

A Stochastic Quasi-Newton Method with Nesterov’s Accelerated Gradient

S. Indrapriyadarsini¹, Shahrzad Mahboubi², Hiroshi Ninomiya², and Hideki Asai¹ ✉

¹ Shizuoka University, Hamamatsu, Shizuoka Pre., Japan
{s.indrapriyadarsini.17,asai.hideki}@shizuoka.ac.jp

² Shonan Institute of Technology, Fujisawa, Kanagawa Pre., Japan
{18T2012@sit,ninomiya@info}.shonan-it.ac.jp

Abstract. Incorporating second order curvature information in gradient based methods have shown to improve convergence drastically despite its computational intensity. In this paper, we propose a stochastic (online) quasi-Newton method with Nesterov’s accelerated gradient in both its full and limited memory forms for solving large scale non-convex optimization problems in neural networks. The performance of the proposed algorithm is evaluated in Tensorflow on benchmark classification and regression problems. The results show improved performance compared to the classical second order oBFGS and oLBFGS methods and popular first order stochastic methods such as SGD and Adam. The performance with different momentum rates and batch sizes have also been illustrated.

Keywords: Neural networks · stochastic method · online training · Nesterov’s accelerated gradient · quasi-Newton method · limited memory · Tensorflow

1 Introduction

Neural networks have shown to be effective in innumerable real-world applications. Most of these applications require large neural network models with massive amounts of training data to achieve good accuracies and low errors. Neural network optimization poses several challenges such as ill-conditioning, vanishing and exploding gradients, choice of hyperparameters, etc. Thus choice of the optimization algorithm employed on the neural network model plays an important role. It is expected that the neural network training imposes relatively lower computational and memory demands, in which case a full-batch approach is not suitable. Thus, in large scale optimization problems, a stochastic approach is more desirable. Stochastic optimization algorithms use a small subset of data (mini-batch) in its evaluations of the objective function. These methods are particularly of relevance in examples of a continuous stream of data, where the partial data is to be modelled as it arrives. Since the stochastic or online methods operate on small subsamples of the data and its gradients, they significantly reduce the computational and memory requirements.

1.1 Related Works

Gradient based algorithms are popularly used in training neural network models. These algorithms can be broadly classified into first order and second order methods [1]. Several works have been devoted to stochastic first-order methods such as stochastic gradient descent (SGD) [2, 3] and its variance-reduced forms [4–6], AdaGrad [7], RMSprop [8] and Adam [9]. First order methods are popular due to its simplicity and optimal complexity. However, incorporating the second order curvature information have shown to improve convergence. But one of the major drawbacks in second order methods is its need for high computational and memory resources. Thus several approximations have been proposed under Newton [10, 11] and quasi-Newton [12] methods in order to make use of the second order information while keeping the computational load minimal.

Unlike the first order methods, getting quasi-Newton methods to work in a stochastic setting is challenging and has been an active area of research. The oBFGS method [13] is one of the early stable stochastic quasi-Newton methods, in which the gradients are computed twice using the same sub-sample, to ensure stability and scalability. Recently there has been a surge of interest in designing efficient stochastic second order variants which are better suited for large scale problems. [14] proposed a regularized stochastic BFGS method (RES) that modifies the proximity condition of BFGS. [15] further analyzed the global convergence properties of stochastic BFGS and proposed an online L-BFGS method. [16] proposed a stochastic limited memory BFGS (SQN) through sub-sampled Hessian vector products. [17] proposed a general framework for stochastic quasi-Newton methods that assume noisy gradient information through first order oracle (SFO) and extended it to a stochastic damped L-BFGS method (SdLBFGS). This was further modified in [18] by reinitializing the Hessian matrix at each iteration to improve convergence and normalizing the search direction to improve stability. There are also several other studies on stochastic quasi-Newton methods with variance reduction [19–21], sub-sampling [11, 22] and block updates [23]. Most of these methods have been proposed for solving convex optimization problems, but training of neural networks for non-convex problems have not been mentioned in their scopes. The focus of this paper is on training neural networks for non-convex problems with methods similar to that of the oBFGS in [13] and RES [14, 15], as they are stochastic extensions of the classical quasi-Newton method. Thus, the other sophisticated algorithms [11, 16–23] are excluded from comparison in this paper and will be studied in future works.

In this paper, we introduce a novel stochastic quasi-Newton method that is accelerated using Nesterov’s accelerated gradient. Acceleration of quasi-Newton method with Nesterov’s accelerated gradient have shown to improve convergence [24, 25]. The proposed algorithm is a stochastic extension of the accelerated methods in [24, 25] with changes similar to the oBFGS method. The proposed method is also discussed both in its full and limited memory forms. The performance of the proposed methods are evaluated on benchmark classification and regression problems and compared with the conventional SGD, Adam and o(L)BFGS methods.

2 Background

$$\min_{\mathbf{w} \in \mathbb{R}^d} E(\mathbf{w}) = \frac{1}{b} \sum_{p \in X} E_p(\mathbf{w}), \tag{1}$$

Training in neural networks is an iterative process in which the parameters are updated in order to minimize an objective function. Given a mini-batch $X \subseteq T_r$ with samples $(x_p, d_p)_{p \in X}$ drawn at random from the training set T_r and error function $E_p(\mathbf{w}; x_p, d_p)$ parameterized by a vector $\mathbf{w} \in \mathbb{R}^d$, the objective function is defined as in (1) where $b = |X|$, is the batch size. In full batch, $X = T_r$ and $b = n$ where $n = |T_r|$. In gradient based methods, the objective function $E(\mathbf{w})$ under consideration is minimized by the iterative formula (2) where k is the iteration count and \mathbf{v}_{k+1} is the update vector, which is defined for each gradient algorithm.

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \mathbf{v}_{k+1}. \tag{2}$$

In the following sections, we briefly discuss the full-batch BFGS quasi-Newton method and full-batch Nesterov’s Accelerated quasi-Newton method in its full and limited memory forms. We further extend to briefly discuss a stochastic BFGS method.

Algorithm 1 BFGS Method

Require: ε and k_{max}
Initialize: $\mathbf{w}_k \in \mathbb{R}^d$ and $\mathbf{H}_k = \mathbf{I}$.

- 1: $k \leftarrow 1$
- 2: Calculate $\nabla E(\mathbf{w}_k)$
- 3: **while** $\|E(\mathbf{w}_k)\| > \varepsilon$ and $k < k_{max}$
do
- 4: $\mathbf{g}_k \leftarrow -\mathbf{H}_k \nabla E(\mathbf{w}_k)$
- 5: Determine α_k by line search
- 6: $\mathbf{v}_{k+1} \leftarrow \alpha_k \mathbf{g}_k$
- 7: $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + \mathbf{v}_{k+1}$
- 8: Calculate $\nabla E(\mathbf{w}_{k+1})$
- 9: Update \mathbf{H}_{k+1} using (4)
- 10: $k \leftarrow k + 1$
- 11: **end while**

Algorithm 2 NAQ Method

Require: $0 < \mu < 1$, ε and k_{max}
Initialize: $\mathbf{w}_k \in \mathbb{R}^d$, $\mathbf{H}_k = \mathbf{I}$ and $\mathbf{v}_k = 0$.

