } k \text{ is odd,} \\
\mu_k &= \mu_{k-1}, & \text{otherwise,}
\end{align*}
$$

We assume when $\mu_k < \mu_{k-1}$ rule is employed, $\mu_k$ is chosen such that $\nabla F(x_k, \xi_k) + \mu_k(x_k - x_0) \neq 0$. We construct the update rule in terms of the following two factors defined for any odd $i$:

$$
\begin{align*}
s_i &:= x_i - x_{i-1}, \\
y_i &:= \nabla F(x_i, \xi_{i-1}) - \nabla F(x_{i-1}, \xi_{i-1}) + \mu_i^\delta s_i,
\end{align*}
$$

where $0 < \delta \leq 1$ is a parameter that controls the level of regularization in the matrix $H_k$. It is worth emphasizing that there are two instances of regularization in the (R-SQN) scheme: one in the stochastic gradient map, and one in $H_k$. The latter is initiated by the regularization in $y_i$. Here, $\delta$ only controls the regularization for matrix $H_k$, but not that of the gradient direction. It is assumed that $\delta > 0$ to make

\footnote{Note that we can ensure this non-zero condition, as follows. If we have $\nabla F(x_k, \xi_k) + \mu_k(x_k - x_0) = 0$ and $x_k \neq 0$, then by replacing $\mu_k$ with a smaller value the relation will hold. If $x_k = x_0$, then we can draw a new sample of $\xi_k$ to satisfy the relation.}
Assumption 1 stated as follows: obtaining the pair \((\nabla F(x), y)\) and establishing the secant condition. (ii) Second, instead of \(O(n)\) of the order \(n\) versus the full-memory variants where the computational effort per iteration becomes \(O(mn)\), for any \(k \geq 2m - 1\), is updated using the following recursive formula:

\[
H_k := \begin{cases} 
H_{k,m}, & \text{if } k \text{ is odd}, \\
H_{k-1}, & \text{otherwise}, 
\end{cases}
\]

where \(m < n\) (in the large scale settings, \(m \ll n\)) is a fixed integer that determines the number of pairs \((x_i, y_i)\) to be used to estimate \(H_k\). Matrix \(H_{k,m}\), for any \(k \geq 2m - 1\), is updated using the following recursive formula:

\[
H_{k,j} := \left( I - \frac{y_i s_i^T}{y_i^T s_i} \right)^T H_{k,j-1} \left( I - \frac{y_i s_i^T}{y_i^T s_i} \right) + \frac{s_i s_i^T}{y_i^T s_i}, \quad i := k - 2(m - j), \quad 1 \leq j \leq m,
\]

where we set \(H_{k,0} := \frac{y_i s_i^T}{y_i^T s_i} I\). Here, at odd iterations, matrix \(H_k\) is obtained recursively from \(H_{k,0}, H_{k,1}, \ldots, H_{k,m-1}\). Note that through a calculation of each \(H_k\) at an odd \(k\), \(m\) pairs \((x_i, y_i)\) are used. More precisely, \(H_k\) uses the following curvature information

\[
\{(s_i, y_i) \mid i = k - 2m + 2, k - 2m + 4, \ldots, k \}\.
\]

For convenience, in the first \(2m - 2\) iterations, we use the identity matrix as \(H_k\). Doing so, we can collect the first \(m\) set of pairs \((x_i, y_i)\) where \(i = 1, 3, \ldots, 2m - 1\) used to obtain \(H_{2m-1}\). There are two main differences between update rule (21) and the rules considered in [3, 18]: (i) The first distinction is with respect to the definition of \(y_i\) in (19). Here the term \(\mu_i s_i\) is to compensate for the lack of strong monotonicity of the gradient mapping and establishing the secant condition. (ii) Second, instead of obtaining the pair \((s_i, y_i)\) at every iteration, we calculate these at only odd iterations to allow for updating the regularization parameter satisfying (18). It is worth mentioning that \(H_k \nabla F_k(x_k, \xi_k)\) at the \(k\)th iteration of (R-SQN) can be calculated through a two-loop recursion with \(\mathcal{O}(nm)\) number of operations (see Ch. 7, Pg. 178 in [21]). This is an important computational advantage of the limited-memory BFGS schemes versus the full-memory variants where the computational effort per iteration becomes of the order \(\mathcal{O}(n^2)\). In our analysis in this section, we consider a stronger variant of Assumption 1 stated as follows:

**Assumption 6.** (a) The function \(F(x, \xi)\) is convex over \(\mathbb{R}^n\) for any \(\xi \in \Omega\).

(b) For any \(\xi \in \Omega\), \(F(\cdot, \xi)\) is continuously differentiable with Lipschitz continuous gradients over \(\mathbb{R}^n\) with parameter \(L_\xi > 0\). Moreover, \(L := \sup_{\xi \in \Omega} L_\xi < \infty\).

(c) The optimal solution set \(X^*\) of problem (SO) is nonempty.

In the following result, we derive the lower and upper bounds for the eigenvalues of the matrix \(H_k\) and show that both the secant condition and the secant equation hold.

**Lemma 6** (Properties of update rule (20)-(21)). Consider the (R-SQN) method. Let \(H_k\) be given by the update rule (20)-(21), where \(s_i\) and \(y_i\) are defined in (19) and \(\mu_k\) is updated according to the procedure (18). Let Assumption 6(a,b) hold. Then, the following results hold:

(a) For any odd \(k > 2m\), the secant condition holds, i.e., \(s_k^T y_k > 0\).

(b) For any odd \(k > 2m\), the secant equation holds, i.e., \(H_k y_k = s_k\).
(c) For any $k > 2m$, $H_k$ satisfies Assumption 3 for the following values:

$$\lambda_{\text{min}} = \frac{1}{(m+n)(L+\mu_0^\delta)}; \quad \lambda = \sqrt{(m+n)^{n+m-1}(L+\mu_0^\delta)^{n+m-1}},$$

$$\alpha = -\frac{(n+m)\delta}{n}.$$  

More precisely, $H_k$ is symmetric, $E[H_k | F_k] = H_k$ and

$$\frac{1}{(m+n)(L+\mu_0^\delta)} I \preceq H_k \preceq \lambda I.$$  

Proof. It can be seen, by induction on $k$, that $H_k$ is symmetric and $F_k$ measurable, assuming that all matrices are well-defined. We use induction on odd values of $k > 2m$ to show that the statements of part (a), (b) and (c) hold and that the matrices are well-defined. Suppose $k > 2m$ is odd and for any odd value of $t < k$, we have $s_t^T y_t > 0$, $H_t y_t = s_t$, and (23) for $t$. We show that these statements also hold for $k$. First, we prove that the secant condition holds. We can write

