Acceleration Technique of Two-Phase Quasi-Newton Method with Momentum for Optimization Problems

Mastiyage Don Sudeera Hasaranga Gunathilaka, Shahrzad Mahboubi and Hiroshi Ninomiya
Shonan Institute of Technology
1-1-25 Tsujido-nishikaigan, Fujisawa, Kanagawa, 251-8511 Japan
Email: fujisanjp@live.com, 18T2012@shonan-it.ac.jp and ninomiya@info.shonan-it.ac.jp

***Abstract*—This paper describes a novel acceleration technique of the Two-Phase Quasi-Newton method using momentum terms for optimization problems. The performance of the proposed algorithm is evaluated on an unconstrained optimization problem in neural network training. The results show that the proposed algorithm has a much faster convergence than the conventional Two-Phase Quasi-Newton method.**

Keywords–neural networks; training algorithm; Two-Phase Quasi-Newton method; Nesterov's accelerated gradient; momentum terms.

I. Introduction

From the past to the recent years, much research has been conducted to improve the accuracy and speed of optimization for various problems. Neural Network (NN) training has a high performance when it comes to an unconstrained optimization problem. However, this ability highly depends on the training algorithm used in NNs. In this research, we aimed at constructing a training algorithm with higher performance and higher speed. Many algorithms have been proposed for this research task. In this paper, we consider the optimization problem of

$$\min_{\mathbf{w}\in R^n} \; E(\mathbf{w}), \tag{1}$$

where $\mathbf{w}$ and $E(\mathbf{w})$ denote the parameter and the objective function, respectively. Optimization problems have been solved with high precision by algorithms based on quadratic convergence characteristics, such as Newton or Quasi-Newton (QN) methods. This is because a solution can be obtained fast and with higher accuracy than algorithms with first-order convergence characteristics [1].

In recent years, for further acceleration and accurate optimization, the Two-Phase Newton method, which is based on third-order approximation, has been proposed [2]. However, this method requires a Hessian of (1) and matrix solutions for each iteration. Therefore, the Two-Phase Newton method takes time to derive a solution. To deal with this problem, the Two-Phase Quasi-Newton (Two-Phase QN) method has been proposed, in which the inverse Hessian is approximated by the gradient of (1) using an iterative formula [3]. This method is more effective for the unconstrained optimization problems, such as (1), than conventional algorithms. On the other hand, the acceleration of the standard QN with momentum terms was proposed as Nesterov's Accelerated Quasi-Newton (NAQ) method [4]. NAQ succeeded in drastically reducing the number of iterations and computational time compared to QN.

In this research, a new Quasi-Newton algorithm based on the third-order approximation is proposed for the acceleration of Two-Phase QN incorporating the momentum terms in the same way as NAQ. This method is referred to as Two-Phase Nesterov's Accelerated Quasi-Newton (Two-Phase NAQ) method. In this paper, the performance of the proposed algorithm is demonstrated through computer simulations using NN training for a simple unconstrained optimization problem and compared with the conventional method.

The contents of this paper are structured as follows: Section II introduces the conventional algorithms, such as Two-Phase Newton and Two-Phase QN. Section III proposes the novel algorithm, Two-Phase NAQ, which is the acceleration method of Two-Phase QN by introducing the momentum term. Section IV provides simulation results to demonstrate the validity of the proposed Two-Phase NAQ. Section V concludes this paper and describes future works.

II. Two-Phase Quasi-Newton method

The Two-Phase Newton method for optimization utilizes the gradient and Hessian of the objective function to result in the third-order approximation [2]. The iterative formulae of Two-Phase Newton are defined using the two-phase update scheme of the parameters (2) and (3).

$$\mathbf{z}_k = \mathbf{w}_k - \overline{\alpha}_k\big[\mathscr{H}(\mathbf{w}_k)\big]^{-1}\nabla E(\mathbf{w}_k), \tag{2}$$

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha_k\big[(1/2)\big(\mathscr{H}(\mathbf{w}_k) + \mathscr{H}(\mathbf{z}_k)\big)\big]^{-1}\nabla E(\mathbf{w}_k). \tag{3}$$

where $\nabla E(\mathbf{w}_k)$ and $\mathscr{H}(\mathbf{w}_k)$ are the gradient and the Hessian of (1). α_k and $\overline{\alpha}_k$ are the stepsizes at $\mathbf{w}_k$ and $\mathbf{z}_k$ along the directions $[\mathscr{H}(\mathbf{w}_k)]^{-1}\nabla E(\mathbf{w}_k)$ and $[(1/2)(\mathscr{H}(\mathbf{w}_k) + \mathscr{H}(\mathbf{z}_k))]^{-1}\nabla E(\mathbf{w}_k)$, respectively. Two-Phase Newton needs to calculate the Hessian and its inverse. Therefore, Two-Phase QN was proposed to reduce these computational costs by the approximation of the inverse Hessian using Broyden-Fletcher-Goldfarb-Shanno (BFGS) formulae [3] of QN. Actually, $\mathbf{H}(\mathbf{z}_k)$ which is the approximated inverse Hessian of $\mathscr{H}(\mathbf{z}_k)^{-1}$, is obtained by