- 1: $k \leftarrow 1$
- 2: **while** $\|E(\mathbf{w}_k)\| > \varepsilon$ and $k < k_{max}$
do
- 3: Calculate $\nabla E(\mathbf{w}_k + \mu \mathbf{v}_k)$
- 4: $\hat{\mathbf{g}}_k \leftarrow -\hat{\mathbf{H}}_k \nabla E(\mathbf{w}_k + \mu \mathbf{v}_k)$
- 5: Determine α_k by line search
- 6: $\mathbf{v}_{k+1} \leftarrow \mu \mathbf{v}_k + \alpha_k \hat{\mathbf{g}}_k$
- 7: $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + \mathbf{v}_{k+1}$
- 8: Calculate $\nabla E(\mathbf{w}_{k+1})$
- 9: Update $\hat{\mathbf{H}}_k$ using (9)
- 10: $k \leftarrow k + 1$
- 11: **end while**

2.1 BFGS quasi-Newton Method

Quasi-Newton methods utilize the gradient of the objective function to achieve superlinear or quadratic convergence. The Broyden-Fletcher-Goldfarb-Shanon (BFGS) algorithm is one of the most popular quasi-Newton methods for unconstrained optimization. The update vector of the quasi-Newton method is given as

$$\mathbf{v}_{k+1} = \alpha_k \mathbf{g}_k, \tag{3}$$

where $\mathbf{g}_k = -\mathbf{H}_k \nabla E(\mathbf{w}_k)$ is the search direction. The hessian matrix \mathbf{H}_k is symmetric positive definite and is iteratively approximated by the following BFGS formula [26].

$$\mathbf{H}_{k+1} = (\mathbf{I} - \mathbf{s}_k \mathbf{y}_k^T / \mathbf{y}_k^T \mathbf{s}_k) \mathbf{H}_k (\mathbf{I} - \mathbf{y}_k \mathbf{s}_k^T / \mathbf{y}_k^T \mathbf{s}_k) + \mathbf{s}_k \mathbf{s}_k^T / \mathbf{y}_k^T \mathbf{s}_k, \quad (4)$$

where \mathbf{I} denotes identity matrix,

$$\mathbf{s}_k = \mathbf{w}_{k+1} - \mathbf{w}_k \quad \text{and} \quad \mathbf{y}_k = \nabla E(\mathbf{w}_{k+1}) - \nabla E(\mathbf{w}_k). \quad (5)$$

The BFGS quasi-Newton algorithm is shown in Algorithm 1.

Limited Memory BFGS (LBFGS): LBFGS is a variant of the BFGS quasi-Newton method, designed for solving large-scale optimization problems. As the scale of the neural network model increases, the $O(d^2)$ cost of storing and updating the Hessian matrix \mathbf{H}_k is expensive [13]. In the limited memory version, the Hessian matrix is defined by applying m BFGS updates using only the last m curvature pairs $\{\mathbf{s}_k, \mathbf{y}_k\}$. As a result, the computational cost is significantly reduced and the storage cost is down to $O(md)$ where d is the number of parameters and m is the memory size.

2.2 Nesterov's Accelerated Quasi-Newton Method

Several modifications have been proposed to the quasi-Newton method to obtain stronger convergence. The Nesterov's Accelerated Quasi-Newton (NAQ) [24] method achieves faster convergence compared to the standard quasi-Newton methods by quadratic approximation of the objective function at $\mathbf{w}_k + \mu \mathbf{v}_k$ and by incorporating the Nesterov's accelerated gradient $\nabla E(\mathbf{w}_k + \mu \mathbf{v}_k)$ in its Hessian update. The derivation of NAQ is briefly discussed as follows.

Let $\Delta \mathbf{w}$ be the vector $\Delta \mathbf{w} = \mathbf{w} - (\mathbf{w}_k + \mu \mathbf{v}_k)$. The quadratic approximation of the objective function at $\mathbf{w}_k + \mu \mathbf{v}_k$ is defined as,

$$E(\mathbf{w}) \simeq E(\mathbf{w}_k + \mu \mathbf{v}_k) + \nabla E(\mathbf{w}_k + \mu \mathbf{v}_k)^T \Delta \mathbf{w} + \frac{1}{2} \Delta \mathbf{w}^T \nabla^2 E(\mathbf{w}_k + \mu \mathbf{v}_k) \Delta \mathbf{w}. \quad (6)$$

The minimizer of this quadratic function is explicitly given by

$$\Delta \mathbf{w} = -\nabla^2 E(\mathbf{w}_k + \mu \mathbf{v}_k)^{-1} \nabla E(\mathbf{w}_k + \mu \mathbf{v}_k). \quad (7)$$

Therefore the new iterate is defined as

$$\mathbf{w}_{k+1} = (\mathbf{w}_k + \mu \mathbf{v}_k) - \nabla^2 E(\mathbf{w}_k + \mu \mathbf{v}_k)^{-1} \nabla E(\mathbf{w}_k + \mu \mathbf{v}_k). \quad (8)$$

This iteration is considered as Newton method with the momentum term $\mu \mathbf{v}_k$. The inverse of Hessian $\nabla^2 E(\mathbf{w}_k + \mu \mathbf{v}_k)$ is approximated by the matrix $\hat{\mathbf{H}}_{k+1}$ using the update equation (9)

$$\hat{\mathbf{H}}_{k+1} = (\mathbf{I} - \mathbf{p}_k \mathbf{q}_k^T / \mathbf{q}_k^T \mathbf{p}_k) \hat{\mathbf{H}}_k (\mathbf{I} - \mathbf{q}_k \mathbf{p}_k^T / \mathbf{q}_k^T \mathbf{p}_k) + \mathbf{p}_k \mathbf{p}_k^T / \mathbf{q}_k^T \mathbf{p}_k, \quad (9)$$

Algorithm 3 Direction Update

Require: current gradient $\nabla E(\theta_k)$, memory size m , curvature pair $(\sigma_{k-i}, \gamma_{k-i})$
 $\forall i = 1, 2, \dots, \min(k-1, m)$ where σ_k is the difference of current and previous weight vector and γ_k is the difference of current and previous gradient vector