$$s_k^T y_k = (x_k - x_{k-1})^T (\nabla F(x_k, \xi_{k-1}) - \nabla F(x_{k-1}, \xi_{k-1}) + \mu_0^\delta (x_k - x_{k-1}))$$

$$= (x_k - x_{k-1})^T (\nabla F(x_k, \xi_{k-1}) - \nabla F(x_{k-1}, \xi_{k-1})) + \mu_0^\delta \|x_k - x_{k-1}\|^2$$

$$\geq \mu_k^\delta \|x_k - x_{k-1}\|^2,$$

where the inequality follows from the monotonicity of the gradient map $\nabla F(\cdot, \xi)$. From the induction hypothesis, $H_{k-2}$ is positive definite, since $k-2$ is odd. Furthermore, since $k-2$ is odd, we have $H_{k-1} = H_{k-2}$ by the update rule (20). Therefore, $H_{k-1}$ is positive definite. Note that since $k-2$ is odd, the choice of $\mu_{k-1}$ is such that $\nabla F(x_{k-1}, \xi_{k-1}) + \mu_{k-1} x_{k-1} \neq 0$ (see the discussion following (18)). Since $H_{k-1}$ is positive definite, we have $H_{k-1} (\nabla F(x_{k-1}, \xi_{k-1}) + \mu_{k-1} x_{k-1}) \neq 0$, implying that $x_k \neq x_{k-1}$. Hence $s_k^T y_k \geq \mu_k^\delta \|x_k - x_{k-1}\|^2 > 0$, where the second inequality is a consequence of $\mu_k > 0$. Thus, the secant condition holds. Next, we show that (23) holds for $k$. Let $\lambda_{k, \text{min}}$ and $\lambda_{k, \text{max}}$ denote the minimum and maximum eigenvalues of $H_k$, respectively. Also, let us denote the inverse of matrix $H_k$ in (21) by $B_k$. It is well-known that using the Sherman-Morrison-Woodbury formula, $B_k$ is equal to $B_{k,m}$ given by

$$B_{k,j} = B_{k,j-1} - \frac{B_{k,j-1}s_i s_i^T B_{k,j-1}}{s_i^T B_{k,j-1} s_i} + \frac{y_i y_i^T}{y_i^T s_i}, \quad i := k-2(m-j), \quad 1 \leq j \leq m,$$

where $s_i$ and $y_i$ are defined by (19) and $B_{k,0} = \frac{y_k y_k^T}{s_k^T y_k} I$. First, we show that for any $i$,

$$\mu_k^\delta \leq \frac{\|y_i\|^2}{y_i^T s_i} \leq L + \mu_k^\delta,$$

where $L$ is the Lipschitzian parameter of the gradient mapping $\nabla F$. Let us consider the function $h(x) := F(x, \xi_{i-1}) + \mu_k^\delta \|x\|^2$ for fixed $i$ and $k$. Note that this function is strongly convex and has a gradient mapping of the form $\nabla F + \mu_k^\delta I$ that is Lipschitz with parameter $L + \mu_k^\delta$. For a convex function $h$ with Lipschitz gradient with parameter $L + \mu_k^\delta$, the following inequality, referred to as co-coercivity property, holds for any $x_1, x_2 \in \mathbb{R}^n$ (see [22], Pg. 24, Lemma 2):

$$\|\nabla h(x_2) - \nabla h(x_1)\|^2 \leq (L + \mu_k^\delta)(x_2 - x_1)^T (\nabla h(x_2) - \nabla h(x_1)).$$
Substituting $x_2$ by $x_1$, $x_1$ by $x_{i-1}$, and recalling (19), the preceding inequality yields

\[(26) \quad \|y_i\|^2 \leq (L + \mu_k^0)s_i^T y_i.\]

Note that function $h$ is strongly convex by parameter $\mu_k^0$. Applying the Cauchy-Schwarz inequality, we can write

\[
\frac{\|y_i\|^2}{s_i^T y_i} \geq \frac{\|y_i\|^2}{\|s_i\| \|y_i\|} = \frac{\|y_i\|}{\|s_i\|} \geq \frac{y_i^T s_i}{\|s_i\|^2} \geq \mu_k^0.
\]

Combining this relation with (26), we obtain (25). Next, we show that the maximum eigenvalue of $B_k$ is bounded. Let $\text{Trace}(\cdot)$ denote the trace of a matrix. Taking trace from both sides of (24) and summing up over index $j$, we obtain

\[
\text{Trace}(B_{k,m}) = \text{Trace}(B_{k,0}) - \sum_{j=1}^{m} \text{Trace} \left( \frac{B_{k,j-1}s_i s_i^T B_{k,j-1}}{s_i^T B_{k,j-1} s_i} \right) + \sum_{j=1}^{m} \text{Trace} \left( \frac{y_i y_i^T}{y_i^T s_i} \right)
\]

\[
= \text{Trace} \left( \|y_i\|^2 I \right) - \sum_{j=1}^{m} \frac{\|B_{k,j-1}s_i\|^2}{s_i^T B_{k,j-1} s_i} + \sum_{j=1}^{m} \frac{\|y_i\|^2}{y_i^T s_i}
\]

\[(27) \quad \leq n \frac{\|y_i\|^2}{y_i^T s_i} + \sum_{j=1}^{m} (L + \mu_k^0) = (m + n)(L + \mu_k^0),
\]

where the third relation is obtained by positive-definiteness of $B_k$ (this can be seen by induction on $k$, and using (24) and $B_{k,0} > 0$). Since $B_k = B_{k,m}$, the maximum eigenvalue of the matrix $B_k$ is bounded. As a result,

\[(28) \quad \lambda_{k,\text{min}} \geq \frac{1}{(m + n)(L + \mu_k^0)}.\]

In the next part of the proof, we establish the bound for $\lambda_{k,\text{max}}$. The following relation can be shown (e.g., see Lemma 3 in [18])

\[
det(B_{k,m}) = det(B_{k,0}) \prod_{j=1}^{m} \frac{s_j^T y_i}{s_j^T B_{k,j-1} s_i}.
\]

