View Reviews

Paper ID

592

Paper Title

Optimistic Acceleration for Adaptive Optimization

Reviewer #1

Questions

1. Please enter a summary of the paper, its contributions, and its potential impact inside and outside of the UAI community. (See http://www.auai.org/~w-auai/uai2020/reviewer_instructions.php)

This paper proposed a new variant of AMSGrad by using the idea of Optimistic Online Learning called Optimistic-AMSGrad. The author(s) provide theoretical analysis for minimizing the regret. Numerical experiments are conducted on various data sets (MNIST, CIFAR-10, CIFAR-100, IMDB) and network architectures to show the benefits of the proposed algorithm. I think this paper may have some potential impact on the practical perspective.

2. Please enter detailed comments of the strengths and weaknesses of the submission. (See http://www.auai.org/~w-auai/uai2020/reviewer_instructions.php)

Strengths:
- Propose a new variant of AMSGrad by using the idea of Optimistic Online Learning
- Numerical experiments show good overall performance over AMSGrad on various data sets and network architectures.

Weaknesses:
- It is unclear about the theoretical contribution of this paper. Could you please provide some advantages of your theory over the existing ones?
- Theoretical results are regret analysis for convex loss but experiments are on deep learning applications, which are non-convex.
- The analysis for beta_1 = 0 is much weaker than the general case.

Comments:
- I think the goal is to show Regret_T / T -> 0. However, in Theorem 1 and Corollary 1, it is unclear to me why Regret_T / T -> 0 unless you need to make some additional assumptions.
- I understand that the analysis for \beta_1 \neq 0 could be hard but you can consider the case where beta_1 is dependent on t and diminishing.

At this moment, I would like to receive the response from the author(s) to see how they address my concerns. I am happy to raise my score after the rebuttal period if it is reasonable.

3. Please provide an overall score for the submission.

Weak Reject: Borderline, tending to reject

6. Please rate your confidence in the score assigned.

The reviewer is confident but not absolutely certain that the evaluation is correct.

Reviewer #2

Questions

1. Please enter a summary of the paper, its contributions, and its potential impact inside and outside of the UAI community. (See http://www.auai.org/~w-auai/uai2020/reviewer_instructions.php)

This paper proposes an adaptive algorithm that incorporates the idea of optimistic online learning into the well-known adaptive method, AMSGrad. When there is always a good guess of the gradient at each step, the authors show an accelerated convergence of the proposed algorithm over AMSGrad. The proposed algorithm is also compared experimentally with AMSGrad under a common epoch basis.

2. Please enter detailed comments of the strengths and weaknesses of the submission. (See http://www.auai.org/~w-auai/uai2020/reviewer_instructions.php)

Pro:
The experimental results are encouraging. It shows that Optimistic-AMSGrad converges faster than its counterpart, AMSGrad, on deep neural networks. Moreover, it seems to improve the test performance as well.

Con:
Theoretical analysis of the paper does not really show acceleration of the proposed algorithm. Note that the “acceleration” analysis in the paper stopped at Theorem 1 and Corollary 1. There is no further analysis based on the proposed gradient prediction strategy.
Note that whether the proposed method accelerates depends on how the norm of g_t-m_t compares to the norm of g_t. Indeed, under stochastic setting, there is always a significant stochastic noise on g_t, which is independent of, hence not predictable by, the history of the gradients, g_{t-1}, g_{t-2}, ….g_{t-r}. Namely, the norm of g_t-m_t can not be very small compared to norm of g_t, no matter how the gradient is predicted. Therefore, "acceleration" is not achieved.

Minor:
Eight lines above Eq. 1: m_t should be a guess of the gradient, not the loss function l_t.

3. Please provide an overall score for the submission.

Weak Reject: Borderline, tending to reject

6. Please rate your confidence in the score assigned.

The reviewer is fairly confident that the evaluation is correct.