- 1: $\eta_k = -\nabla E(\theta_k)$
- 2: **for** $i := 1, 2, \dots, \min(m, k-1)$ **do**
- 3: $\beta_i = (\sigma_{k-i}^\top \eta_k) / (\sigma_{k-i}^\top \gamma_{k-i})$
- 4: $\eta_k = \eta_k - \beta_i \gamma_{k-i}$
- 5: **end for**
- 6: **if** $k > 1$ **then**
- 7: $\eta_k = \eta_k (\sigma_k^\top \gamma_k / \gamma_k^\top \gamma_k)$
- 8: **end if**
- 9: **for** $i : k - \min(m, (k-1)), \dots, k-1, k$ **do**
- 10: $\tau = (\gamma_i^\top \eta_k) / (\gamma_i^\top \sigma_i)$
- 11: $\eta_k = \eta_k - (\beta_i - \tau) \sigma_i$
- 12: **end for**
- 13: **return** η_k

where

$$\mathbf{p}_k = \mathbf{w}_{k+1} - (\mathbf{w}_k + \mu \mathbf{v}_k) \quad \text{and} \quad \mathbf{q}_k = \nabla E(\mathbf{w}_{k+1}) - \nabla E(\mathbf{w}_k + \mu \mathbf{v}_k). \quad (10)$$

(9) is derived from the secant condition $\mathbf{q}_k = (\hat{\mathbf{H}}_{k+1})^{-1} \mathbf{p}_k$ and the rank-2 updating formula [24]. It is proved that the Hessian matrix $\hat{\mathbf{H}}_{k+1}$ updated by (9) is a positive definite symmetric matrix given $\hat{\mathbf{H}}_k$ is initialized to identity matrix [24]. Therefore, the update vector of NAQ can be written as:

$$\mathbf{v}_{k+1} = \mu \mathbf{v}_k + \alpha_k \hat{\mathbf{g}}_k, \quad (11)$$

where $\hat{\mathbf{g}}_k = -\hat{\mathbf{H}}_k \nabla E(\mathbf{w}_k + \mu \mathbf{v}_k)$ is the search direction. The NAQ algorithm is given in Algorithm 2. Note that the gradient is computed twice in one iteration. This increases the computational cost compared to the BFGS quasi-Newton method. However, due to acceleration by the momentum and Nesterov’s gradient term, NAQ is faster in convergence compared to BFGS.

Limited Memory NAQ (LNAQ) Similar to LBFGS method, LNAQ [25] is the limited memory variant of NAQ that uses the last m curvature pairs $\{\mathbf{p}_k, \mathbf{q}_k\}$. In the limited-memory form note that the curvature pairs that are used incorporate the momentum and Nesterov’s accelerated gradient term, thus accelerating LBFGS. Implementation of LNAQ algorithm can be realized by omitting steps 4 and 9 of Algorithm 2 and determining the search direction $\hat{\mathbf{g}}_k$ using the two-loop recursion [26] shown in Algorithm 3. The last m vectors of \mathbf{p}_k and \mathbf{q}_k are stored and used in the direction update.

2.3 Stochastic BFGS quasi-Newton Method (oBFGS)

The online BFGS method proposed by Schraudolph et al in [13] is a fast and scalable stochastic quasi-Newton method suitable for convex functions. The changes

proposed to the BFGS method in [13] to work well in a stochastic setting are discussed as follows. The line search is replaced with a gain schedule such as

$$\alpha_k = \tau/(\tau + k) \cdot \alpha_0, \quad (12)$$

where $\alpha_0, \tau > 0$ provided the Hessian matrix is positive definite, thus restricting to convex optimization problems. Since line search is eliminated, the first parameter update is scaled by a small value. Further, to improve the performance of oBFGS, the step size is divided by an analytically determined constant c . An important modification is the computation of \mathbf{y}_k , the difference of the last two gradients is computed on the same sub-sample X_k [13,14] as given below,

$$\mathbf{y}_k = \nabla E(\mathbf{w}_{k+1}, X_k) - \nabla E(\mathbf{w}_k, X_k). \quad (13)$$

This however doubles the cost of gradient computation per iteration but is shown to outperform natural gradient descent for all batch sizes [13]. The oBFGS algorithm is shown in Algorithm 4. In this paper, we introduce direction normalization as shown in step 5, details of which are discussed in the next section.

Stochastic Limited Memory BFGS (oLBFGS) [13] further extends the oBFGS method to limited memory form by determining the search direction \mathbf{g}_k using the two-loop recursion (Algorithm 3). The Hessian update is omitted and instead the last m curvature pairs \mathbf{s}_k and \mathbf{y}_k are stored. This brings down the computation complexity to $2bd + 6md$ where b is the batch size, d is the number of parameters, and m is the memory size. To improve the performance by averaging sampling noise step 7 of Algorithm 3 is replaced by (14) where σ_k is \mathbf{s}_k and γ_k is \mathbf{y}_k .

$$\eta_k = \begin{cases} \epsilon \eta_k & \text{if } k = 1, \\ \frac{\eta_k}{\min(k, m)} \sum_{i=1}^{\min(k, m)} \frac{\sigma_{k-i}^T \gamma_{k-i}}{\gamma_{k-i}^T \gamma_{k-i}} & \text{otherwise.} \end{cases} \quad (14)$$

3 Proposed Algorithm - oNAQ and oLNAQ

The oBFGS method proposed in [13] computes the gradient of a sub-sample minibatch X_k twice in one iteration. This is comparable with the inherent nature of NAQ which also computes the gradient twice in one iteration. Thus by applying suitable modifications to the original NAQ algorithm, we achieve a stochastic version of the Nesterov's Accelerated Quasi-Newton method. The proposed modifications for a stochastic NAQ method is discussed below in its full and limited memory forms.

3.1 Stochastic NAQ (oNAQ)

The NAQ algorithm computes two gradients, $\nabla E(\mathbf{w}_k + \mu \mathbf{v}_k)$ and $\nabla E(\mathbf{w}_{k+1})$ to calculate \mathbf{q}_k as shown in (10). On the other hand, the oBFGS method proposed

in [13] computes the gradient $\nabla E(\mathbf{w}_k, X_k)$ and $\nabla E(\mathbf{w}_{k+1}, X_k)$ to calculate \mathbf{y}_k as shown in (13). Therefore, oNAQ can be realised by changing steps 3 and 8 of Algorithm 2 to calculate $\nabla E(\mathbf{w}_k + \mu \mathbf{v}_k, X_k)$ and $\nabla E(\mathbf{w}_{k+1}, X_k)$. Thus in oNAQ, the \mathbf{q}_k vector is given by (15) where $\lambda \mathbf{p}_k$ is used to guarantee numerical stability [27–29].

$$\mathbf{q}_k = \nabla E(\mathbf{w}_{k+1}, X_k) - \nabla E(\mathbf{w}_k + \mu \mathbf{v}_k, X_k) + \lambda \mathbf{p}_k, \quad (15)$$

Further, unlike in full batch methods, the updates in stochastic methods have high variance resulting in the objective function to fluctuate heavily. This is due to the updates being performed based on small sub-samples of data. This can be seen more prominently in case of the limited memory version where the updates are based only on m recent curvature pairs. Thus in order to improve the stability of the algorithm, we introduce direction normalization as

$$\hat{\mathbf{g}}_k = \mathbf{g}_k / \|\mathbf{g}_k\|_2, \quad (16)$$

where $\|\mathbf{g}_k\|_2$ is the l_2 norm of the search direction \mathbf{g}_k . Normalizing the search direction at each iteration ensures that the algorithm does not move too far away from the current objective [18]. Fig.1 illustrates the effect of direction normalization on oBFGS and the proposed oNAQ method. The solid lines indicate the moving average. As seen from the figure, direction normalization improves the performance of both oBFGS and oNAQ. Therefore, in this paper we include direction normalization for oBFGS also.