Multiplying and dividing by $s_i^T s_i$, using the strong convexity of the function $h$, and invoking (25) and the result of (27), we obtain

\[
det(B_k) = det \left( \frac{y_i^T y_k}{s_k^T y_k} \right) \prod_{j=1}^{m} \left( \frac{s_j^T y_i}{s_j^T s_i} \left( \frac{s_j^T s_i}{s_j^T B_{k,j-1} s_i} \right) \right)
\]

\[
\geq \left( \frac{y_i^T y_k}{s_k^T y_k} \right) \prod_{j=1}^{m} \mu_k \left( \frac{s_j^T s_i}{s_j^T B_{k,j-1} s_i} \right)
\]

\[(29) \quad \geq (\mu_k)^{(n+m)\delta} \prod_{j=1}^{m} \frac{1}{(m + n)(L + \mu_k^0)} = \frac{\mu_k^{(n+m)\delta}}{(m + n)^m(L + \mu_k^0)^m}.
\]

Let $\alpha_{k,1}, \alpha_{k,2}, \ldots, \alpha_{k,n}$ be the eigenvalues of $B_k$ sorted non-decreasingly. Note that since $B_k > 0$, all the eigenvalues are positive. Also, from (27), we know that $\alpha_{k,\ell} \leq \frac{1}{(m + n)(L + \mu_k^0)} \geq \frac{1}{\mu_k^0}$. Therefore, the maximum eigenvalue $\alpha_{k,max}$ is bounded by

\[
\alpha_{k,\text{max}} \leq \frac{1}{\mu_k^0}.
\]
\((m + n)(L + \mu_0^\delta)\). Therefore
\[
\prod_{\ell=1}^{n} \alpha_{k,\ell} \leq \alpha_{k,1}^{}(m + n)^{n-1}(L + \mu_0^\delta)^{n-1}.
\]
From this relation and (29), we obtain
\[
\begin{align*}
\alpha_{1,k} & \leq \frac{\mu_k^{(n+m)^\delta}}{(m + n)^{n+m-1}(L + \mu_k^\delta)^{n+m-1}}. \\
\end{align*}
\]
This relation and that \(\alpha_{k,1} = \lambda_{k,\text{max}}^{-1}\) imply that
\[
\lambda_{k,\text{max}} \leq \sqrt{(m + n)^{n+m-1}(L + \mu_k^\delta)^{n+m-1}}.
\]
Therefore, from (28) and (30) and that \(\mu_k^\delta\) is non-increasing, we conclude that (23) holds for any odd \(k > 1\), and using (24), for \(j = m\) we obtain
\[
B_{k,m} = B_{k,m-1} - \frac{B_{k,m-1}s_k^TB_{k,m-1} s_k}{s_k^TB_{k,m-1}s_k} + \frac{y_k y_k^T}{y_k^T s_k},
\]
where we used \(i = k - 2(m - m) = k\). Multiplying both sides of the preceding equation by \(s_k\), and using \(B_k = B_{k,m}\), we have \(B_k s_k^T = B_{k,m-1}s_k - B_{k,m-1}s_k + y_k = y_k\). Multiplying both sides of the preceding relation by \(H_k\) and invoking \(H_k = B_k^{-1}\), we conclude that \(H_k y_k = s_k\). Therefore, we showed that the statements of (a), (b), and (c) hold for \(k\), assuming that they hold for any odd \(2m < t < k\). In a similar fashion to this analysis, it can be seen that the statements hold for \(t = 2m + 1\). Thus, by induction, we conclude that the statements hold for any odd \(k > 2m\). To complete the proof, it is enough to show that (23) holds for any even value of \(k > 2m\). Let \(t = k - 1\). Since \(t > 2m\) is odd, relation (23) holds. Writing (23) for \(k - 1\), and taking into account that \(H_k = H_{k-1}\), and \(\mu_k < \mu_{k-1}\), we can conclude that (23) holds for any even value of \(k > 2m\) and this completes the proof.

**Lemma 7** (Feasible tuning sequences for a.s. convergence). Let the sequences \(\gamma_k\) and \(\mu_k\) be given by the following rules:
\[
\begin{align*}
\gamma_k & = \frac{\gamma_0}{(k + 1)^a}, \\
\mu_k & = \frac{\mu_02^b}{(k + \kappa)^b},
\end{align*}
\]
where \(\kappa = 2\) if \(k\) is even and \(\kappa = 1\) otherwise, \(\gamma_0^\text{and} \mu_0^\text{are positive scalars such that} \gamma_0 \leq L(m + n), \text{and} a \text{and} b \text{are positive scalars that satisfy the following conditions:}
\[
\frac{a}{b} > 2 \left(\frac{n + m}{n}\right), \quad a + b \leq 1, \quad a + 2b > 1, \quad a - \left(\frac{m + n}{n}\right) > 0.5.
\]
Then, for any \(0 < \delta \leq 1\), the sequences \(\gamma_k\) and \(\mu_k\) satisfy Assumption 4 with \(\lambda_{\text{min}}\) and \(\alpha\) given by (22). Also, \(\mu_k\) satisfies (18).

**Proof.** In the following, we show that the presented class of sequences satisfy each of the conditions listed in Assumption 4. Throughout, we let \(\alpha\) denote \(-\frac{(m + n)\delta}{n}\).

(a) Replacing the sequences by their given rules, we obtain
\[
\begin{align*}
\gamma_k \mu_k^{2n} & = \frac{\gamma_0}{(2^b \mu_0)^{-2a}}(k + 1)^{-a}(k + \kappa)^{-2ab} \leq \frac{\gamma_0}{(2^b \mu_0)^{2a}}(k + 1)^{-a-2ab}.
\end{align*}
\]
Since we assumed \(\frac{a}{b} > 2 \left(\frac{m + n}{n}\right)\), for any \(\delta \in (0, 1]\), we have \(\frac{a}{b} > 2 \left(\frac{m + n}{n}\right)\) implying that \(-a - 2ab < 0.\) Thus, the preceding term goes to zero verifying Assumption 4(a).
(b) Let $k$ be an even number. Thus, $\kappa = 2$. From (31) we have $\mu_k = \mu_{k+1} = \frac{\mu_0 \gamma^b}{(k+2)^b}$. Now, let $k$ be an odd number. Again, according to (31) can write

$$
\mu_{k+1} = \frac{\mu_0 \gamma^b}{((k+1) + 2)^b} < \frac{\mu_0 \gamma^b}{(k+1)^b} = \frac{\mu_0 \gamma^b}{(k+\kappa)^b} = \mu_k.
$$

Therefore, $\mu_k$ given by (31) satisfies (18). Also, from (31) we have $\mu_k \to 0$. Thus, Assumption 4(b) holds.

