← Back to **Author Console** (/group?id=AAAI.org/2025/Conference/Authors#your-submissions)

ASRO: Sign-Based Optimization for Stochastic Learning

PDF (/pdf?id=u57ZMDYari)

Rhugved Chaudhari (/profile?id=~Rhugved_Chaudhari1), Yashodhara V Haribhakta (/profile?id=~Yashodhara_V_Haribhakta1)

06 Aug 2024 (modified: 10 Dec 2024) Submitted to AAAI 2025 Conference, Area Chairs, Senior Program Committee, Program Committee, Authors Revisions (/revisions?id=u57ZMDYari) BibTeX CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

Primary Keyword: Machine Learning (ML) -> ML: Optimization
Secondary Keywords: Reasoning under Uncertainty (RU) -> RU: Stochastic Optimization, Machine Learning (ML) -> ML: Deep Learning Algorithms

Abstract:

Over the years, numerous stochastic optimization algorithms have emerged for minimizing objective functions with respect to their parameters, and these have been extensively used to train neural networks. Among these, adaptive learning rate algorithms have seen significant success in recent years. Despite the advent and success of adaptive algorithms, there remains a need for separate and dedicated global learning rate scheduling strategies to train very deep networks on large datasets and achieve optimal performance. This paper introduces "Asro" - Adaptive Sign Reversal Optimization, a straightforward stochastic optimization algorithm that eliminates the need for a dedicated learning rate scheduler. Asro presents a novel principle, Sign-directed rate adaptation, which aims to detect the vicinity of local minima and adjust the learning rate based on the sign reversal of successive first moments (exponential averages) of the gradients. Additionally, the paper proposes 'AccAsro,' a variant of Asro that utilizes the sign preservation of gradients across successive timesteps to dynamically increase the learning rate. This approach also simplifies the search for optimal hyper-parameters, as explained in the paper. We empirically demonstrate that Asro achieves improved loss convergence compared to mainstream optimization algorithms across a wide variety of tasks, such as image classification and language modeling.

Supplementary Material: zip (/attachment?id=u57ZMDYari&name=supplementary_material)
iThenticate Agreement: Yes, I agree to iThenticate's EULA agreement version: v1beta
Submission Number: 7406

Filter by reply type... Filter by author... Search keywords...

Sort: Newest First – = ≡

Everyone Program Chairs Submission7406 Area... Submission7406... *10 / 10 replies shown*

Submission7406 Authors Submission7406... ✖

Add: **Withdrawal**

Paper Decision

Decision by Program Chairs 10 Dec 2024, 03:41 (modified: 10 Dec 2024, 04:38)

Program Chairs, Area Chairs, Senior Program Committee, Authors Revisions (/revisions?id=n25zlaICDh)

Decision: Reject

Comment:
This paper introduces novel stochastic optimization algorithms that can eliminate the need for a dedicated learning rate scheduler in adaptive optimization methods. The authors also conduct experiments aiming to empirically demonstrate the proposed Asro method can achieve improved loss convergence. Based on the reviewers' ratings and feedback, we recommend rejecting this manuscript. Moreover, this paper also exceeds the page limit.

A Novel Optimizer with Limited Empirical Evidence

Official Review by Program Committee Pbni 02 Nov 2024, 11:10 (modified: 10 Dec 2024, 04:08)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors
Revisions (/revisions?id=9pAW4JuVCn)

Review:
The paper introduces a novel optimization algorithm called ASRO (Adaptive Sign Reversal Optimization). This algorithm aims to enhance the training of deep learning models by eliminating the need for a dedicated learning rate scheduler. The core principle of ASRO is Sign-directed rate adaptation, which adjusts the learning rate based on the sign reversal of successive first moments (exponential averages) of the gradients.

The experimental results show that Asro can solve the challenges posed by saddle points, plateaus, and other complex features of the loss landscape that typically hinder convergence in deep learning models.

