View Reviews

Paper ID
6346

Paper Title
Understanding and detecting convergence for stochastic gradient descent with momentum

Reviewer #1

Questions

1. Please summarize the main claim(s) of this paper in two or three sentences.
The paper studies the SGD with heavy ball momentum and shows that there exists a transient phase in which iterates a shows that there exists a transient phase in which iterates move towards a region of interest, and a stationary phase in which iterates remain bounded in that region around a minimum point. The authors provide theoretical and empirical support for the design choices for the convergence diagnostics and demonstrate the effect of momentum on the statistic of the diagnostic.

2. Merits of the Paper. What would be the main benefits to the machine learning community if this paper were presented at the conference? Please list at least one.
The paper is well written and the contributions are clear.
The statements of the Theorems are robust and the proofs are nicely presented.

3. Please provide an overall evaluation for this submission.
Borderline paper, but the flaws may outweigh the merits.

4. Score Justification Beyond what you've written above as "merits", what were the major considerations that led you to your overall score for this paper?
There is a major issue on the assumptions used that does not allow me to suggest acceptance. Please find below my argument. If the authors are able to handle the main issue raised below I will be happy to suggest acceptance however I believe that this will require new theoretical approach which could be time-consuming and not trivial.

Major Issue:
For all theoretical results the authors assume assumption 1 (strongly convex function) and assumption 3 which requires $\|\nabla f(\theta)\| \leq G$ and $\Exp[\|\nabla \ell(\theta ,\xi) 0\nabla (\theta)\|^2]\leq\delta$. In other words assumption 3 implies that $\Exp[\|\nabla \ell(\theta ,\xi) \|^2] \leq \delta ^2 +G^2 $ (bounded gradients)
However as explained recently in [1] and [2] (see references below) these two assumptions lead to contradiction and to an empty set of functions. A function cannot be strongly convex and at the same time has bounded gradients at least in the unconstrained case.

5. Detailed Comments for Authors Please comment on the following, as relevant: - The significance and novelty of the paper's contributions. - The paper's potential impact on the field of machine learning. - The degree to which the paper substantiates its main claims. - Constructive criticism and feedback that could help improve the work or its presentation. - The degree to which the results in the paper are reproducible. - Missing references, presentation suggestions, and typos or grammar improvements.
Other comments:
1) The work of Loizou and Richtrarik on SGD with momentum mentioned in Appendix A of the paper did not make the above two assumptions in order to prove convergence. On the other hand they focus on the quadratic case with interpolation assumption. This is the reason for the linear convergence to the exact solution.

2) The work [2] bellow also provide theorem explaining how one can select the step-size of SGD (without momentum) for the transient phase and stationary phase.

3) For SGD with momentum very closely work is [3] bellow where the transient and stationary phase are explained.

Important missing references:
[1] Nguyen, Lam M., Phuong Ha Nguyen, Marten van Dijk, Peter Richtárik, Katya Scheinberg, and Martin Takáč. "SGD and Hogwild! convergence without the bounded gradients assumption." ICML 2018
[2] Gower, Robert Mansel, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtarik. "SGD: General Analysis and Improved Rates." ICML 2019
[3] Goh, "Why Momentum Really Works", Distill, 2017. http://doi.org/10.23915/distill.00006

6. Please rate your expertise on the topic of this submission, picking the closest match.
I have published one or more papers in the narrow area of this submission.

7. Please rate your confidence in your evaluation of this paper, picking the closest match.
I am very confident in my evaluation of the paper. I read the paper very carefully and I am very familiar with related work.

8. Datasets If this paper introduces a new dataset, which of the following norms are addressed? (For ICML 2020, lack of adherence is not grounds for rejection and should not affect your score; however, we have encouraged authors to follow these suggestions.)
This paper does not introduce a new dataset (skip the remainder of this question).

12. I agree to keep the paper and supplementary materials (including code submissions and Latex source) confidential, and delete any submitted code at the end of the review cycle to comply with the confidentiality requirements.
Agreement accepted

13. I acknowledge that my review accords with the ICML code of conduct (see https://icml.cc/public/CodeOfConduct).
Agreement accepted

Reviewer #2

Questions

1. Please summarize the main claim(s) of this paper in two or three sentences.
The authors analyze the impact of momentum in the detection of the stationary phase of SGD.

2. Merits of the Paper. What would be the main benefits to the machine learning community if this paper were presented at the conference? Please list at least one.
Detecting the stationary phase of SGD is crucial for the design of the learning rate and momentum schedules for minimizing functions.