The next proposed modification is with respect to the step size. In full batch methods, the step size or the learning rate is usually determined by line search methods satisfying either Armijo or Wolfe conditions. However, in stochastic methods, line searches are not quite effective since search conditions apply global validity. This cannot be assumed when using small local sub-samples [13]. Several studies show that line search methods does not necessarily ensure global convergence and have proposed methods that eliminate line search [27–29]. Moreover, determining step size using line search methods involves additional function computations until the search conditions such as the Armijo or Wolfe condition is satisfied. Hence we determine the step size using a simple learning rate schedule. Common learning rate schedules are polynomial decays and exponential decay functions. In this paper, we determine the step size using a polynomial decay schedule [30]

$$\alpha_k = \alpha_0 / \sqrt{k}, \quad (17)$$

where α_0 is usually set to 1. If the step size is too large, which is the case in the initial iterations, the learning can become unstable. This is stabilized by direction normalization. A comparison of common learning rate schedules are illustrated in Fig. 2

The proposed stochastic NAQ algorithm is shown in Algorithm 5. Note that the gradient is computed twice in one iteration, thus making the computational cost same as that of the stochastic BFGS (oBFGS) proposed in [13].

Algorithm 4 oBFGS Method

Require: minibatch X_k , k_{max} and $\lambda \geq 0$,
Initialize: $\mathbf{w}_k \in \mathbb{R}^d$, $\mathbf{H}_k = \epsilon \mathbf{I}$ and $\mathbf{v}_k = 0$

- 1: $k \leftarrow 1$
- 2: **while** $k < k_{max}$ **do**
- 3: $\nabla \mathbf{E}_1 \leftarrow \nabla E(\mathbf{w}_k, X_k)$
- 4: $\mathbf{g}_k \leftarrow -\mathbf{H}_k \nabla E(\mathbf{w}_k, X_k)$
- 5: $\mathbf{g}_k = \mathbf{g}_k / \|\mathbf{g}_k\|_2$
- 6: Determine α_k using (12)
- 7: $\mathbf{v}_{k+1} \leftarrow \alpha_k \mathbf{g}_k$
- 8: $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + \mathbf{v}_{k+1}$
- 9: $\nabla \mathbf{E}_2 \leftarrow \nabla E(\mathbf{w}_{k+1}, X_k)$
- 10: $\mathbf{s}_k \leftarrow \mathbf{w}_{k+1} - \mathbf{w}_k$
- 11: $\mathbf{y}_k \leftarrow \nabla \mathbf{E}_2 - \nabla \mathbf{E}_1 + \lambda \mathbf{s}_k$
- 12: Update \mathbf{H}_k using (4)
- 13: $k \leftarrow k + 1$
- 14: **end while**

Algorithm 5 Proposed oNAQ Method

Require: minibatch X_k , $0 < \mu < 1$ and k_{max}
Initialize: $\mathbf{w}_k \in \mathbb{R}^d$, $\hat{\mathbf{H}}_k = \epsilon \mathbf{I}$ and $\mathbf{v}_k = 0$

- 1: $k \leftarrow 1$
- 2: **while** $k < k_{max}$ **do**
- 3: $\nabla \mathbf{E}_1 \leftarrow \nabla E(\mathbf{w}_k + \mu \mathbf{v}_k, X_k)$
- 4: $\hat{\mathbf{g}}_k \leftarrow -\hat{\mathbf{H}}_k \nabla E(\mathbf{w}_k + \mu \mathbf{v}_k, X_k)$
- 5: $\hat{\mathbf{g}}_k = \hat{\mathbf{g}}_k / \|\hat{\mathbf{g}}_k\|_2$
- 6: Determine α_k using (17)
- 7: $\mathbf{v}_{k+1} \leftarrow \mu \mathbf{v}_k + \alpha_k \hat{\mathbf{g}}_k$
- 8: $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + \mathbf{v}_{k+1}$
- 9: $\nabla \mathbf{E}_2 \leftarrow \nabla E(\mathbf{w}_{k+1}, X_k)$
- 10: $\mathbf{p}_k \leftarrow \mathbf{w}_{k+1} - (\mathbf{w}_k + \mu \mathbf{v}_k)$
- 11: $\mathbf{q}_k \leftarrow \nabla \mathbf{E}_2 - \nabla \mathbf{E}_1 + \lambda \mathbf{p}_k$
- 12: Update $\hat{\mathbf{H}}_k$ using (9)
- 13: $k \leftarrow k + 1$
- 14: **end while**

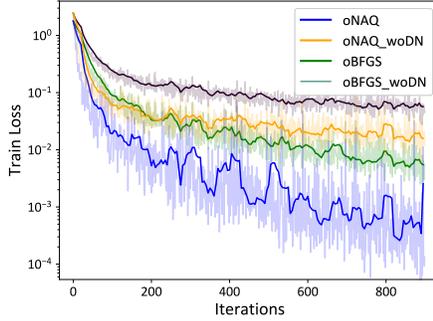


Fig. 1: Effect of direction normalization on 8x8 MNIST with $b = 64$ and $\mu = 0.8$.

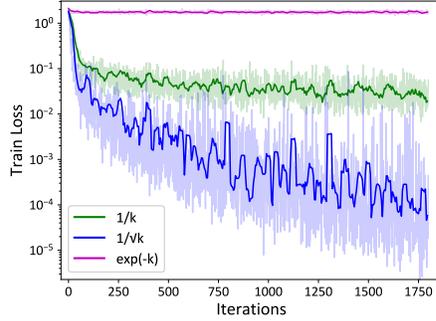


Fig. 2: Comparison of α_k schedules on 8x8 MNIST with $b = 64$ and $\mu = 0.8$.