(c) The given rules (31) imply that $\gamma_k$ and $\mu_k$ are both non-increasing sequences. Therefore, we have $\gamma_k \mu_k \leq \gamma_0 \mu_0$ for any $k \geq 0$. So, to show that Assumption 4(c) holds, it is enough to show that $\lambda_{\min} \gamma_0 \mu_0 \leq 1$ where $\lambda_{\min}$ is given by (22). Since we assumed that $\gamma_0 \mu_0 \leq L(m+n)$, for any $\delta \in (0,1]$, we have $\gamma_0 \mu_0 \leq (m+n)(L + \mu_0^2)$, implying that $\lambda_{\min} \gamma_0 \mu_0 \leq 1$ and that Assumption 4(c) holds.

(d) From (31), we can write

$$
\sum_{k=0}^{\infty} \gamma_k \mu_k = \gamma_0 \mu_0 \sum_{k=0}^{\infty} (k + 1)^{-a} (k + \kappa)^{-b} \geq \gamma_0 \mu_0 \gamma^b \sum_{k=0}^{\infty} (k + 2)^{-(a+b)} = \infty,
$$

where the last relation is due to $a + b \leq 1$. Therefore, Assumption 4(d) holds.

(e) Using (31), it follows

$$
\sum_{k=0}^{\infty} \gamma_k \mu_k^2 = \gamma_0 \mu_0 \sum_{k=0}^{\infty} (k + 1)^{-a} (k + \kappa)^{-2b} \leq \gamma_0 \mu_0 \gamma^b \sum_{k=0}^{\infty} (k + 1)^{-(a+2b)} < \infty,
$$

where the last inequality is due to $a + 2b > 1$. Therefore, Assumption 4(e) holds.

(f) From (31), we have

$$
\sum_{k=0}^{\infty} \gamma_k \mu_k^{2a} \leq \gamma_0 \mu_0 \gamma^b \sum_{k=0}^{\infty} (k + \kappa)^{-2ab} \left( \sum_{k=0}^{1} \frac{(k + \kappa)^{-ab}}{(k+1)^{2a}} + \sum_{k=2}^{\infty} \frac{(2k)^{-2ab}}{k^{2a}} \right) < \infty,
$$

where the first inequality is due to $\alpha < 0$ and the last inequality is due to $a + \alpha b > a + \frac{a}{\beta} > 0.5$. Therefore, Assumption 4(f) is verified.

**Remark** 3 (An example for feasible sequences). Note that, in Lemma 7, the conditions on parameters $a$, $b$, $\gamma_0$, and $\mu_0$ are not restrictive. In fact, if we choose $m < \frac{n}{2}$, and $\gamma_0 = \mu_0 \leq \sqrt{L}$, it can be simply seen that $a = \frac{2}{6}$ and $b = \frac{1}{6}$ satisfy the conditions of Lemma 7.

**Lemma 8** (Feasible tuning sequences for convergence in mean). Let the sequences $\gamma_k$ and $\mu_k$ be given by (31), where $\gamma_0$ and $\mu_0$ are positive scalars such that $\gamma_0 \mu_0 \leq L(m+n)$. Let $m, 0 < \delta < 1, a > 0$ and $b > 0$ satisfy the following conditions:

$$
\frac{a}{b} > 1 + 2\delta \left( \frac{m+n}{n} \right), \quad a + b < 1, \quad a \leq 2b \left( 1 + \delta \left( \frac{m+n}{n} \right) \right).
$$

Then, there exists $K > 0$ such that for any $k \geq K$, the sequences $\gamma_k$ and $\mu_k$ satisfy Assumption 5 with an arbitrary $0 < \beta < 1$ and for $\rho = (\mu_0 \gamma^b)^{2+2b(m+n)} \gamma_0^{-1}$, and with $\lambda_{\min}$, $\lambda$ and $\alpha$ given by (22). Also, $\mu_k$ satisfies (18).

**Proof.** In the following, we verify the conditions of Assumption 5.

(a) Note that the relation $\frac{a}{b} > 1 + 2\delta \left( \frac{m+n}{n} \right)$ is equivalent to $a > b(1-2\alpha)$. Using this
and the update rules of $γ_k$ and $μ_k$, it can be easily seen that $γ_kμ_k^{2α−1} → 0$ verifying Assumption 5(a). Conditions (b), (c) and (18): This is already shown in parts (b, c) of the proof of Lemma 7.

(d) It suffices to show there exists $K_1$ such that for any $k ≥ K_0$ and $0 < β < 1$, we have

\[ \gamma_{k−1} \frac{μ_k^{1−2α}}{μ_k^{1−2α}} − 1 ≤ βλ_{min}γ_kμ_k. \]

From (31) and that $α = −\frac{(m+n)δ}{n} < 0$, we obtain

\[ \frac{γ_{k−1}μ_k^{1−2α}}{γ_kμ_k^{1−2α}} − 1 ≤ \frac{γ_{k−1}}{γ_k} − 1 = \left(1 + \frac{\alpha}{κ}\right) − 1 = 1 + \frac{a(\frac{1}{κ})}{κ} = O\left(\frac{1}{κ}\right), \]

where the first inequality is implied due to $μ_k$ being a non-increasing sequence, and in the second equation we used the Taylor’s expansion of $\left(1 + \frac{\alpha}{κ}\right)^a$. Therefore, since the right-hand side of the relation (32) is of the order $\frac{1}{κ}$ and that $a + b < 1$, the preceding inequality shows that such $K_0$ exists for which Assumption 5(d) holds for all $0 < β < 1$.

(e) From (31), we have

\[ \frac{μ_k^{2α−1}}{γ_k} = δ_0^{-1} (μ_0^{2b})^{2α−1}(k + κ)^{−b(2−2α)}(k + 1)^a ≤ δ_0^{-1} (μ_0^{2b})^{2α−1} ≤ \delta_0^{-1} (μ_0^{2b})^{2α−1} = ρ, \]

where the first inequality is due to the assumption $−a + 2(1 − α)b ≥ 0$. Therefore, Assumption 5(e) is satisfied.

We conclude that the desired results hold for $K := K_0$. □

The following result, provides a set of feasible values for $a$, $b$ and ranges for $δ$, and $m$ satisfying the conditions of Lemma 8. Moreover, we establish the convergence rate of the regularized stochastic L-BFGS scheme.

