Future Generation Computer Systems 89 (2018) 192–200

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

An adaptive control momentum method as an optimizer in the cloud

Jianhao Ding [1], Lansheng Han *,[2], Dan Li

School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

HIGHLIGHTS

- An optimizer Adacom is proposed based on an adaptive control system and momentum.
- Adacom is designed to alleviate oscillation and decrease the curvature.
- Adacom introduce reference model interacted with momentum to generate the update.
- The method can be used for optimizations in Autonomous Cloud and pervasive computing.
- Theoretical demonstration and evaluations prove its feasibility among the methods.

ARTICLE INFO

Article history:
Received 29 December 2017
Received in revised form 11 May 2018
Accepted 24 June 2018
Available online 2 July 2018

Keywords:
Gradient descent
Adaptive control system
Machine learning
Autonomous cloud

ABSTRACT

Many issues in the cloud can be transformed into optimization problems, where data is of high dimension and randomness. Thus, stochastic optimizing is a key to Autonomous Cloud. And one of the most significant discussions in this field is how to adapt the learning rate and convergent path dynamically. This paper proposes a gradient-based algorithm called Adacom, that is based on an adaptive control system and momentum. Critically inheriting the previous studies, a reference model is introduced to generate the update. The method reduces noise and decides on paths with less oscillation, while maintaining the accumulated learning rate. Due to system design properties, the method requires fewer hyper-parameters for tuning. We state the prospect of Adacom as a general optimizer in Autonomous Cloud, and explore the potential of Adacom for pervasive computing by the assumption of transition data. Then we demonstrate the convergence of Adacom theoretically. The evaluations over the simulated transition data prove the feasibility and superiority of Adacom with other gradient-based methods.

1. Introduction

The data in the cloud is varied, multidimensional and unstructured [1]. At present, the ecosystem of cloud computing is developing to multi-level, and the resource allocation of distributed cloud computing is changing strategically with the demand of industry and client [2]. The multiplication and mobility of devices that users access to the cloud have impacted the traditional network topology [3]. The traditional allocation response algorithm cannot cope with the mobile scene, and the computing originally in cloud servers has been gradually marginalized. Thus, the intelligent cloud response issue of edge devices needs to be addressed [4]. In many aspects, the client input and the operation have inconsistency in the representation of data [5]. In cloud environments, the realization of unified data platform will be the foundation of pervasive computing. In the future, with the trend of pervasive computing, the ability of cloud computing will be, to some extent, limited by general optimizers.

Demonstrated as a basic and efficient method for stochastic optimization, stochastic gradient descent has been widely applied as an optimizer in many scientific fields. On the one hand, many machine learning models can be ultimately projected to some objective function with the intent of searching local minima with respect to the parameters [6]. It has been argued that quasi-Newtonian methods work well in searching for the extrema, when the first and second derivatives are available in the domain of a small-scaled optimization. However, as the data size gets larger or the function becomes discrete, stochastic gradient descent performs efficiently and ideally. On the other hand, taking cyber security into consideration, the application of gradient optimizers can reduce the risk of leakage of raw information during transmission and improve data confidentiality [7].

* Corresponding author.
E-mail addresses: dingjh1998@hust.edu.cn (J. Ding), 1998010309@hust.edu.cn (L. Han).

[1] Academic Areas: Security of Big Data, Artificial Intelligence, Machine Learning.

[2] Academic Areas: Information Security, Security of Network, Security of Big Data. He received his PHD degree of Information Security in 2006. He published more than 30 papers and took part in more than 10 international conference.

https://doi.org/10.1016/j.future.2018.06.039
0167-739X/© 2018 Elsevier B.V. All rights reserved.

Stochastic gradient descent has been operating as by far the most common core optimization strategy in deep learning [8]. It is worth mentioning that the extension of its application is currently both explosive and inspiring. For instance, in neural computation, data flow between different neuron levels has been shown to be achieved through the algorithm [9]. Additionally, we can utilize stochastic gradient descent to boost the classification effects on statistical methods like the naive Bayes classifier [10]. In the area of deep learning, dropout regularization helps to prevent the emergence of overfitting for generalization with noisy data [11]. The recent advanced Deep Residual Network outperforms other Convolutional Neural Networks (CNN) in certain jobs [12]. It is widely accepted that almost all the machine learning frameworks mentioned above call for the stochastic gradient optimizer for optimization. A dedicated deep learning model can help adjust the parameters of a model in the cloud. For example, in [13], researchers use deep recurrent neural network to detect malware software. This paper seeks to introduce a novel algorithm to address these issues, which can dynamically adapt to the iterative gradient with the lower-order momentum parameter.