3.2 Stochastic Limited-Memory NAQ (oLNAQ)

Stochastic LNAQ can be realized by making modifications to Algorithm 5 similar to LNAQ. The search direction $\hat{\mathbf{g}}_k$ in step 4 is determined by Algorithm 3. oLNAQ like LNAQ uses the last m curvature pairs $\{\mathbf{p}_k, \mathbf{q}_k\}$ to estimate the Hessian matrix instead of storing and computing on a $d \times d$ matrix. Therefore, the

implementation of oLNAQ does not require initializing or updating the Hessian matrix. Hence step 12 of Algorithm 5 is replaced by storing the last m curvature pairs $\{\mathbf{p}_k, \mathbf{q}_k\}$. Finally, in order to average out the sampling noise in the last m steps, we replace step 7 of Algorithm 3 by eq. (14) where σ_k is \mathbf{p}_k and γ_k is \mathbf{q}_k . Note that an additional $2md$ evaluations are required to compute (14). However the overall computation cost of oLNAQ is much lesser than that of oNAQ and the same as oLBFGS.

4 Simulation Results

We illustrate the performance of the proposed stochastic methods oNAQ and oLNAQ on four benchmark datasets - two classification and two regression problems. For the classification problem we use the 8x8 MNIST and 28x28 MNIST datasets and for the regression problem we use the Wine Quality [31] and CASP [32] datasets. We evaluate the performance of the classification tasks on a multi-layer neural network (MLNN) and a simple convolution neural network (CNN). The algorithms oNAQ, oBFGS, oLNAQ and oLBFGS are implemented in Tensorflow using the ScipyOptimizerInterface class. Details of the simulation are given in Table 1.

4.1 Multi-Layer Neural Networks - Classification Problem

We evaluate the performance of the proposed algorithms for classification of handwritten digits using the 8x8 MNIST [33] and 28x28 MNIST dataset [34]. We consider a simple MLNN with two hidden layers. ReLU activation function and softmax cross-entropy loss function is used. Each layer except the output layer is batch normalized.

Table 1: Details of the Simulation - MLNN.

	8x8 MNIST	28x28 MNIST	Wine Quality	CASP
task	classification	classification	regression	regression
input	8x8	28x28	11	9
MLNN structure	64-20-10-10	784-100-50-10	11-10-4-1	9-10-6-1
parameters (d)	1,620	84,060	169	173
train set	1,198	55,000	3,918	36,584
test set	599	10,000	980	9,146
classes/output	10	10	1	1
momentum (μ)	0.8	0.85	0.95	0.95
batch size (b)	64	64/128	32/64	64/128
memory (m)	4	4	4	4

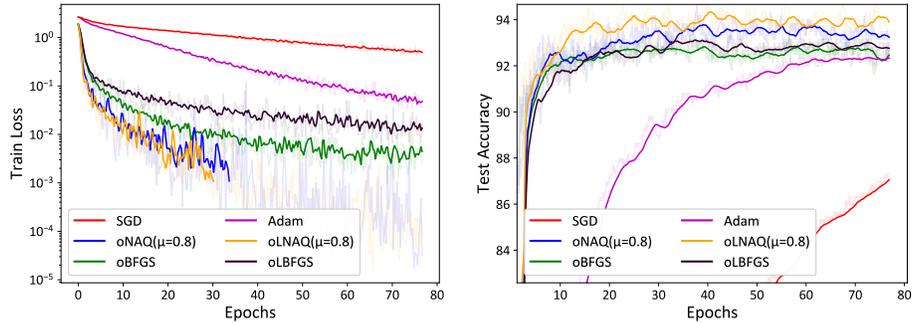


Fig. 3: Comparison of train loss and test accuracy versus number of epochs required for convergence of 8x8 MNIST data with a maximum of 80 epochs.

Results on 8x8 MNIST Dataset We evaluate the performance of oNAQ and oLNAQ on a reduced version of the MNIST dataset in which each sample is an 8x8 image representing a handwritten digit [33]. Fig. 3 shows the number of epochs required to converge to a train loss of $< 10^{-3}$ and its corresponding test accuracy for a batch size $b = 64$. The maximum number of epochs is set to 80. As seen from the figure, it is clear that oNAQ and oLNAQ require fewer epochs compared to oBFGS, oLBFGS, Adam and SGD. In terms of computation time, o(L)BFGS and o(L)NAQ require longer time compared to the first order methods. This is due to the Hessian computation and twice gradient calculation. Further, the oBFGS and oNAQ per iteration time difference compared to first order methods is much larger than that of the limited memory algorithms with memory $m = 4$. This can be seen from Fig. 4 which shows the comparison of train loss and test accuracy versus time for 80 epochs. It can be observed that for the same time, the second order methods perform significantly better compared to the first order methods, thus confirming that the extra time taken by the second order methods does not adversely affect its performance. Thus, in the subsequent sections we compare the train loss and test accuracy versus time to evaluate the performance of the proposed method.

Results on 28x28 MNIST Dataset Next, we evaluate the performance of the proposed algorithm on the standard 28x28 pixel MNIST dataset [34]. Due to system constraints and large number of parameters, we illustrate the performance of only the limited memory methods. Fig.5 shows the results of oLNAQ on the 28x28 MNIST dataset for batch size $b = 64$ and $b = 128$. The results indicate that oLNAQ clearly outperforms oLBFGS and SGD for even small batch sizes. On comparing with Adam, oLNAQ is in close competition with Adam for small batch sizes such as $b = 64$ and performs better for larger batch sizes such as $b = 128$.

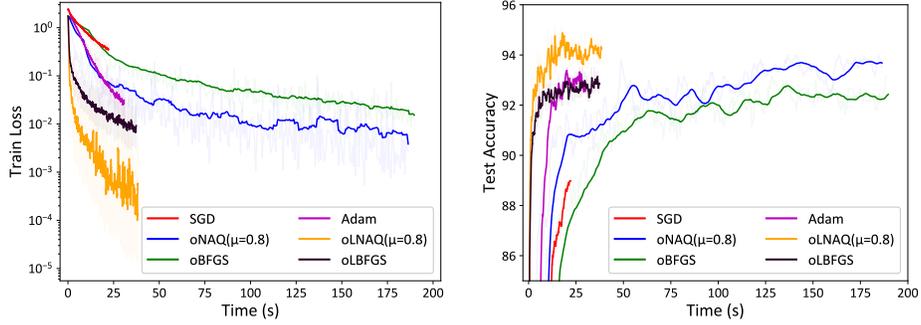


Fig. 4: Comparison of train loss and test accuracy over time on 8x8 MNIST (80 epochs).