**THEOREM 9.** [Rate of convergence of the regularized stochastic L-BFGS method]
Consider the (R-SQN) method where $H_k$ is given by the update rule (20)-(21). Suppose Assumptions 2 and 6 hold. Let $0 < γ < \frac{1}{3}$ be an arbitrary scalar and let $δ ∈ \left(0, \frac{\sqrt{m}}{\sqrt{n+\frac{m}{n}}}\right)$. Let the sequences $γ_k$ and $μ_k$ be given by (31) with $a = \frac{2}{3} − ε + \frac{2δ(n+m)}{3n}$, $b = \frac{1}{3}$, and $γ_0μ_0 ≤ L(n+m).$ Then, there exists $K > 2m$ such that

\[E[f(x_k)] − f^∗ ≤ \left(\frac{θγ_0}{(μ_0^{3/2})^{1−2α}}\right) \frac{1}{K^{1/2−ε}}, \quad \text{for all } k > K,\]

where $θ$ is given by (14), and $λ_{min}$, $λ$, and $α$ are given by (22).

**Proof.** First, we show that the conditions of Theorem 5 are satisfied. Assumption 1 holds as a consequence of Assumption 6. From Lemma 6(c), Assumption 3 holds for any $k > 2m$ as well. To show that Assumption 5 holds, we apply Lemma 8. We have

\[ \frac{a}{b} = \frac{2}{3} − ε + \frac{2δ(n+m)}{3n} = 2 − 3ε + \frac{2δ(n+m)}{n} > 1 + 2δ \left(\frac{m+n}{n}\right), \]

20
where we used $\epsilon < \frac{1}{5}$. Moreover, since $\delta < \frac{1.5\rho a}{n+m}$, we have $a + b = 1 - \epsilon + \frac{2\delta(n+m)}{3n} < 1$. Also, from the values of $a$ and $b$ we have

$$2b \left(1 + \delta \left(\frac{m+n}{n}\right)\right) = \frac{2}{3} - \frac{2\delta(n+m)}{3n} = a + \epsilon > a.$$  

We conclude that the conditions of Lemma 8 hold. This implies that there exists $K_0 > 0$ such that for any $k \geq K_0$, the sequences $\gamma_k$ and $\mu_k$ satisfy Assumption 5 with any arbitrary $0 < \beta < 1$ and for $\rho = (\mu_0 2^k)^{2 + \frac{2(m+n)d}{n}} \gamma_0^{-1}$, and with $\lambda_{\min}$, $\lambda$ and $\alpha$ given by (22). Let us define $K := \max\{K_0, 2m\}$. Because all conditions of Theorem 5 are satisfied, from (13), (31), and substituting values of $a$, $b$, and $\alpha$, for any $k \geq K$ we obtain

$$E[f(x_{k+1})] - f^* \leq \theta \gamma_{k-1} \mu_k^{2\alpha - 1} \leq \frac{\theta \gamma_0 (k + 1)^{(1 - 2\alpha)/3}}{(\mu_0 \sqrt{2})^{1 - 2\alpha} (k + 1)^{2/3 - \epsilon - \frac{3}{2} \alpha}} = \left(\frac{\theta \gamma_0}{(\mu_0 \sqrt{2})^{1 - 2\alpha}}\right) \frac{1}{(k + 1)^{\frac{3}{2} - \epsilon}}.$$  

Through a change of variable from $k + 1$ to $k$, we conclude the result.

Our goal in the remainder of this section lies in establishing the convergence and rate statement for the deterministic L-BFGS scheme. Consider the following regularized deterministic L-BFGS method:

(R-L-BFGS) \hspace{1cm} x_{k+1} := x_k - \gamma_k H_k (\nabla f(x_k) + \mu_k x_k), \hspace{1cm} \text{for all } k \geq 0,

where $H_k$ is given by the update rule (20), $\mu_k$ is updated according to (18) and

$$s_i := x_i - x_{i-1},$$

$$y_i := \nabla f(x_i) - \nabla f(x_{i-1}) + \mu_i^s s_i.$$  

\hspace{1cm} (33)

**Theorem 10** (Convergence and rate analysis of regularized deterministic L-BFGS method). Let $x_k$ be generated by the R-L-BFGS method. Suppose Assumption 1 holds. Let $\lambda_{\min}$, $\lambda$ and $\alpha$ be given by (22). Then,

(a) Let $\mu_k$ satisfies (18). If $\gamma_k$ and $\mu_k$ satisfy the following relation:

$$\gamma_k \mu_k^{2\alpha} \leq \frac{\lambda_{\min}}{\lambda^2 (L + \mu_0)}, \hspace{1cm} \text{for all } k \geq 0,$$

then, for any $k \geq 0$, we have

$$f_{k+1}(x_{k+1}) - f^* \leq (1 - \lambda_{\min} \mu_k \gamma_k)(f_k(x_k) - f^*) + \frac{\lambda_{\min} \text{dist}^2(x_0, X^*)}{2} \mu_k^2 \gamma_k.$$  

(b) Let $\gamma_k$ and $\mu_k$ be given by the update rule (31) where $a > 0$ and $b > 0$ satisfy

$$\frac{a}{b} > 2 \left(\frac{n+m}{n}\right), \hspace{1cm} a + b \leq 1, \hspace{1cm} a + 2b > 1.$$  

Then, $\lim_{k \to \infty} f(x_k) = f^*$ for any $\delta \in (0, 1]$. Specifically, for $a = \frac{4}{5}$ and $b = \frac{1}{5}$ and $m < n$, this result holds.
(c) Let $\epsilon \in (0, 1)$ be an arbitrary given small scalar. Let $\gamma_k$ and $\mu_k$ be given by the update rule (31) where $a = \epsilon$, $b = 1 - \epsilon$. Also, assume $\delta \in \left(0, \frac{\epsilon}{2(\alpha + m)(1 - \epsilon)}\right)$. Let $\gamma_0$ and $\mu_0$ satisfy the following condition:

$$\gamma_0 \mu_0 \geq (n + m)(L + \mu_0^2).$$

Then, there exists $K$ such that

$$f(x_k) - f^* \leq \frac{\Gamma}{(k + 1)^{1-\epsilon}}, \quad \text{for all } k \geq K,$$

where $\Gamma := \max \left\{ (K + 1)^{1-\epsilon} (f_k(x_k) - f^*), \frac{\lambda_{\text{max}} \mu_0^2 \delta b \text{dist}_0^2(x_0, X^*)}{4(\lambda_{\text{min}} \gamma_0 \mu_0 - b)} \right\}$.