$$\mathbf{H}(\mathbf{z}_k) = \Big(\mathbf{I} - \frac{\mathbf{s}_k\mathbf{y}_k^{\mathrm{T}}}{\mathbf{y}_k^{\mathrm{T}}\mathbf{s}_k}\Big)\mathbf{H}_k\Big(\mathbf{I} - \frac{\mathbf{y}_k\mathbf{s}_k^{\mathrm{T}}}{\mathbf{y}_k^{\mathrm{T}}\mathbf{s}_k}\Big) + \frac{\mathbf{s}_k\mathbf{s}_k^{\mathrm{T}}}{\mathbf{y}_k^{\mathrm{T}}\mathbf{s}_k}, \tag{4}$$

where $\mathbf{s}_k = \mathbf{z}_k - \mathbf{w}_k$ and $\mathbf{y}_k = \nabla E(\mathbf{z}_k) - \nabla E(\mathbf{w}_k)$. As a result, the iteration formulae of Two-Phase QN is shown as (5) and (6) based on [3].

$$\mathbf{z}_k = \mathbf{w}_k - \overline{\alpha}_k\mathbf{H}_k\nabla E(\mathbf{w}_k), \tag{5}$$

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha_k\mathbf{H}_{k+1}\nabla E(\mathbf{w}_k), \tag{6}$$

where $\mathbf{H}_{k+1}$ is calculated by

$$\mathbf{H}_{k+1} = \lambda\mathbf{H}_k + (1-\lambda)\mathbf{H}(\mathbf{z}_k). \tag{7}$$

III. Proposed algorithm - Two-Phase Nesterov's Accelerated Quasi-Newton method

In this section, the Two-Phase NAQ is proposed. Two-Phase QN is accelerated by using the momentum acceleration technique in the same way as NAQ. Specifically, Two-Phase NAQ is derived by the third-order approximation of (1) around $\mathbf{w}_k+\mu_k\mathbf{v}_k$, whereas (1) was approximated around $\mathbf{w}_k$ in Two-Phase QN [3]. The proposed method drastically improves the convergence speed of Two-Phase QN using the gradient vector at $\mathbf{w}_k+\mu_k\mathbf{v}_k$ of $\nabla E(\mathbf{w}_k+\mu_k\mathbf{v}_k)$ called Nesterov's accelerated gradient vector [4]. The iterative formulae of the proposed Two-Phase NAQ are defined as

$$\hat{\mathbf{z}}_k = \mathbf{w}_k + \mu_k\mathbf{v}_k - \overline{\alpha}_k\hat{\mathbf{H}}_k\nabla E(\mathbf{w}_k + \mu_k\mathbf{v}_k), \tag{8}$$

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \mu_k\mathbf{v}_k - \alpha_k\hat{\mathbf{H}}_{k+1}\nabla E(\mathbf{w}_k + \mu_k\mathbf{v}_k), \tag{9}$$

where, $\mu_k\mathbf{v}_k$ is the momentum term, in which $\mathbf{v}_k = \mathbf{w}_k - \mathbf{w}_{k-1}$ and μ_k is the momentum coefficient. In (8), $\hat{\mathbf{z}}_k$ denotes the middle-step suggested by the two-phase technique, in order to accelerate using the third-order approximation. (9) is considered as Two-Phase Newton method with the momentum term. The $\hat{\mathbf{H}}_k$ matrix is iteratively updated by

$$\hat{\mathbf{H}}_{k+1} = \lambda\hat{\mathbf{H}}_k + (1-\lambda)\hat{\mathbf{H}}(\hat{\mathbf{z}}_k), \tag{10}$$

Here, $\hat{\mathbf{H}}(\hat{\mathbf{z}}_k)$ is iteratively approximated by (11).