Pros:

Elimination of Dedicated Learning Rate Scheduler: ASRO eliminates the need for a separate global learning rate scheduling strategy by incorporating a novel principle called Sign-directed rate adaptation. This principle adjusts the learning rate based on the sign reversal of successive gradient moments, which helps in detecting the vicinity of local minima and adjusting the learning rate accordingly.

Efficient Navigation of Saddle Points and Plateaus: ASRO's ability to detect and navigate saddle points and plateaus in the loss landscape more efficiently is a key advantage. By monitoring the sign changes in the first moments (exponential moving averages) of gradients, ASRO can adjust learning rates adaptively, allowing the optimizer to accelerate out of flat regions and decelerate appropriately near minima. This approach mitigates the risk of oscillations and promotes steady convergence, especially in high-dimensional, non-convex optimization problems typical in deep learning applications

Cons:

Sensitivity to Hyperparameters: Although ASRO aims to simplify the search for optimal hyperparameters, it still requires careful tuning of parameters such as the learning rate scaler and the decrement factor. Incorrect settings can lead to suboptimal performance or even divergence during training.

Limited Empirical Validation: While the paper demonstrates ASRO's effectiveness across various tasks, a more extensive empirical evaluation on a wider range of applications and datasets would further solidify its generalizability. Both Asro and AccAsro may exhibit initial oscillations, particularly AccAsro due to its reliance on direct gradient signs. While Figure 1 presents results for AccAsro, a more detailed comparison between Asro and AccAsro would provide further insights into their relative performance and stability.

Rating: 5: Marginally below acceptance threshold
Confidence: 2: The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper

Rebuttal by Authors

Rebuttal
by Authors (Yashodhara V Haribhakta (/profile?id=~Yashodhara_V_Haribhakta1), Rhugved Chaudhari (/profile?id=~Rhugved_Chaudhari1))
09 Nov 2024, 17:12 (modified: 11 Nov 2024, 22:08)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors
Revisions (/revisions?id=aY3DAsM9MD)

Rebuttal:
Thank you very much for your insightful and constructive review of our paper.

1. Elimination of Learning Rate Scheduler
2. Efficient Navigation of Saddle Points and Plateaus

- Thanks a lot for acknowledging this.
- Note: Also, faster convergence and ease of hyperparameter tuning.

3. Sensitivity to Hyperparameters

- Thank you for your feedback. While tuning is still necessary, as noted in Section 3.3.1, Asro simplifies this by allowing a range for the learning rate rather than a fixed value. To support users further, we will add guidance on selecting the learning rate scaler and decrement/increment factors in the revision. We kindly refer you to our response to Reviewer "3opC" (Point 5).
- Regarding divergence it is extreme and cannot occur unless the upper LR threshold is set to an extremely high unconventional value. In the context of other optimizers, this case is similar to setting the learning rate extremely high, which is almost negligibly done in real practice.

4. Regarding “Limited Empirical Validation”:

- Thank you for your feedback regarding the empirical validation. While we agree that an evaluation across a wider range of tasks could have been included in the paper, we believe our paper presents “sufficient” empirical validation, for the following reasons:
- Many widely adopted optimizers, including Adam, were originally validated on only a few tasks (e.g., logistic regression, DNNs, CNNs (in case of Adam)). Despite this limited scope, Adam has demonstrated broad applicability, which we believe similarly supports Asro’s generalizability based on our results.
- Since Asro is built on Adam’s framework with the primary modification being the adaptive “learning rate” adjustment (sign reversal optimization), the established reliability of Adam in deep learning contexts further supports Asro's applicability.
- Regarding the inclusion of Asro in Figure 1, we clarify that Asro works on decelerating the learning rate based on the sign reversal rate adaptation principle. The effect of which is seen in a longer training duration only, when the parameters are near their optimum values. Hence, we found it out of context to include Asro in this particular experiment.
- Rigorous Evaluation Across Key Tasks: We have included rigorous evaluations across two highly relevant fields: Language Modelling (NLP) and Computer Vision (CNNs). Our experiments encompass complete boundary conditions such as higher and lower learning rates and with and without cosine decay.