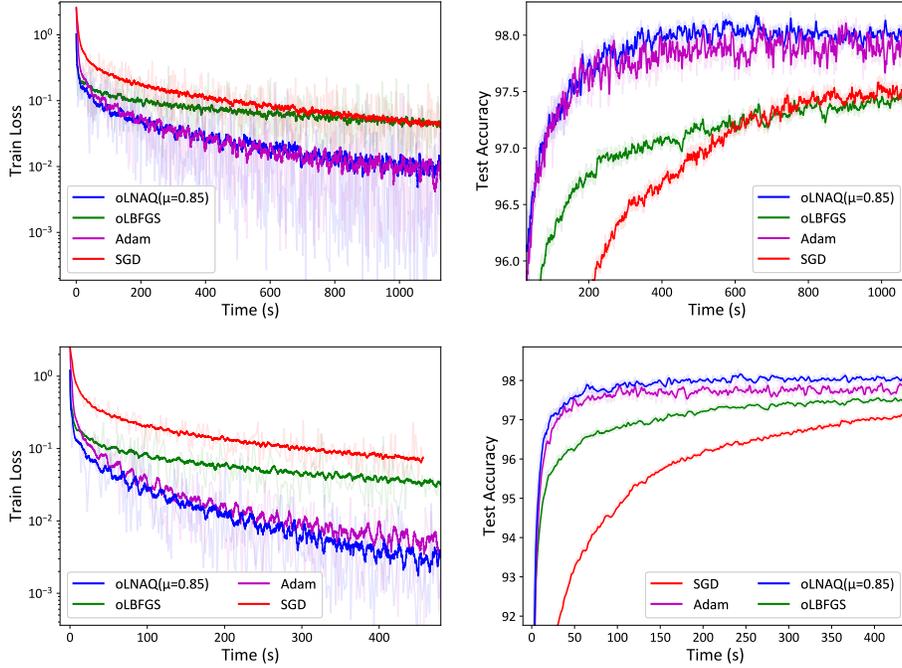


Fig. 5: Results on 28x28 MNIST for $b = 64$ (top) and $b = 128$ (bottom).

4.2 Convolution Neural Network - Classification Task

We study the performance of the proposed algorithm on a simple convolution neural network (CNN) with two convolution layers followed by a fully connected layer. We use sigmoid activation functions and softmax cross-entropy error function. We evaluate the performance of oNAQ using the 8x8 MNIST dataset with

a batch size of 64 and $\mu = 0.8$ and number of parameters $d = 778$. The CNN architecture comprises of two convolution layers of 3 and 5 5×5 filters respectively, each followed by 2×2 max pooling layer with stride 2. The convolution layers are followed by a fully connected layer with 10 hidden neurons. Fig. 6 shows the CNN results of 8×8 MNIST. Calculation of the gradient twice per iteration increases the time per iteration when compared to the first order methods. However this is compensated well since the overall performance of the algorithm is much better compared to Adam and SGD. Also the number of epochs required to converge to low error and high accuracies is much lesser than the other algorithms. In other words, the same accuracy or error can be achieved with lesser amount of training data. Further, we evaluate the performance of oLNAQ using the 28×28 MNIST dataset with batch size $b = 128$, $m = 4$ and $d = 260,068$. The CNN architecture is similar to that as described above except that the fully connected layer has 100 hidden neurons. Fig.7 shows the results of oLNAQ on the simple CNN. The CNN results show similar performance as that of the results on multi-layer neural network where oLNAQ outperforms SGD and oBFGS. Comparing with Adam, oLNAQ is much faster in the first few epochs and becomes closely competitive to Adam as the number of epochs increases.

4.3 Multi-layer Neural Network - Regression Problem

We further extend to study the performance of the proposed stochastic methods on regression problems. For this task, we choose two benchmark datasets - prediction of white wine quality [31] and CASP [32] dataset. We evaluate the performance of oNAQ and oLNAQ on multi-layer neural network as shown in Table 1. Sigmoid activation function and mean squared error (MSE) function is used. Each layer except the output layer is batch normalized. Both datasets were z-normalized to have zero mean and unit variance.

Results on Wine Quality Dataset We evaluate the performance of oNAQ and oLNAQ on the Wine Quality [31] dataset to predict the quality of the white wine on a scale of 3 to 9 based on 11 physiochemical test values. We split the dataset in 80-20 % for train and test set. For the regression problems, oNAQ with smaller values of momentum $\mu = 0.8$ and $\mu = 0.85$ show similar performance as that of oBFGS. Larger values of momentum resulted in better performance. Hence we choose a value of $\mu = 0.95$ which shows faster convergence compared to the other methods. Further comparing the performance for different batch sizes, we observe that for smaller batch sizes such as $b = 32$, oNAQ is close in performance with Adam and oLNAQ is initially fast and gradually becomes close to Adam. For bigger batch sizes such as $b = 64$, oNAQ and oLNAQ are faster in convergence initially. Over time, oLNAQ continues to result in lower error while oNAQ gradually becomes close to Adam. Fig. 8 shows the root mean squared error (RMSE) versus time for batch sizes $b = 32$ and $b = 64$.

Results on CASP Dataset The next regression problem under consideration is the CASP (Critical Assessment of protein Structure Prediction) dataset from

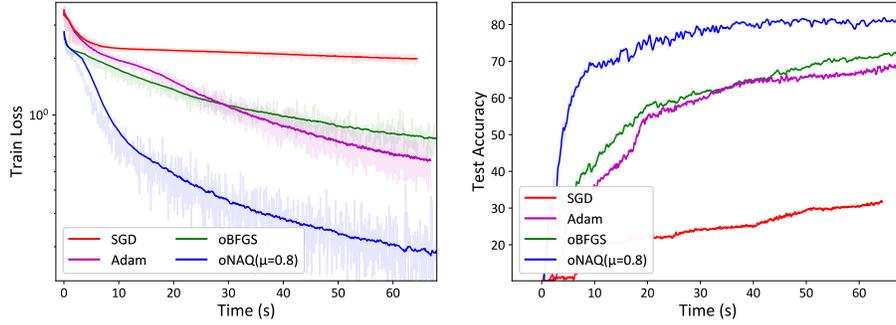


Fig. 6: Convolution Neural Network results on 8x8 MNIST with $b = 64$.

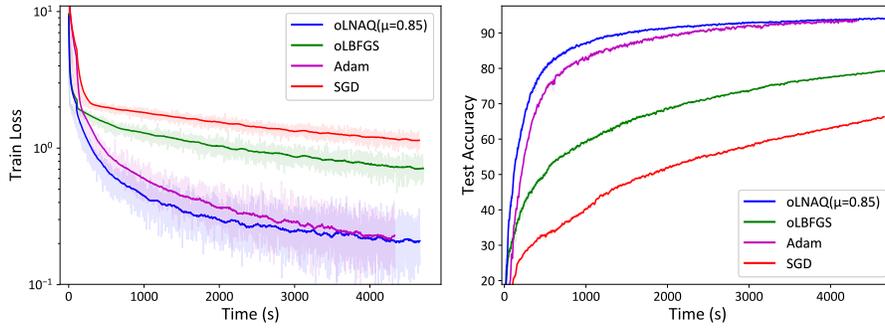


Fig. 7: CNN Results on 28x28 MNIST with $b = 128$.