Reviewer #3

Questions

1. Please enter a summary of the paper, its contributions, and its potential impact inside and outside of the UAI community. (See http://www.auai.org/~w-auai/uai2020/reviewer_instructions.php)

This paper proposes a new adaptive gradient method named Optimistic-AMSGrad. Compared with AMSGrad, this algorithm utilizes an additional estimator m_t to estimate the gradient g_t at round t. By existing proof techniques from optimistic online learning, the authors show that Optimistic-AMSGrad has a lower regret bound than AMSGrad. The authors also carry out some numerical experiments to show the superiority of their algorithm.

2. Please enter detailed comments of the strengths and weaknesses of the submission. (See http://www.auai.org/~w-auai/uai2020/reviewer_instructions.php)

In Section 3.1, the authors derived the convergence result of Optimistic-AMSGrad and compare the result with that of AMSGrad. Then the authors claim that when the gradient estimator m_t is ‘sufficiently close’ to the true gradient g_t, the last term of (4) is smaller than that of (5), which implies that the regret of Optimistic-AMSGrad is smaller than that of AMSGrad. I wonder what is the precise definition of ‘sufficiently close’ ? This claim confuses me because it is easy to show that when the true gradient g_t is sparse, the last term of (5) will be very small (as \hat v_t is defined by g_t recursively), while the last term of (4) still remain large for most estimators m_t (for instance, m_t = g_{t-1}). Can the authors show some sufficient or necessary conditions that m_t should satisfy to make their claim valid?

In Section 3.2, the authors mentioned several works on the converge analysis of Adam-type algorithms. However, the authors only commented on one of these works in detail. Can the authors elaborate other works?

In the experiment part, the authors only compare their algorithm with AMSgrad and optimistic Adam, with respect to the number of iterations. Although the number of iterations has been regarded as a standard measure for the comparison of different optimization algorithms in the deep learning training, the wall-clock running time is more prefered because it can reflect the true computational cost in practice. Given the discussion of the computational cost in Section 6.1, can the authors also add one or two additional experiments and plot the running time for each algorithm? Furthermore, since the proposed algorithm is an improvement version of AMSgrad, it is also important to compare with other variants of AMSgrad/Adam, such as Padam [1] and Adabound [2].

[1] Chen, Jinghui, et al. "Closing the generalization gap of adaptive gradient methods in training deep neural networks." arXiv preprint

arXiv:1806.06763 (2018).
[2] Luo, Liangchen, et al. "Adaptive gradient methods with dynamic bound of learning rate." arXiv preprint arXiv:1902.09843 (2019).

3. Please provide an overall score for the submission.
Weak Reject: Borderline, tending to reject

6. Please rate your confidence in the score assigned.
The reviewer is confident but not absolutely certain that the evaluation is correct.

Reviewer #5

Questions

1. Please enter a summary of the paper, its contributions, and its potential impact inside and outside of the UAI community. (See http://www.auai.org/~w-auai/uai2020/reviewer_instructions.php)
The paper proposes a modification to AMSGrad that incorporates gradient prediction, and demonstrates its efficacy for two toy tasks, and training deep neural networks on image classification tasks.

2. Please enter detailed comments of the strengths and weaknesses of the submission. (See http://www.auai.org/~w-auai/uai2020/reviewer_instructions.php)
The paper is well written and easy to follow. The paper extends AMSGrad to utilize a prediction of the gradient. Arguably, the method is very closely related to variance reduction methods [1, 2]. For example, the paper should compare to [1] (cited over 1000 times). Additionally, it would be useful to include (1) training accuracy as a function of training time, and (2) test accuracy at convergence. Without these experiments, it is impossible to judge the empirical value of the proposed optimizer, which forces me to vote for rejection.

To expand on the connection of the proposed optimizer to variance reduction methods. As shown by the authors in Theorem 1, the regret of Optimistic-AMSGrad depends on the ||m_t - g_t|| term. In offline setting, this is minimized if m_t=g, where g is the full-batch gradient. Variance reduction methods focus on reducing ||g - \hat g_t|| and \hat g_t is an estimate of the gradient at timestep t. This term is also minimized if g=\hat g_t.

3. Please provide an overall score for the submission.
Weak Reject: Borderline, tending to reject

6. Please rate your confidence in the score assigned.
The reviewer is fairly confident that the evaluation is correct.