Official Review

Official Review by Program Committee LjCb 15 Oct 2024, 15:56 (modified: 10 Dec 2024, 04:08)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors

Revisions (/revisions?id=05WnkhUi5r)

Review:
This paper is out of the page length. It has 8 pages of main text while the permitted max page length is 7. It should be desk-rejected.

Rating: 3: Clear rejection
Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Rebuttal by Authors

Rebuttal
by Authors (Yashodhara V Haribhakta (/profile?id=~Yashodhara_V_Haribhakta1), Rhugved Chaudhari (/profile?id=~Rhugved_Chaudhari1))

09 Nov 2024, 17:11 (modified: 11 Nov 2024, 22:08)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors

Revisions (/revisions?id=hSRFhDx04Z)

Rebuttal:

1. “This paper is out of the page length. It has 8 pages of main text, while the permitted max page length is 7. It should be desk-rejected.”

- We sincerely apologize for this mistake and any inconvenience it may have caused. We clarify and accept that the main text spans approximately 7.25 pages. We would like to kindly bring forward that this was not an intentional oversight but rather a misunderstanding regarding what all constitutes “main text” under the submission guidelines. We mistakenly thought that the “main text” includes the abstract, algorithms, and experiments but not necessarily the conclusion. As a result, a few short paragraphs of the conclusion exceeded the limit by about a quarter page.
- We sincerely apologize for this and assure you that we will correct it in a revision. To ensure that we do not change the paper's relevant contents, we will cut short the conclusion itself to a great extent.
- Given this context, we request you to forgive this and humbly and respectfully request that the paper still be considered for your review. As reviewers ‘3opC’ and ‘Pbni’ noted, the paper introduces Asro and Sign Reversal Optimization—a novel approach that effectively navigates saddle points and plateaus, achieving faster and lower convergence and eliminating the need for a dedicated learning rate scheduler. Additionally, the method significantly eases hyperparameter tuning, especially for the learning rate. Through evaluations in language modelling and image recognition tasks, we have observed that Asro consistently outperforms the baselines.
- We hope you find the contribution valuable and well-presented, and we would be grateful if you review the paper. We humbly request you to not let the formatting error overshadow the paper’s contributions, with a rejection.

Humble Request for Consideration Despite Formatting Issue Regarding Page Limit and Concern Regarding Desk-Rejection

Author SPC Confidential Comment

by Authors (Yashodhara V Haribhakta (/profile?id=~Yashodhara_V_Haribhakta1), Rhugved Chaudhari (/profile?id=~Rhugved_Chaudhari1))

09 Nov 2024, 17:22 (modified: 09 Nov 2024, 17:24)

Program Chairs, Area Chairs, Senior Program Committee, Authors

Revisions (/revisions?id=aWSIKhySUN)

Comment:

- We respectfully bring to your attention a concern regarding the review (LjCb), which recommends desk-rejection due to a page limit issue. We fully understand the importance of adhering to submission guidelines, and we sincerely apologize for this. We would like to politely clarify that this was not an oversight of the rules, but rather a misunderstanding about what content constitutes “main text.” The paper’s main text currently spans approximately 7.25 pages due to the misunderstanding. We were unsure of and hence mistakenly believed that the abstract, algorithms, and experiments were qualified as the “main text” but not the conclusion, resulting in a brief overflow of the conclusion into the eighth page (quarter page).
- We take full responsibility for this mistake and have assured the reviewer that, if given the chance, we will adjust the formatting and reduce the conclusion to meet the page limit precisely, in a revision. To ensure that we do not change the paper's relevant contents, we will cut short the conclusion itself to a great extent.
- This minor formatting issue was in no way intended to circumvent submission rules, and we hope it does not detract from the value of our contribution. As highlighted by reviewers 3opC and Pbni, our paper introduces Asro and Sign Reversal Optimization, a novel and impactful approach that effectively accelerates through plateaus and saddle points, achieves faster convergence, and eases hyperparameter tuning while eliminating the need for a dedicated learning rate scheduler. Our method consistently outperforms the baselines across tasks like language modeling and image recognition.