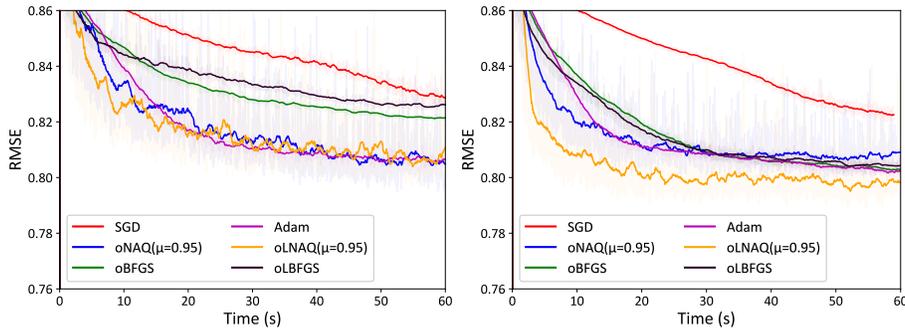
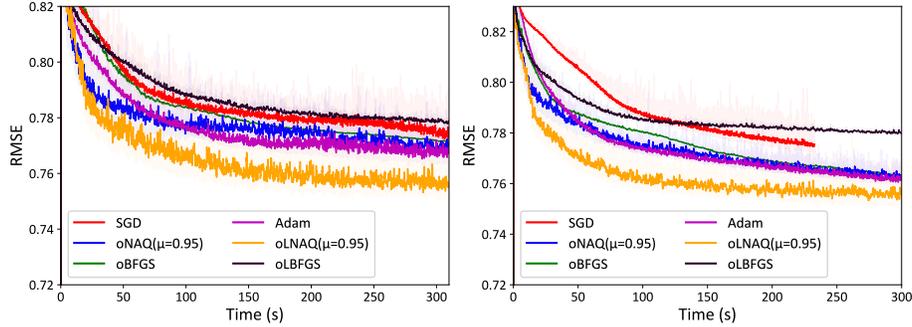
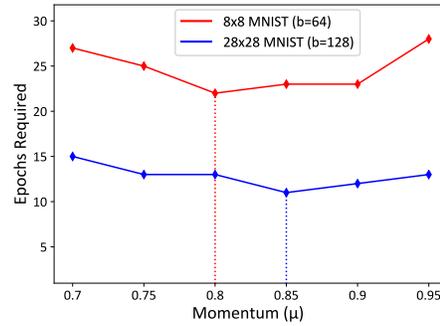
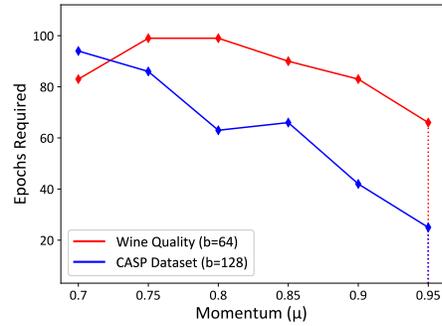


Fig. 8: Results of Wine Quality Dataset for $b = 32$ (left) and $b = 64$ (right).

[32]. It gives the physicochemical properties of protein tertiary structure. We split the dataset in 80-20% for train and test set. Similar to the wine quality problem, a momentum of $\mu = 0.95$ was fixed. Fig. 9 shows the root mean squared error (RMSE) versus time for batch sizes $b = 64$ and $b = 128$. For both batch sizes,

Fig. 9: Results of CASP Dataset for batch size $b = 64$ (left) and $b = 128$ (right).Fig. 10: No. of epochs required to converge for different values of μ with $m = 4$ for oLNAQ classification problems.Fig. 11: No. of epochs required to converge for different values of μ with $m = 4$ for oLNAQ regression problems.

oNAQ is initially fast and becomes close to Adam and shows better performance compared to oBFGS and oLBFGS. On the other hand, we observe that oLNAQ consistently shows decrease in error and outperforms the other algorithms for both batch sizes.

4.4 Discussions on choice of parameters

The momentum term μ is a hyperparameter with a value in the range $0 < \mu < 1$ and is usually chosen closer to 1 [24, 35]. The performance for different values of the momentum term have been studied for all the four problem sets in this paper. Fig. 10 and Fig. 11 show the number of epochs required for convergence for different values of μ for the classification and regression datasets respectively. For the limited memory schemes, a memory size of $m = 4$ showed optimum results for all the four problem datasets with different batch sizes. Larger memory sizes also show good performance. However considering computational efficiency, memory

Table 2: Summary of Computational Cost and Storage.

	Algorithm	Computational Cost	Storage
full batch	BFGS	$nd + d^2 + \zeta nd$	d^2
	NAQ	$2nd + d^2 + \zeta nd$	d^2
	LBFGS	$nd + 4md + 2d + \zeta nd$	$2md$
	LNAQ	$2nd + 4md + 2d + \zeta nd$	$2md$
online	oBFGS	$2bd + d^2$	d^2
	oNAQ	$2bd + d^2$	d^2
	oLBFGS	$2bd + 6md$	$2md$
	oLNAQ	$2bd + 6md$	$2md$

size is usually maintained smaller than the batch size. Since the computation cost is $2bd + 6md$, if $b \approx m$ the computation cost would increase to $8bd$. Hence a smaller memory is desired. Memory sizes less than $m = 4$ does not perform well for small batch sizes and hence $m = 4$ was chosen.

4.5 Computation and Storage Cost

The summary of the computational cost and storage for full batch and stochastic (online) methods are illustrated in Table 2. The cost of function and gradient evaluations can be considered to be nd , where n is the number of training samples involved and d is the number of parameters. The Nesterov’s Accelerated quasi-Newton (NAQ) method computes the gradient twice per iteration compared to the BFGS quasi-Newton method which computes the gradient only once per iteration. Thus NAQ has an additional nd computation cost. In both BFGS and NAQ algorithms, the step length is determined by line search methods which involves ζ function evaluations until the search condition is satisfied. In the limited memory forms the Hessian update is approximated using the two-loop recursion scheme, which requires $4md + 2d$ multiplications. In the stochastic setting, both oBFGS and oNAQ compute the gradient twice per iteration, making the computational cost the same in both. Both methods do not use line search and due to smaller number of training samples (minibatch) in each iteration, the computational cost is smaller compared to full batch. Further, in stochastic limited memory methods, an additional $2md$ evaluations are required to compute the search direction as given (14). In stochastic methods the computational complexity is reduced significantly due to smaller batch sizes ($b < n$).