$$\hat{\mathbf{H}}(\hat{\mathbf{z}}_k) = \left(\mathbf{I} - \frac{\mathbf{p}_k\mathbf{q}_k^{\mathrm{T}}}{\mathbf{q}_k^{\mathrm{T}}\mathbf{p}_k}\right)\hat{\mathbf{H}}_k\left(\mathbf{I} - \frac{\mathbf{q}_k\mathbf{p}_k^{\mathrm{T}}}{\mathbf{q}_k^{\mathrm{T}}\mathbf{p}_k}\right) + \frac{\mathbf{p}_k\mathbf{p}_k^{\mathrm{T}}}{\mathbf{q}_k^{\mathrm{T}}\mathbf{p}_k}. \tag{11}$$

where, $\mathbf{p}_k = \hat{\mathbf{z}}_k - (\mathbf{w}_k+\mu_k\mathbf{v}_k)$ and $\mathbf{q}_k = \nabla E(\hat{\mathbf{z}}_k) - \nabla E(\mathbf{w}_k + \mu_k\mathbf{v}_k)$. In this paper, the momentum coefficient μ_k is set to the adaptive scheme suggested in [4]. The algorithm of Two-Phase NAQ is illustrated in Figure 1. In this research, λ is set to 0.5.

```
1. $k = 1$;
2. Initialize $\mathbf{w}_k$= random$[-0.5, 0.5]$, $\hat{\mathbf{H}}_k$ and $\hat{\mathbf{H}}(\hat{\mathbf{z}}_k) = \mathbf{I}$ (unit matrix) and $\mathbf{v}_k = \mathbf{0}$;
3. While($||\nabla E(\mathbf{w}_k)|| > \epsilon$ and $k < k_{max}$)
    (a) Update $\mu_k$;
    (b) Calculate $\nabla E(\mathbf{w}_k + \mu_k\mathbf{v}_k)$;
    (c) Update $\hat{\mathbf{z}}_k$ using (8);
    (d) Calculate $\nabla E(\hat{\mathbf{z}}_k)$;
    (e) Update $\hat{\mathbf{H}}_{k+1}$ using (10) and (11);
    (f) Update $\mathbf{w}_{k+1}$ using (9);
    (i) $k = k + 1$;
4. return $\mathbf{w}_k$;
```

Figure 1. Algorithm of the proposed Two-Phase NAQ.

IV. Simulation Results

Computer simulations are conducted to demonstrate the validity of the proposed Two-Phase NAQ for optimization problems. In this simulation, NN training is considered as an example for optimization problems. The function (12) is approximated using a feedforward NN with one hidden layer [4][5].

$$f(x) = 1 + (x + 2x^2)\mathrm{sin}(-x^2). \tag{12}$$

The input and output are x and $f(x)$, respectively. The sample dataset includes 400 training and 10,000 test points. The training and the test datasets are generated with 0.02 intervals and the random sampling in $x \in [-4, 4)$, respectively. The trained network has a hidden layer with 7 neurons. Therefore, the structure of the NN is 1-7-1. Each hidden neuron has a sigmoid function as the activation. In this research, the Mean Squared Error (MSE) is considered as the objective function of (1) for the training of the NN. 10 independent runs with $\mathbf{w}$ initialized by uniform random numbers in the range $[-0.5, 0.5]$ are conducted. The trained NN is estimated by the average, best and worst of $E_{train}(\mathbf{w})$ and $E_{test}(\mathbf{w})$, with the average of computational time (s) and the average of iteration count (k). The termination conditions are set to $\epsilon = 1.0\times10^{-6}$ and $k_{max} = 30,000$. The performance of Two-Phase NAQ is compared with the conventional Two-Phase QN [3]. The stepsizes α_k and $\overline{\alpha}_k$ for each algorithm are determined according to Armijo's conditions [4]. The simulation results of (12) are summarized in Table I. The table shows that the proposed Two-Phase NAQ converges faster than Two-Phase QN without loss of optimization properties. That is, the iteration counts and time of Two-Phase NAQ are much smaller than Two-Phase QN. At the same time, both algorithms here have comparable results for $E_{train}(\mathbf{w})$ and $E_{test}(\mathbf{w})$. The effect of increasing speed by Two-Phase NAQ is shown in Figure 2. Figure 2 shows the best training errors $E(\mathbf{w})$ for the iteration count (k) of Two-Phase QN and Two-Phase NAQ, which are $E_{train}(\mathbf{w}) = 0.67\times10^{-3}$ and $E_{train}(\mathbf{w}) = 0.31\times10^{-3}$, respectively. From this figure, it is shown that the errors of Two-Phase NAQ drastically decrease in the early stages of the training compared to Two-Phase QN. Furthermore, the calculation times per iteration of Two-Phase QN and Two-Phase NAQ are 0.25×10^{-3} and 0.27×10^{-3} (s), respectively. As a result, it is confirmed that the total simulation time of Two-Phase NAQ is faster than Two-Phase QN. This result is obvious from the following consideration. The summary of the computational cost is illustrated in Table II. 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 Two-Phase NAQ and Two-

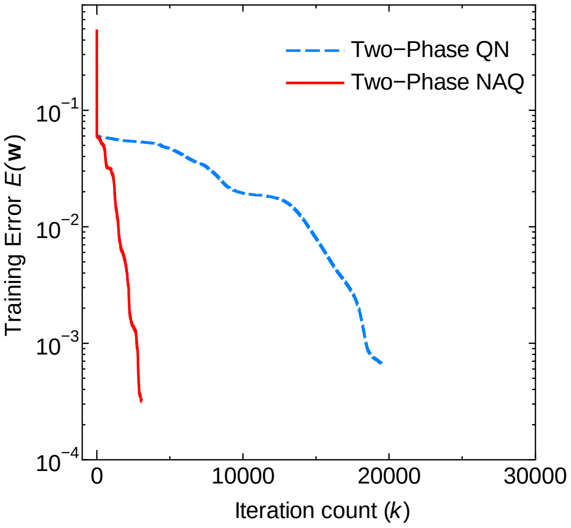

Figure 2. Plot of training error vs iteration count.