3. Please provide an overall evaluation for this submission.
Borderline paper, but the flaws may outweigh the merits.

4. Score Justification Beyond what you've written above as "merits", what were the major considerations that led you to your overall score for this paper?
One main contribution consists in the adaptation of the Pflug diagnostic is a straightforward way. Instead of considering [theta_n-theta_{n-1}]^T[theta_{n-1}-theta_{n-2}] (original test for SGD), they consider the scalar products of gradients when momentum is used. This is a very limited contribution.
The second contribution is the analysis of the impact of momentum in the detection. However, the results are quite unclear and this is difficult to identify the limit of the theory.

5. Detailed Comments for Authors Please comment on the following, as relevant: - The significance and novelty of the paper's contributions. - The paper's potential impact on the field of machine learning. - The degree to which the paper substantiates its main claims. - Constructive criticism and feedback that could help improve the work or its presentation. - The degree to which the results in the paper are reproducible. - Missing references, presentation suggestions, and typos or grammar improvements.
Here is a bunch of remarks I have on the paper and appendices.

- Precise in theorem 3.1 how the bound can be negative.

Although this is common knowledge, as there is no space limitation in the appendix, write explicitly the properties used for strong convexity.

Notation clash: checking period and strong convexity parameter both use "c" as a notation. I recommend using "\mu" for strong convexity, as this is more common in optimization.

Theorem 3.1: technical remark: On what random variable the assumption is taken?
- Only xi_{n+1}? If this is the case, the proof can be simplified by considering only
E[nabla l(theta_n,xi_{n+1})]^T nabla l(theta_n-1,xi_n) (i.e., the expectation only on the first gradient).
This would make the proof clearer, but I haven't check if the rest of the proof also holds in this case.
- On both xi_n and xi_n+1? In this case, then theta_n also depends on xi_n. Thus it should be written E[theta_n]. Moreover, the gradient also depends on theta_n, itself function of x_n, which means $E[\nabla l(theta_n,xi_n+1)]$ is not equal to $\nabla l(theta_n)$.

In the second case, this seems to be a big flaw as this completely break the proof. In the (highly probable) case I am wrong, I suggest the author clarify this part of the proof.

Also, theorem 3.1 in the appendix, why is this smaller than zero? This is stated as property. However, this seems to be more a condition over gamma and beta.

Theorem 3.1: As this uses lemma B1, it should also use Assumption 5.

Theorem 3.2, why not considering something else than beta in the expectation? Maybe this can lead to a better expression. I.e., we can track nabla l + coef*(theta_n-theta_n-1) for some coefficient.

Quadratic loss model: the noise model seems multiplicative (the first iteration reads epsilon x1). This is known to have better properties than additive noise. Why do you not consider standard noise like nabla L(x,y) + epsilon?

Theorem 3.4 is somewhat unclear: what is the model of the noise used here?

The algorithm is pretty unclear: there are too many details that are better explained in plain English rather than with pseudo-code. Like, what is the function h? what is the purpose of burnin? Etc.

Why proposition 4.1 is a proposition? I do not see any proof of this claim. This looks more like an observation.

6. Please rate your expertise on the topic of this submission, picking the closest match.
I have closely read papers on this topic, and written papers in the broad area of this submission.

7. Please rate your confidence in your evaluation of this paper, picking the closest match.
I tried to check the important points carefully. It is unlikely, though possible, that I missed something that could affect my ratings.

8. Datasets If this paper introduces a new dataset, which of the following norms are addressed? (For ICML 2020, lack of adherence is not grounds for rejection and should not affect your score; however, we have encouraged authors to follow these suggestions.)
This paper does not introduce a new dataset (skip the remainder of this question).

12. I agree to keep the paper and supplementary materials (including code submissions and Latex source) confidential, and delete any submitted code at the end of the review cycle to comply with the confidentiality requirements.
Agreement accepted

13. I acknowledge that my review accords with the ICML code of conduct (see https://icml.cc/public/CodeOfConduct).
Agreement accepted

Reviewer #3

Questions

1. Please summarize the main claim(s) of this paper in two or three sentences.
This paper designs a convergence diagnostic for SGD with momentum. This diagnostic is to detect the entrance of the iterates to the stationary phase. Some heuristical theoretical claims are provided.