5 Conclusion

In this paper we have introduced a stochastic quasi-Newton method with Nesterov’s accelerated gradient. The proposed algorithm is shown to be efficient

compared to the state of the art algorithms such Adam and classical quasi-Newton methods. From the results presented above, we can conclude that the proposed o(L)NAQ methods clearly outperforms the conventional o(L)BFGS methods with both having the same computation and storage costs. However the computation time taken by oBFGS and oNAQ are much higher compared to the first order methods due to Hessian computation. On the other hand, we observe that the per iteration computation of Adam, oLBFGS and oLNAQ are comparable. By tuning the momentum parameter μ , oLNAQ is seen to perform better and faster compared to Adam. Hence we can conclude that with an appropriate value of μ , oLNAQ can achieve better results. Further, the limited memory form of the proposed algorithm can efficiently reduce the memory requirements and computational cost while incorporating second order curvature information. Another observation is that the proposed oNAQ and oLNAQ methods significantly accelerates the training especially in the first few epochs when compared to both, first order Adam and second order o(L)BFGS method. Several studies propose pretrained models. oNAQ and oLNAQ can possibly be suitable for pretraining. Also, the computational speeds of oNAQ could be improved further by approximations which we leave for future work. Further studying the performance of the proposed algorithm on bigger problem sets, including that of convex problems and on popular NN architectures such as AlexNet, LeNet and ResNet could test the limits of the algorithm. Furthermore, theoretical analysis of the convergence properties of the proposed algorithms will also be studied in future works.

References

1. Haykin, S.: Neural Networks and Learning Machines. 3rd edn. Pearson Prentice Hall, (2009)
2. Bottou, L., Cun, Y.L.: Large scale online learning. In: Advances in neural information processing systems. (2004) 217–224
3. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT’2010. Springer (2010) 177–186
4. Robbins, H., Monro, S.: A stochastic approximation method. The annals of mathematical statistics (1951) 400–407
5. Peng, X., Li, L., Wang, F.Y.: Accelerating minibatch stochastic gradient descent using typicality sampling. arXiv preprint arXiv:1903.04192 (2019)
6. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: Advances in neural information processing systems. (2013) 315–323
7. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research **12**(Jul) (2011) 2121–2159
8. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. University of Toronto, Technical Report (2012)
9. Kingma, D.P., Ba, J.: Adam : A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Martens, J.: Deep learning via hessian-free optimization. In: ICML. Volume 27. (2010) 735–742

11. Roosta-Khorasani, F., Mahoney, M.W.: Sub-sampled newton methods i: globally convergent algorithms. arXiv preprint arXiv:1601.04737 (2016)
12. Dennis, Jr, J.E., Moré, J.J.: Quasi-newton methods, motivation and theory. *SIAM review* **19**(1) (1977) 46–89
13. Schraudolph, N.N., Yu, J., Günter, S.: A stochastic quasi-newton method for online convex optimization. In: *Artificial Intelligence and Statistics*. (2007) 436–443
14. Mokhtari, A., Ribeiro, A.: Res: Regularized stochastic bfgs algorithm. *IEEE Transactions on Signal Processing* **62**(23) (2014) 6089–6104
15. Mokhtari, A., Ribeiro, A.: Global convergence of online limited memory bfgs. *The Journal of Machine Learning Research* **16**(1) (2015) 3151–3181
16. Byrd, R.H., Hansen, S.L., Nocedal, J., Singer, Y.: A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization* **26**(2) (2016) 1008–1031
17. Wang, X., Ma, S., Goldfarb, D., Liu, W.: Stochastic quasi-newton methods for nonconvex stochastic optimization. *SIAM Journal on Optimization* **27**(2) (2017) 927–956
18. Li, Y., Liu, H.: Implementation of stochastic quasi-newton's method in pytorch. arXiv preprint arXiv:1805.02338 (2018)
19. Lucchi, A., McWilliams, B., Hofmann, T.: A variance reduced stochastic newton method. arXiv preprint arXiv:1503.08316 (2015)
20. Moritz, P., Nishihara, R., Jordan, M.: A linearly-convergent stochastic l-bfgs algorithm. In: *Artificial Intelligence and Statistics*. (2016) 249–258
21. Bollapragada, R., Mudigere, D., Nocedal, J., Shi, H.J.M., Tang, P.T.P.: A progressive batching l-bfgs method for machine learning. arXiv preprint arXiv:1802.05374 (2018)
22. Byrd, R.H., Chin, G.M., Neveitt, W., Nocedal, J.: On the use of stochastic hessian information in optimization methods for machine learning. *SIAM Journal on Optimization* **21**(3) (2011) 977–995
23. Gower, R., Goldfarb, D., Richtárik, P.: Stochastic block bfgs: Squeezing more curvature out of data. In: *International Conference on Machine Learning*. (2016) 1869–1878
24. Ninomiya, H.: A novel quasi-newton-based optimization for neural network training incorporating nesterov's accelerated gradient. *Nonlinear Theory and Its Applications, IEICE* **8**(4) (2017) 289–301
25. Mahboubi, S., Ninomiya, H.: A novel training algorithm based on limited-memory quasi-newton method with nesterov's accelerated gradient in neural networks and its application to highly-nonlinear modeling of microwave circuit. *IARIA International Journal on Advances in Software* **11**(3-4) (2018) 323–334
26. Nocedal, J., Wright, S.J.: *Numerical Optimization*. Springer Series in Operations Research. Springer, second edition (2006)
27. Zhang, L.: A globally convergent bfgs method for nonconvex minimization without line searches. *Optimization Methods and Software* **20**(6) (2005) 737–747
28. Dai, Y.H.: Convergence properties of the bfgs algorithm. *SIAM Journal on Optimization* **13**(3) (2002) 693–701
29. Indrapriyadarsini, S., Mahboubi, S., Ninomiya, H., Asai, H.: Implementation of a modified nesterov's accelerated quasi-newton method on tensorflow. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE (2018) 1147–1154
30. Zinkevich, M.: Online convex programming and generalized infinitesimal gradient ascent. In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. (2003) 928–936

31. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J.: Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems* **47**(4) (2009) 547–553 <https://archive.ics.uci.edu/ml/datasets/wine+quality>
32. Rana, P.: Physicochemical properties of protein tertiary structure data set. UCI Machine Learning Repository (2013) <https://archive.ics.uci.edu/ml/datasets/Physicochemical+Properties+of+Protein+Tertiary+Structure>
33. Alpaydin, E., Kaynak, C.: Optical recognition of handwritten digits data set. UCI Machine Learning Repository (1998) <https://archive.ics.uci.edu/ml/datasets/optical+recognition+of+handwritten+digits>
34. LeCun, Y., Cortes, C., Burges, C.: Mnist handwritten digit database. AT&T Labs [Online] Available: <http://yann.lecun.com/exdb/mnist> (2010)
35. Sutskever, I., Martens, J., Dahl, G.E., Hinton, G.E.: On the importance of initialization and momentum in deep learning. *ICML* (3) **28**(1139-1147) (2013) 5