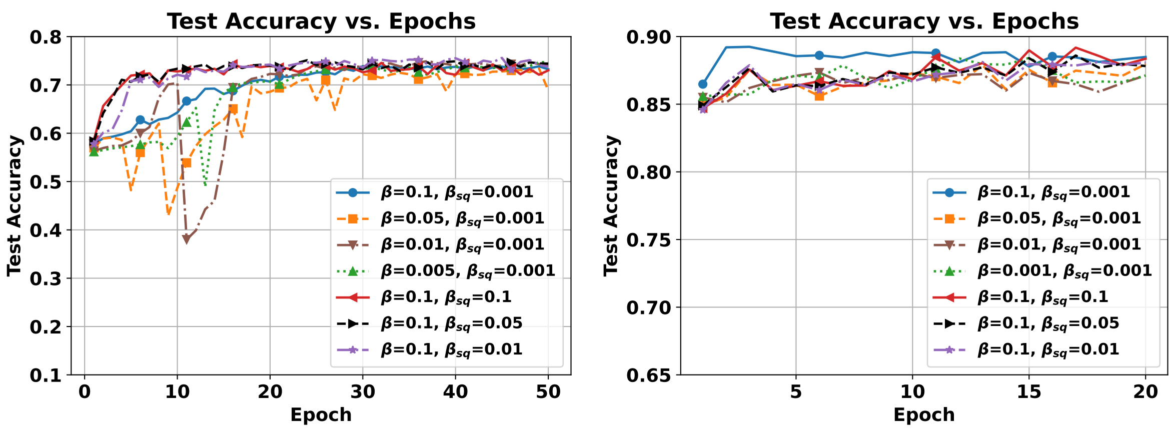
Rebuttal Response

Anonymous Authors[1]



(a) Transformer on AUC maximization

(b) BERT on hyper-representation

Figure 1. Test accuracy of different models on AUC maximization and hyper-representaion using AdamBO with different $(\beta, \beta_{\text{sq}})$. (a) 2-layer Transformer model on AUC maximization (data imbalanced ratio = 0.9); (b) 8-layer BERT model on hyper-representation.

Table 1. Comparison of Adam-related papers under different settings and assumptions. ✓̸ represents dropping the bias correction term for the first-order momentum while keeping it for the second-order momentum. d denotes the dimension. Only the key assumptions are listed here.

Adam Paper	Problem	Stochastic Setting	Assumptions	Choice of β	Bias Correction	Complexity
(De et al., 2018)	Single-Level	Deterministic	F.1(A) + F.2	$1 - O(\epsilon)$	✗	$O(\epsilon^{-6})$
(Défossez et al., 2020)	Single-Level	Stochastic (Expectation)	F.1(A) + F.2	$(\beta_{\text{sq}}, 1]$	✓̸	$\widetilde{O}(d\epsilon^{-4})$
(Guo et al., 2021)	Single-Level	Stochastic (Expectation)	F.1(A) + F.2 [1]	$O(\epsilon^2)$	✗	$O(\epsilon^{-4})$
(Zhang et al., 2022)	Single-Level	Stochastic (Finite Sum)	F.1(A)	$(1 - \sqrt{1-\beta_{\text{sq}}}, 1]$	✓ (Randomly Reshuffled)	Not Converge [2]
(Wang et al., 2022)	Single-Level	Stochastic (Finite Sum)	F.1(B)	$(1 - \sqrt{1-\beta_{\text{sq}}}, 1]$	✗ (Randomly Reshuffled)	Not Converge
(Li et al., 2023)	Single-Level	Stochastic (Expectation)	F.1(C)	$O(\epsilon^2)$	✓	$O(\epsilon^{-4})$
AdamBO (This work, Theorem 4.1)	Bilevel	Stochastic (Expectation)	F.1(B) [3]	$\widetilde{\Theta}(\epsilon^2)$	✓	$\widetilde{O}(\epsilon^{-4})$

References

De, S., Mukherjee, A., and Ullah, E. Convergence guarantees for rmsprop and adam in non-convex optimization and an empirical comparison to nesterov acceleration. *arXiv preprint arXiv:1807.06766*, 2018.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1](Guo et al., 2021, Assumption 2) can be implied by Assumption F.2, although it is weaker.

[2]Adam can converge with an additional strong growth condition (Zhang et al., 2022; Wang et al., 2022).

[3]Under Assumption 3.2, the objective function Φ is (L_0, L_1)-smooth, see Lemma B.10 for details.

Défossez, A., Bottou, L., Bach, F., and Usunier, N. A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

Guo, Z., Xu, Y., Yin, W., Jin, R., and Yang, T. A novel convergence analysis for algorithms of the adam family. *arXiv preprint arXiv:2112.03459*, 2021.

Li, H., Jadbabaie, A., and Rakhlin, A. Convergence of adam under relaxed assumptions. *arXiv preprint arXiv:2304.13972*, 2023.

Wang, B., Zhang, Y., Zhang, H., Meng, Q., Ma, Z.-M., Liu, T.-Y., and Chen, W. Provable adaptivity in adam. *arXiv preprint arXiv:2208.09900*, 2022.

Zhang, Y., Chen, C., Shi, N., Sun, R., and Luo, Z.-Q. Adam can converge without any modification on update rules. *Advances in neural information processing systems*, 35:28386–28399, 2022.