- We are very concerned that a desk-rejection for a formatting error may overshadow the paper's contributions and would be deeply grateful if it could still be considered for review.
- Thank you very much for your understanding and for considering our request.

Review

Official Review by Program Committee AtZs 20 Sept 2024, 12:43 (modified: 10 Dec 2024, 04:08)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors

Revisions (/revisions?id=V6uiJ5qSVC)

Review:

I am sorry that I am not familiar with this research topic, so the following comments may be not correct:

1. first of all, this paper is not well prepared, as it clearly breaks the rule of submission, i.e., exceeding the page limits, which is not respectable manner to the reviewers;
2. the compared baselines seem to be too old-dated, would there be a comparison with more recent state-of-the-art methods?
3. while the empirical results are strong, a discussion of a more rigorous theoretical convergence guarantee for the proposed algorithms would be better.
4. The paper does not discuss the computational overhead of tracking sign changes and adjusting learning rates. It's unclear how this affects training time compared to simpler optimizers.

Rating: 5: Marginally below acceptance threshold

Confidence: 2: The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper

Rebuttal by Authors

Rebuttal

by Authors (Yashodhara V Haribhakta (/profile?id=~Yashodhara_V_Haribhakta1), Rhugved Chaudhari (/profile?id=~Rhugved_Chaudhari1))

09 Nov 2024, 17:11 (modified: 11 Nov 2024, 22:08)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors

Revisions (/revisions?id=9y9MyESWxf)

Rebuttal:

Thank you for your feedback. We appreciate your comments and suggestions, and we hope the following responses clarify our approach and address your concerns.

1. Page Limit

- We apologize for any confusion regarding the page limit. This was not an intentional oversight or deliberate rule break but rather a misunderstanding about the content allowed within the first seven pages. We have clarified this further in our response to Reviewer LjCb, and we kindly request that you refer to that explanation for additional context. We appreciate your understanding and will ensure the paper adheres to the guidelines in the final submission.

2. Regarding the Baselines used.

- Thank you for your feedback. While we agree that some baselines were proposed a few years ago, we respectfully believe they are not old-dated. Our selection process was driven by their widespread use and practical relevance today (Adam and SGD), and their historical importance in optimization research (Adagrad). Additionally, we would like to note that RAdam (2021) and AMSGrad (2019) are fairly recent methods that continue to serve as strong benchmarks in recent optimizer papers. We hope this clarifies our rationale for the selection of baselines.

3. Theoretical Convergence Proof

- Thank you for your feedback. We note that (Acc)Asro builds on Adam, with our main contribution being the sign-reversal learning rate adaptation within thresholds. Given Adam's theoretical convergence guarantee, we believe this foundation suffices, similar to learning rate schedulers that typically don’t require additional convergence proofs. We hope this addresses your concern.

4. Computational Overhead

- Thank you for highlighting the computational impact of our optimizer. We agree that an overhead analysis could clarify feasibility, though we omitted it as most optimizer papers do not include this, along with the following reasons.
- Element-wise Operations: The operations for tracking sign changes, updating scalers, and adjusting learning rates are element-wise with complexity O(n) (where n is the number of parameters), similar to other adaptive optimizers like Adam.
- Training Efficiency in Practice: Empirically, the additional operations have minimal impact on training time. Our results show improved convergence efficiency, requiring fewer iterations to achieve comparable or better performance than baseline optimizers, making training time favourable even with the extra operations.

The paper proposes a novel optimizer for faster and lower convergence over popular optimizers like Adam. The paper is well presented and claims are supported by good experiments. While there are flaws and room for improvement, the paper merits a rating above the acceptance threshold. I would like to see some of the cons claimed below in a finali revised work.