**Proof.** (a) The conditions of Lemma 6 are met indicating that Assumption 6 holds. Assumption 2 is clearly met with $\nu = 0$ as the problem is deterministic. Therefore, all of the conditions of Lemma 3 are satisfied and thus (7) holds. Substituting $\nu = 0$ in (7) and eliminating the expectation operator yields the desired inequality.

(b) First, we show that (35) holds. We can write

$$\gamma_k \mu_k^{2a} = \frac{\gamma_0}{(2b \mu_0)^{2a}} (k + 1)^{-a} (k + \kappa)^{-2ab} \leq \frac{\gamma_0}{(2b \mu_0)^{2a}} (k + 1)^{-a - 2ab}.$$

Note that for any $\delta \in (0, 1]$, we have $\frac{\alpha}{\beta} > 2 \left( \frac{\alpha + m}{n} \right)$ implying that $-a - 2ab < 0$. Therefore, $\gamma_k \mu_k^{2a} \to 0$ showing that there exists $K_0$ such that for any $k \geq K_0$, (35) holds. We apply Lemma 2 to the inequality (35) by setting $\alpha_k := \lambda_{\text{min}} \gamma_0 \mu_0, \beta_k := \frac{\lambda_{\text{max}} \mu_0 \delta b \text{dist}_0^2(x_0, X^*)}{2(\alpha \mu_0)}, v_k := f_k(x_k) - f^*$. Since $a + b \leq 1$, we have $\sum_{k=0}^{\infty} \beta_k = \infty$. Also, $a + 2b > 1$ indicates that $\sum_{k=0}^{\infty} \beta_k < \infty$. Since all conditions of Lemma 2 are met, we have $f_k(x_k) \to f^*$. Recalling Definition 1, this implies that $f(x_k) \to f^*$.

(c) First, we show that by the given update rules for $\gamma_k$ and $\mu_k$, relation (35) holds. Note that $\alpha = -\delta \frac{a + m}{n}$. Therefore, we can write

$$\gamma_k \mu_k^{2a} = \frac{\gamma_0 (k + \kappa)^{\frac{a + m}{n} \delta b}}{(\mu_0^{2a})^\delta (k + 1)^a} \leq \frac{\gamma_0 (k + 2)^{\frac{a + m}{n} \delta b}}{(\mu_0^{2a})^\delta (k + 1)^a},$$

$$= \frac{\gamma_0 (1 + \frac{1}{k+1})^{2 \frac{a + m}{n} \delta b} (k + 1)^a}{(\mu_0^{2a})^\delta (k + 1)^a} - \frac{\delta b (m + n)}{2a}.$$

Using the condition on $\delta$, we have $a - \frac{2b \delta (m + n)}{n} = \epsilon - \frac{2(1 - \epsilon) \delta (m + n)}{n} > 0$. Thus, relation (38) indicates that there exists $K_1$ such that for any $k \geq K_1$, (35) holds. Besides, since $a$ and $b$ are positive, there exits $K_2$ such that for any $k \geq K_2$, we have $1 - \lambda_{\text{min}} \gamma_k \mu_k > 0$. Let us now define $K := \max\{K_1, K_2, 2m\}$. Next, we use induction on $k$ to show (37). For $k = K$, it clearly holds. Let us assume (37) holds for $k > K$. Let $e_k$ denote $f_k(x_k) - f^*$. From (35) and the update rules of $\gamma_k$ and $\mu_k$ we can write

$$e_k \leq \left( 1 - \frac{\lambda_{\text{min}} \gamma_0 \mu_0}{k^a (k + \kappa - 1)^b} \right) e_{k-1} + \frac{\lambda_{\text{min}} \mu_0 \text{dist}_0^2(x_0, X^*) \gamma_0 \mu_0^{2a} \delta b}{k^a (k + \kappa - 1)^{2b}}$$

$$\leq \left( 1 - \frac{\lambda_{\text{min}} \gamma_0 \mu_0 \delta b}{k^a (k + \kappa)^b} \right) e_{k-1} + \frac{\lambda_{\text{min}} \mu_0 \text{dist}_0^2(x_0, X^*) \gamma_0 \mu_0^{2a} \delta b}{k^a (k + 1)^{2b}},$$

$$\leq \left( 1 - \frac{\lambda_{\text{min}} \gamma_0 \mu_0 \delta b}{k} \right) e_{k-1} + \frac{\lambda_{\text{min}} \mu_0 \text{dist}_0^2(x_0, X^*) \gamma_0 \mu_0^{2a} \delta b}{k^a (k + 1)^{2b}},$$

\begin{align}
&\leq \left( 1 - \frac{\lambda_{\text{min}} \gamma_0 \mu_0 \delta b}{k} \right) e_{k-1} + \frac{\lambda_{\text{min}} \mu_0 \text{dist}_0^2(x_0, X^*) \gamma_0 \mu_0^{2a} \delta b}{k^a (k + 1)^{2b}} \tag{39}.
\end{align}
where $\kappa$ is defined in (31), and the last inequality is implied by $\frac{k^n(k+1)^b}{k^{a+b}} \leq 2^b$ for $k \geq 1$. Note that since $k \geq K_2$, the term $\left(1 - \frac{\lambda_{\min} \gamma_0 \mu_0}{k^{a+1}}\right)$ in (39) is non-negative. Therefore, we can replace $e_{k-1}$ by its upper bound $\frac{\Gamma}{k}$ in (39). Doing so and noticing that $a + b = 1$, we obtain

$$
(40) \quad e_k \leq \left(1 - \frac{C_1}{k}\right) \frac{\Gamma}{k^b} + \frac{C_2}{k^{b+1}},
$$

where we used the following definitions

$$
C_1 := \lambda_{\min} \gamma_0 \mu_0, \quad C_2 := \lambda_{\min} \text{dist}^2(x_0, X^*) \gamma_0 \mu_0^2 2^{2b-1}.
$$

Using (40), to show that $e_k \leq \frac{\Gamma}{(k+1)^{b+1}}$, it is enough to show that

$$
\Gamma \left(\frac{1}{k^b} - \frac{1}{(k+1)^b}\right) \leq \frac{C_1 \Gamma - C_2}{k^{b+1}}.
$$