TABLE I. SUMMARY OF SIMULATION RESULTS OF (12).

Algorithm	$E_{train}(\mathbf{w})(\times\mathbf{10^{-3}})$	Time	Iteration	$E_{test}(\mathbf{w})(\times 10^{-3})$
	Ave / Best / Worst	(sec)	counts	Ave / Best / Worst
Two-Phase QN	6.85 / 0.67 / 18.64	4.49	17,200	6.66 / 0.66 / 18.12
Two-PhaseNAQ	5.67 / 0.31 / 18.61	**0.80**	**2,854**	5.56 / 0.31 / 18.08

TABLE II. SUMMARY OF THE COMPUTATIONAL COST.

Algorithm	Computational Cost
Two-Phase QN	$2nd + 2d^2 + 2\zeta nd$
Two-PhaseNAQ	$2nd + 2d^2 + 2\zeta nd$

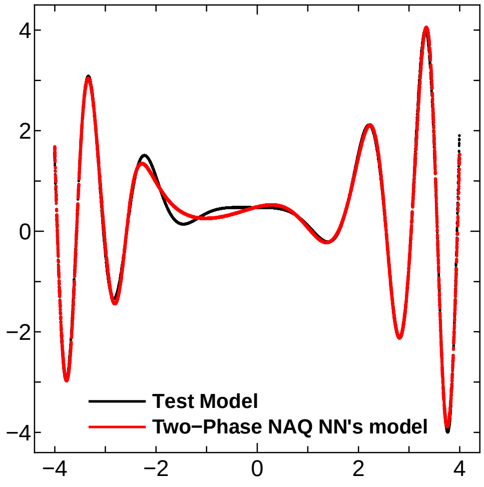

Figure 3. Plot of comparison between the test and the Two-Phase NAQ NN's model.

Phase QN compute the gradient twice per iteration. In both algorithms, the step length is determined by the line search method which involves ζ function evaluations until the search conditions are satisfied. As a result, it can be considered that the Two-Phase QN and the proposed Two-Phase NAQ require the same computational cost. For measuring the accuracy of modeling, the output of the neural model trained by Two-Phase NAQ with $E_{test}(\mathbf{w}) = 0.31 \times 10^{-3}$ is compared with the test data in Figure 3. Figure 3 shows a good match between the neural model and the test data.

V. Conclusion

In this research, we proposed a novel optimization algorithm, which was referred to as Two-Phase Nesterov's Accelerated Quasi-Newton (Two-Phase NAQ) method. The proposed algorithm was developed based on the third-order approximation method incorporating a momentum acceleration technique. The effectiveness of the proposed Two-Phase NAQ was demonstrated through computer simulations compared with the conventional Two-Phase QN for the training of NNs. From the simulation results, it can be concluded that the proposed method succeeded in surpassing the acceleration of Two-Phase QN without increasing the computational cost.

In the future, the convergence properties and further improvements of the proposed algorithm will be studied. Also, the validity of the proposed algorithm for large-scale and complicated real-world optimization problems, such as microwave circuit modeling [4], will be demonstrated.

References

[1] J. Nocedal and S.J. Wright, "*Numerical Optimization Second Edition*", Springer, 2006.

[2] D. K. R. Babajee and M. Z. Dauhoo, "An Analysis of The properties of the Variants of Newton's method with Third Order Convergence", *Applied Mathematics and Computations Journal*, Vol.1, No.1, 2006, pp. 659-684.

[3] S. K. Chakraborty and G. Panda, "A Two-Phase Quasi-Newton Method for Optimization Problem", Jul. 2018, arXiv preprint arXiv: 1807.11001.

[4] S. Mahboubi and H. Ninomiya, "A Novel quasi-Newton with Momentum Training for Microwave Circuit Models using Neural Networks", *Proc. ICECS'18*, Dec. 2018, pp. 629-632.

[5] N. Benoudjit, C. Archambeau, A. Lendasse, J. Lee and M. Verleysen, "Width optimization of the Gaussian kernels in radial basis function networks", *Proc. Eur. Symp. Artif. Neural Netw*, 2002, pp. 425-432.