Official Review by Program Committee 3opC 20 Sept 2024, 06:34 (modified: 10 Dec 2024, 04:08)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors

Revisions (/revisions?id=QDHiHp0boz)

Review:

Pros:

- The benefits of this method to navigating saddle points and plateaus seems clear and is substantiated sufficiently by the experimental results
- The paper proposes a novel optimizer that leads to faster and lower convergence, demonstrating it's significance and potential impact
- The method is applicable to both Transformer and CNN based architectures, which is an important feature not commonly present in many optimization works
- The paper is well written, clear, and figures/tables support the arguments made well

Cons:

- While the evaluation covers important domains and architectures, it is limited. Optimizer works often need rigorous evaluations across an array of architecture and experimental configurations for full validation, and it would have been nice to cover a greater gammet of gamut of experiments
    - More so, error bars are not presented. It is not clear whether this performance would be maintained across random seeds
- While the paper does an okay job at validating the claims made in section 3.3.1, it is not clear or presented where the hyper-parameters for (Acc)Asro came from and what background efforts were made to arrive at them.

- While AccAsro was introduced to "accelerate convergence" and simply the hyper-parameter tuning process, it also consistently outperforms the regular Asro. While the introduction of Asro is useful, the paper should make it clear that AccAsro is the primary contribution and optimizer to use, which I found the wording disd not fully suggest.
- **Question**: how do we mitigate settle in a local optimum, avoiding the convergence to a (more) global optimum. While the method focuses on settling in local minima, it may unecessarily harm performance by missing lower convergence points. I am wondering if such a problem even may arise.

Summary

The authors present a novel optimize with proven performance and improvements over the popular Adam optimizer. While there are cons in the paper that I would like to see addressed in a final revision, the paper merits an acceptance rating.

Rating: 6: Marginally above acceptance threshold
Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

−
=
≡

Rebuttal by Authors

Rebuttal
by Authors (👁 Yashodhara V Haribhakta (/profile?id=~Yashodhara_V_Haribhakta1), Rhugved Chaudhari (/profile?id=~Rhugved_Chaudhari1))
📅 09 Nov 2024, 17:10 (modified: 11 Nov 2024, 22:08)
👁 Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors
🗐 Revisions (/revisions?id=PHmnWF9PnF)

Rebuttal:
Thank you for the positive assessment and constructive feedback; we will address concerns in a final revision.

1. Navigating Saddle Points and Plateaus
2. Novel Optimizer with Faster and Lower Convergence
3. Applicability to Transformer and CNN Architectures
4. Clarity and Presentation

- We sincerely appreciate your recognition of these strengths.

5. Hyperparameters

- We thank you for highlighting this point. We agree that a more detailed discussion of the hyperparameters for (Acc)Asro would enhance the paper, and this would be included in Sections 3.3.1 and 4 in a revision.
- We addressed this in the technical appendix (A.2.2, A.3.2). For increment factors, we evaluated 1e-3, 1e-2, 5e-2, 1e-1; finding that the vicinity of 1e-2 to 5e-2 works well without extensive tuning. For decrement factors, we tested 1e-5, 1e-4, 1e-3; observing that values around 1e-4 to 5e-5 provided similarly stable results.
- Initial learning rate (α), can be set based on the developer's preferences, with upper and lower LR scalers adjusted one magnitude higher and lower, respectively.

6. AccAsro: Primary Contribution

- We thank you for this insightful observation. We agree that AccAsro is the primary contribution of our work and represents the main optimizer to use, which we will specify in Section 3.3 in the revision.

7. Question

- Thank you for this excellent question. (Acc)Asro is not, only designed to settle at a local minimum. Sign-directed rate adaptation also uses the exponential moving average of gradients m_t as momentum to bypass shallow local minima. Learning rate deceleration only occurs upon sign reversal in m_t, typically in flatter, broader regions, promoting better generalization while avoiding sharp, suboptimal minima.

8. Empirical evaluation and Error bars

- Thank you for your valuable feedback. While we agree that more experiments could be included, we believe our empirical evaluation is “sufficient” to demonstrate our method's effectiveness. Please see our response to Reviewer “Pbni” (Point 4) for details.

- Thank you for your comment on including error bars. While we acknowledge they enhance statistical rigor, we did not include them as we averaged across 3 random seeds, which accounts for variability in the results.
- We observed consistently better performance of (Acc)Asro over baselines, even in multiple cases where seeds were not set at all, meaning the parameter initialization was random, demonstrating reliability despite random initializations.