Rearranging the terms, we need to verify that $\Gamma \geq \frac{C_2}{C_1 - C_3}$ and $C_3 < C_1$, where

$$
C_3 := \sup_{k \geq 1} \left\{k^{b+1} \left(\frac{1}{k^b} - \frac{1}{(k+1)^b}\right)\right\}.
$$

We claim that $C_3 = b$. To prove this, we need to show that $k^{b+1} \left(\frac{1}{k^b} - \frac{1}{(k+1)^b}\right) \leq b$. Equivalently, we need to verify that

$$
\left(1 - \frac{1}{k+1}\right)^b \geq 1 - \frac{b}{k}, \quad \text{for all } k \geq 1.
$$

Consider the function $g(x) := (1 - \frac{1}{1+x})^b + \frac{b}{x} - 1$ for $x \geq 1$. We have

$$
g'(x) = \frac{b}{(1+x)^2} \left(1 - \frac{1}{1+x}\right)^{1-b} - \frac{b}{x^2} = \frac{b}{(1+x)^2} \left(\frac{x+1}{x}\right)^{1-b} - \left(\frac{x+1}{x}\right)^2 \leq 0,
$$

due to $0 < b < 1$. Hence, $g$ is non-increasing implying that it suffices to show $g(1) \geq 0$, i.e., $2^b(1-b) \leq 1$. Let us define $h(x) := 2^x(1-x)$ for $0 < x < 1$. We have $h'(x) = 2^x \ln(2)(1-x) - 1$. This indicates that $h'(x) < 0$ over $x \in (0, 1)$, implying that $h(b) \leq h(0) = 1$. Hence, we conclude that $C_3 = b$. To show that $C_3 < C_1$ holds, we need to verify that $C_1 > b$. This is true due to (36). To complete the proof we need to show $\Gamma \geq \frac{C_2}{C_1 - b}$. This holds from the definition of $\Gamma$.

5. Numerical experiments. In this section, we present the implementation results of the developed regularized stochastic L-BFGS algorithm to solve a big data application. The Reuters Corpus Volume I (RCV1) data set [11] is a collection of news-wire stories produced by Reuters journalists from 1996 to 1997. The articles are categorized into four hierarchical groups including Corporate/Industrial (CCAT), Economics (ECAT), Government/Social (GCAT) and Markets (MCAT). In this data set, the articles represent the training samples. After the tokenization process, each article is converted to a sparse binary vector, in that 1 denotes the existence and 0 denotes nonexistence of a token in the corresponding article. We consider a subset of the data with $N = 30,000$ articles and $n = 138,921$ tokens. Here we focus our
attention on the binary classification of the articles with respect to the MCAT class. We consider the logistic regression loss minimization (LRM) problem given as follows:

$$ (LRM) \quad \min_{x \in \mathbb{R}^n} f(x) := \frac{1}{N} \sum_{i=1}^{N} \ln \left(1 + \exp \left(-u_i^T x v_i \right) \right), $$

where $u_i \in \mathbb{R}^n$ is the input binary vector associated with article $i$, $v_i \in \{-1, 1\}$ represents the class of the $i$th article (i.e., non-MCAT or MCAT). To address this problem, the SQN methods in [3,18] solve an approximate problem of the form

$$ (\text{Regularized LRM}) \quad \min_{x \in \mathbb{R}^n} f(x) := \frac{1}{N} \sum_{i=1}^{N} \ln \left(1 + \exp \left(-u_i^T x v_i \right) \right) + \frac{\mu}{2} \|x\|^2, $$

where $\mu > 0$ is an arbitrary regularization parameter. Throughout this section, we refer to the stochastic L-BFGS method considered in [3] by $\mu$-S-L-BFGS where $\mu$ can take different values, and refer to the developed regularized stochastic L-BFGS algorithm in this paper by R-S-L-BFGS. Note that in $\mu$-S-L-BFGS, stepsizes of the form $\gamma_k = \frac{\gamma_0}{k+1}$ for $k \geq 0$ are used, while in our scheme, we use stepsizes of the form $\gamma_k = \frac{\gamma_0}{\sqrt{k+1}}$. In this section, $\gamma_0$ denotes the initial stepsize, $\mu_0$ denotes the initial regularization parameter in our scheme, and $m$ denotes the memory parameter in the BFGS schemes, that is the number of pairs of vectors used to estimate the inverse of Hessian. We set the initial point $x_0$ to be the origin in both methods. To compare the two schemes, we consider different settings for $\gamma_0$, $\mu_0$, and $m$. We let $\gamma_0$ take
values of 10, 1, and 0.1, $\mu_0$ take values of 1, 0.1, and 0.01, and $m$ take values of 2 and 5. These settings result in 18 different cases that are presented in Figure 1. In each setting, we run the two schemes for 10,000 number of iterations. The $y$-axis denotes the logarithm of average of the logistic regression loss function. In all cases, we used three sample paths and sample sizes of 10,000 to calculate the average value of the objective function.

We observe that in most of the cases, the R-S-L-BFGS method outperforms the $\mu$-S-L-BFGS method. Only the setting with $\mu_0 = 1$, and $\gamma_0 = 0.1$, and the setting where $\mu_0 = 0.1$, and $\gamma_0 = 0.1$, the $\mu$-S-L-BFGS method are slightly better. Recall that the $\mu$-S-L-BFGS method is developed to minimize the (Regularized LRM) problem, but the plots in Figure 1 show the original logistic regression function. This may be the reason that in some of the settings, the graphs of the $\mu$-S-L-BFGS method fluctuate and do not decrease monotonically. Another observation is that we find the R-S-L-BFGS method more robust to the choice of the memory parameter $m$. We conclude that iterative regularization makes the stochastic L-BFGS method more robust to the choice of algorithm settings and does not require additional computational effort for tuning the regularization parameter.

6. Concluding remarks. We consider stochastic quasi-Newton (SQN) methods for solving large scale stochastic optimization problems with smooth but unbounded gradients. Much of the past research on convergence rates of these algorithms relies on the strong convexity of the objective function. Importantly, no rate statement exists for this class of algorithms in absence of this assumption. Motivated by this gap, we develop a new framework for $\ell_2$ regularization where the regularization parameter is updated iteratively within the algorithm. We establish the convergence in an a.s. sense and derive the convergence rate statements for SQN methods and their deterministic counterparts. Particularly, we prove that the iterates generated by the regularized stochastic L-BFGS scheme converges to an optimal solution at the rate of the order $O\left(\frac{1}{k^{\frac{1}{3}}-\epsilon}\right)$ for arbitrary small $\epsilon > 0$. The deterministic variant of this algorithm achieves the rate $O\left(\frac{1}{x^{\frac{1}{3}}-\epsilon}\right)$ for arbitrary small $\epsilon' > 0$. Our empirical analysis on a large scale binary classification problem is promising.