2. Merits of the Paper. What would be the main benefits to the machine learning community if this paper were presented at the conference? Please list at least one.
1. This paper considers the non-zero momentum case, compared to the previous work (Chee & Toulis (2018)).

3. Please provide an overall evaluation for this submission.
Below the acceptance threshold, I would rather not see it at the conference.

4. Score Justification Beyond what you've written above as "merits", what were the major considerations that led you to your overall score for this paper?

Many theoretical claims in this paper are vague, incomplete, and therefore not persuasive.

1. In Theorem 2.1, a convergence bound for SGDM is provided. However, the statement is quite incomplete. E.g., "Under additional assumptions of the loss", but these assumptions are never listed. Furthermore, the stepsize range is not mentioned at all, which I found very confusing. Afterall, stepsize cannot be taken arbitrarily.

G and \delta are used in this theorem but they are actually defined later.

2. On line 158 left column, it is said that "This does not greatly effect the con-
vergence rate as momentum is most useful in the transient phase. " Why is momentum not that much useful in the stationary phase? It may actually affect the radius of the stationary distribution by Theorem 2.1.

3. Many parameters in the main algorithm (Alg 1) are missing. The heuristic convergence function h is not defined, the threshold T is not specified. Checking period c is missing, the parameter burnin is missing.

4. In the main assumption 1-5, many parameters are not specified and they actually play important roles in the latter theory. E.g., M and K. The assumption 3 is not justified. If the variance of the stochastic gradient is lower bounded below by \delta_0, how large can \delta_0 be? In the case of overfitting (think of huge neural networks), \delta_0 should actually be very small around global optima (each training example is fitted, the training accuracy is 100%).

5. Detailed Comments for Authors Please comment on the following, as relevant: - The significance and novelty of the paper's contributions. - The paper's potential impact on the field of machine learning. - The degree to which the paper substantiates its main

claims. - Constructive criticism and feedback that could help improve the work or its presentation. - The degree to which the results in the paper are reproducible. - Missing references, presentation suggestions, and typos or grammar improvements.

Let me continue my comments

5. By theorem 3.1 and 3.2, the authors claim that "The convergence threshold for a good test statistic should not depend too much on momentum". This is not a valid claim.

First, the upper bounds in Theorems 3.1 and 3.2 may not be tight, one cannot say that the bound in Theorem 3.2 is larger so the alternative test statistic is worse.

Second, even if the second test statistic is worse, one cannot say that "a good test statistic should not depend too much on momentum". There may be other ones that do depend on momentum and are better.

6. Corollary 3.3 is important in this paper since it justifies why the proposed statistic is good. However, it requires a large stepsize. The stepsize lower bound depends on the (unspecified) quantities M and K and \sigma_0. Is this feasible? The lower bound may be actually larger than the step size required in Theorem 2.1

7. It is claimed that "Corollary 3.3 suggests that too large momentum may make the learning rate condition too prohibitive, invalidating the negative expectation in practice." But A_0 and A_\beta may not be that different since beta is in [0,1), and the resulting stepsize lower bound is not that different.

8. In section 3.3, the theory is justified on a simple quadratic model. However, this model is too simple and I suggest consider a finite sum of quadratic functions at least.

9. The statements in Proposition 4.1 is too vague to put as a proposition. E.g., "The expectation is negative due to a relatively sparse number of inner products which have high magnitude and negative sign. In these few key inner products the true gradient dominates because of an interaction with the loss curvature." The sparse number of inner products, high magnitude should be quantified.

10 Similarly, the statements on line 315-323 left column are too vague. The authors can consider providing the value of \lambda in Corollary 4.3, how large can it be?

6. Please rate your expertise on the topic of this submission, picking the closest match.
I have closely read papers on this topic, and written papers in the broad area of this submission.

7. Please rate your confidence in your evaluation of this paper, picking the closest match.
I am very confident in my evaluation of the paper. I read the paper very carefully and I am very familiar with related work.

8. Datasets If this paper introduces a new dataset, which of the following norms are addressed? (For ICML 2020, lack of adherence is not grounds for rejection and should not affect your score; however, we have encouraged authors to follow these suggestions.)
This paper does not introduce a new dataset (skip the remainder of this question).

12. I agree to keep the paper and supplementary materials (including code submissions and Latex source) confidential, and delete any submitted code at the end of the review cycle to comply with the confidentiality requirements.
Agreement accepted

13. I acknowledge that my review accords with the ICML code of conduct (see https://icml.cc/public/CodeOfConduct).
Agreement accepted