1.1. Motivation and contributions

Most gradient-based methods have natural limitations: for example, one major problem with the batch gradient descent is that the learning rate is fixed and hard to estimate in advance. When the parameter configuration is close to the local minima, it suffers from oscillation around the destination. Since random sampling leads to the combination of gradient and noise in the stochastic process, the convergent path appears dentate and serrated. Researchers have argued that the matter can be settled by importing a momentum algorithm from Physics. The basic form of a momentum algorithm enables the gradient to be accumulated and does, to some extent, fix the problem [14]. In actual fact, instead of a swift convergence, the speedup accounts for the oscillation or divergence, making the algorithm constrained in a technical sense.

Along with the advancement of deep learning, a series of algorithms have been developed that can automatically adapt to the learning rate, namely Adagrad [15], RMSProp [16], Adadelta [17], Adam [18], and so on. These methods are technically feasible, although some have one hyper-parameter that needs to be set manually, while others require tuning over two hyper-parameters. When tuning over parameters, we are unable to estimate the determinant effect towards the learning rate through direct calculation, as some methods are too robust to provide promising feedback over coefficients.

Although methods like Adam realize a stable annealing learning rate with the hyper-parameters, we discovered one drawback. The hitch is that, owing to the initial bias, the convergent path corresponds to the curve, which has large space curvature in all dimensions around the local extrema and thus costs tremendous time before a complete convergence. We believe that one property of perfect convergent paths is the least average space curvature. It can be proved that the shortest path is the route with the least curvature, and hence the consequence is a decline in processing time.

In this paper, we present an adaptive control momentum gradient-based algorithm called Adacom. The method is based on a first-order gradient that adapts to an adequate element wise learning rate at each iteration. This work attempts to damp the oscillation and obtain a fairly rapid learning rate using the interaction between the natural gradient and momentum. The paper contributes in the following ways:

(1) We propose Adacom for a smooth path and adaptive learning rates in stochastic optimization with only one hyper-parameter.
(2) We restate the focus of Autonomous Cloud in optimization models and point out Adacom as a general optimizer.
(3) We analyze the convergence of Adacom in theory and obtained the convergence rate.
(4) We assess the performance of Adacom, which is used to ascertain the effect as a general optimizer in the cloud.

1.2. Organization

The rest of the paper is organized as follows. In Section 2, some relevant gradient-based stochastic optimization algorithms are reviewed. In Section 3, Adacom is introduced at length in its different design aspects to reveal our original ideas; we also present a graph associated with the theoretical analysis. In Section 4, the vast scenarios in the cloud for implementations of Adacom have been investigated. In Section 5, the convergence of the algorithm is proved analytically based on a convex assumption. In Section 6, evaluations are conducted to illustrate the effectiveness of the algorithm empirically, and experimental results are presented. Finally, in Section 7, conclusions and future work are discussed.

2. Related work

Like other related algorithms, Adacom is a modification originating from natural gradient descent. In this section, we will describe the different updating patterns of each of the other prevalent algorithms [19].

A. Stochastic gradient descent with momentum

The momentum method inherits the momentum concept from Physics, with the intent of speeding up the learning process [20]. And it is especially employed on functions with stochastic noise or an ill-conditioned Hessian matrix. It accumulates the exponential moving averages. Given the gradient g of an objective function $f(\theta)$, the momentum is given by:

$$v_t = \alpha v_{t-1} + g_t \tag{1}$$

where α controls the decay. Theoretical and experimental demonstrations claimed that asymptotic local rate of convergence will be lost [21]. But Ilya et al. argued in [22] that the asymptotic convergence rate derived from momentum will improve the performance of deep network, even dominate the whole learning. Momentum and its modifications [20] reduce stochastic noise by average, but studies tend to overlook the fact that momentum cannot adjust in term of small gradients.

B. Adagrad

Duchi et al. formalized Adagrad (short for Adaptive Subgradient Methods) [15]. The method creates a bound that continually grows by a product gradient matrix, and utilizes the bound as a denominator to approximate the second order information [23], which is given by:

$$\Delta\theta = -\frac{\epsilon}{\delta + \sqrt{r}} g, \tag{2}$$

Although it performs well for sparse gradients over convex objective functions, because of the initial accumulating of larger gradients, learning rates decrease too fast [24]. These signals confuse the algorithm that cannot converge to the extrema. Nishant et al. also argued that Adagrad lacks the attention to correlations between components of gradients [25].