REFERENCES

[1] A. Bordes, L. Bottou, and P. Gallinari, SGD-QN: Careful quasi-Newton stochastic gradient descent, Journal of Machine Learning Research 10 (2009), 1737–1754.
[2] L. Bottou, Large-scale machine learning with stochastic gradient descent, Proceedings of the 19th International Conference on Computational Statistics (Paris, France) (Y. Lechevallier and G. Saporta, eds.), Springer, 2010, pp. 177–187.
[3] R. H. Byrd, S. L. Hansen, J. Nocedal, and Y. Singer, A stochastic Quasi-Newton method for large-scale optimization, SIAM Journal on Optimization 26 (2016), no. 2, 1008–1031.
[4] C. Charitha, J. Dutta, and L. D. Russell, Lagrange multipliers, (exact) regularization and error bounds for monotone variational inequalities, Mathematical Programming 161 (2017), no. 1-2, 519–549.
[5] R. Fletcher, A new approach to variable metric algorithms, The Computer Journal 13 (1970), no. 3, 317–322.
[6] M. P. Friedlander and P. Tseng, Exact regularization of convex programs, SIAM Journal on Optimization 18 (2008), no. 4, 1320–1350.
[7] D. Goldfarb, A family of variable-metric methods derived by variational means, Mathematics of Computation 24 (1970), no. 109, 23–26.
[8] M. Grbzbalaban, A. Ozdaglar, and P. Parrilo, A globally convergent incremental newton method, Mathematical Programming 151 (2015), no. 1, 283–313.
[9] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, An interior-point method for large-scale $l_1$-regularized least squares, IEEE Journal of Selected Topics in Signal Processing 1
[10] J. Koshal, A. Nedić, and U. V Shanbhag, Regularized iterative stochastic approximation methods for stochastic variational inequality problems, IEEE Transactions on Automatic Control 58 (2013), no. 3, 594–609.

[11] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, RCV1: A new benchmark collection for text categorization research, Journal of machine learning research 5 (2004), 361–397.

[12] D. Li and M. Fukushima, A modified BFGS method and its global convergence in nonconvex minimization, Journal of Computational and Applied Mathematics 129 (2001), 15–35.

[13] D. C. Liu and J. Nocedal, On the limited memory BFGS method for large scale optimization, Mathematical programming 45 (1989), 503–528.

[14] A. Lucchi, B. McWilliams, and T. Hofmann, A variance reduced stochastic Newton method, arXiv preprint arXiv:1503.08316 (2015).

[15] O. L. Mangasarian M. C. Ferris, Finite perturbation of convex programs, Applied Mathematics and Optimization 23 (1991), no. 1, 263–273.

[16] O. L. Mangasarian and R. R. Meyerf, Nonlinear perturbation of linear programs, SIAM Journal of Control and Optimization 17 (1979), no. 6, 745–757.

[17] A. Mokhtari and A. Ribeiro, RES: regularized stochastic BFGS algorithm, IEEE Transactions on Signal Processing 62 (2014), no. 23, 6089–6104.

[18] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, Robust stochastic approximation approach to stochastic programming, SIAM Journal on Optimization 19 (2009), no. 4, 1574–1609.

[19] A. Nedić and S. Lee, On stochastic subgradient mirror-descent algorithm with weighted averaging, SIAM Journal on Optimization 24 (2014), no. 1, 84–107.

[20] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, Robust stochastic approximation approach to stochastic programming, SIAM Journal on Optimization 19 (2009), no. 4, 1574–1609.

[21] J. Nocedal and S. J. Wright, Numerical optimization, 2nd ed., Springer, New York, 2006.

[22] B.T. Polyak, Introduction to optimization, Optimization Software, Inc., New York, 1987.

[23] B. Recht, M. Fazel, and P. A. Parrillo, Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, SIAM Review 52 (2012), no. 3, 471–501.

[24] H. Robbins and S. Mono, A stochastic approximation method, Ann. Math. Statistics 22 (1951), 400–407.

[25] N. N. Schraudolph, J. Yu, and S. Gunter, A stochastic quasi-Newton method for online convex optimization, In Proc. 11th Intl. Conf. on Artificial Intelligence and Statistics (Aistats) (2007), 433–440.

[26] A. Shapiro, Monte Carlo sampling methods, Handbook in Operations Research and Management Science, vol. 10, Elsevier Science, Amsterdam, 2003, pp. 353–426.

[27] N. Srebro, J. D. M. Rennie, and T. S. Jaakola, Maximum-margin matrix factorization, Advances in Neural Information Processing Systems 17, MIT Press, 2005, pp. 1329–1336.

[28] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of Royal Statistical Society 58 (1996), no. 1, 267–288.

[29] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the royal statistical society 58 (1996), no. 1, 267–288.

[30] P. Tseng and S. Yun, Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization, Journal of Optimization Theory and Applications 140 (2009), no. 3, 513–535.

[31] X. Wang, S. Ma, and W. Liu, Stochastic quasi-Newton methods for nonconvex stochastic optimization, SIAM Journal on Optimization 27 (2017), no. 2, 927–956.

[32] S. J. Wright, Accelerated block-coordinate relaxation for regularized optimization, SIAM Journal on Optimization 22 (2012), no. 1, 159–186.

[33] F. Yousefian, A. Nedić, and U. V. Shanbhag, On stochastic gradient and subgradient methods with adaptive step-length sequences, Automatica 48 (2012), no. 1, 56–67, An extended version of the paper available at: http://arxiv.org/abs/1105.4549.

[34] F. Yousefian, A. Nedić, and U. V. Shanbhag, Self-tuned stochastic approximation schemes for non-Lipschitzian stochastic multi-user optimization and nash games, IEEE Transactions on Automatic Control 61 (2016), no. 7, 1753–1766.

[35] F. Yousefian, A. Nedić, and U. V. Shanbhag, Stochastic quasi-Newton methods for non-strongly convex problems: Convergence and rate analysis, IEEE 55th Conference on Decision and Control (CDC), IEEE, 2016, pp. 4496–4503.

[36] F. Yousefian, A. Nedić, and U. V. Shanbhag, On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems, Mathematical Programming (Series B.) 165 (2017), no. 1, 391–431.