Adam Exploits ℓ_∞-geometry of Loss Landscape via Coordinate-wise Adaptivity

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2024

Abstract

Adam outperforms SGD in transformer optimization for language modeling tasks. Yet such benefits are not well-understood theoretically – previous theoretical convergence analysis for Adam and SGD mainly focus on the number of steps T and are already minimax-optimal in non-convex cases, which are both $O(T^{-1/4})$. In this work, we argue that the better dependency on the loss smoothness and model dimension is the key that Adam optimizes faster than SGD, which is typically much larger than total steps for modern language modeling tasks. More specifically, we give a new convergence analysis for Adam under novel assumptions that loss is smooth under ℓ_∞ geometry rather than the more common ℓ_2 geometry, which yields a much better empirical smoothness constant for GPT-2 models. Moreover, we show that if we rotate the pretraining loss randomly, Adam can be outperformed by some variants of SGD which is invariant to rotations. This implies that any practically relevant explanation of Adam's optimization benefit must involve non-rotational invariant properties of loss, such as ℓ_∞ smoothness as used in our analysis.

1. Introduction

Large language models (LLMs) have gained phenomenal capabilities as their scale grows [3, 10, 16, 18, 19, 21, 28]. However, pre-training LLMs are incredibly time-consuming. Adaptive Momentum Estimation (`Adam`)[11] is the current to-go optimization algorithm for LLMs due to its fast convergence. In contrast, `SGD`, the default algorithm for training more classic architectures like ResNets [8], optimizes language model loss much slower than `Adam`.

However, the optimization benefit of `Adam` over `SGD` cannot be explained by existing theory. Existing convergence analyses for `Adam` and `SGD` focus on the dependency on the number of steps under assumptions on the smoothness and gradient bounds of the loss [6], and it has been shown that both `Adam` and `SGD` achieve the minimax convergence rate $O(T^{-1/4})$ in the non-convex settings [1]. Thus according to the theory, in the worst case, `SGD` would be more desirable compared to `Adam` because they have the same convergence rate, and yet `Adam` is less memory-efficient due to its coordinate-wise adaptivity, which needs to store the empirical moving average of second-order moments of past stochastic gradients. Therefore, we hypothesize that the coordinate-wise adaptivity in `Adam` is exploiting some unknown properties of LLMs which SGD cannot make use of.

Towards this end, one significant difference between `Adam` and `SGD` we identified in this paper, which is often ignored in the assumptions of the previous works, is that `SGD` is rotation-invariant while `Adam` is only permutation-invariant (see definitions in Appendix C). Intuitively, this means if we rotate the loss landscape, the optimization trajectory of SGD would be the same (up to some rotation), while the trajectory of `Adam` could be completely different. If `Adam` optimizes much slower

after rotation, then it suggests Adam is exploiting some non-rotational-invariant properties of the loss function, which is not captured by standard theoretical assumptions in the convergence analysis.

Figure 1 summarizes our findings by comparing Adam on the original and rotated loss. The performance of Adam does become much worse than Adam on the original loss. We also test some memory-efficient and rotational-invariant variants of SGD, AdaSGD [22] (defined in Algorithm 2)[1]. Surprisingly, the rotated Adam performs even much worse than the SGD variant. The results suggest that it is impossible to separate the superior optimization performance of Adam over SGD just using rotationally invariant assumptions on the loss function, which raises the natural question,

> What are the non-rotation-invariant properties of a loss function that enable faster convergence of Adam than SGD?

We hypothesize the common assumption that the gradient of the loss function is Lipschitz w.r.t. ℓ_2 norm does not provide the best convergence rate of Adam. Inspired by the similarity between Adam and SignGD and the fact that SignGD is the normalized steepest descent with respect to ℓ_∞ norm, we propose ℓ_∞ norm as a better norm for Lipschitzness because each coordinate of the parameter update in a single step t in SignGD has the same magnitude, which is the learning rate η_t. Then we prove a convergence rate of $O(\sqrt{\frac{1}{T}})$ for Adam under this new assumption without noise, or $O((\frac{\log T}{T})^{1/4})$ with noise, which has the same dependency on T as previous results. However, our convergence rate will depend on the $(1,1)$-norm of Hessian instead of the top eigenvalue when assuming the gradient is ℓ_2 Lipschitz. In order to show that $(1,1)$-norm of Hessian as a new metric can affect how fast Adam converges in real tasks, we conduct experiments and find there is a correlation between convergence rate and this metric.

We summarize our contributions below

1. We show by experiments that only using rotation-invariant assumptions cannot explain the empirical optimization advantage of Adam over SGD for optimizing language models. (Figure 1)

2. We propose a new complexity metric for the optimization problem, which is the $(1,1)$-norm of the Hessian matrix of loss, $\left\|\nabla^2 L(x)\right\|_{1,1}$. We present a novel convergence result for Adam depending on this metric in the case of $\beta_1 = 0$. (Theorem 2.6)

3. We empirically verify that when Adam converges slower on the rotated loss on GPT-2 models, the $(1,1)$-norm of Hessian also increases, which suggests that our new complexity metric for Adam's convergence is practically relevant. (Table 1)

Notations and Settings. For a matrix $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$, its $(1,1)$-norm is defined as $\sum_{i=1}^{d_1} \sum_{j=1}^{d_2} |\boldsymbol{A}_{i,j}|$. We independent stochastic loss functions For a deterministic loss function $L(\boldsymbol{x})$, we consider optimization over L with only access independent stochastic functions $\{L_t(\boldsymbol{x})\}_{t=1}^T$ such that $\mathbb{E}L_t(\boldsymbol{x}) = L(\boldsymbol{x})$ for any input $\boldsymbol{x} \in \mathbb{R}^d$.

1. There is one small difference. We use an exponential average of the gradient for $\boldsymbol{m}_t$ instead of momentum. Our definition makes AdaSGD the same as Adam in a one-dimensional problem.

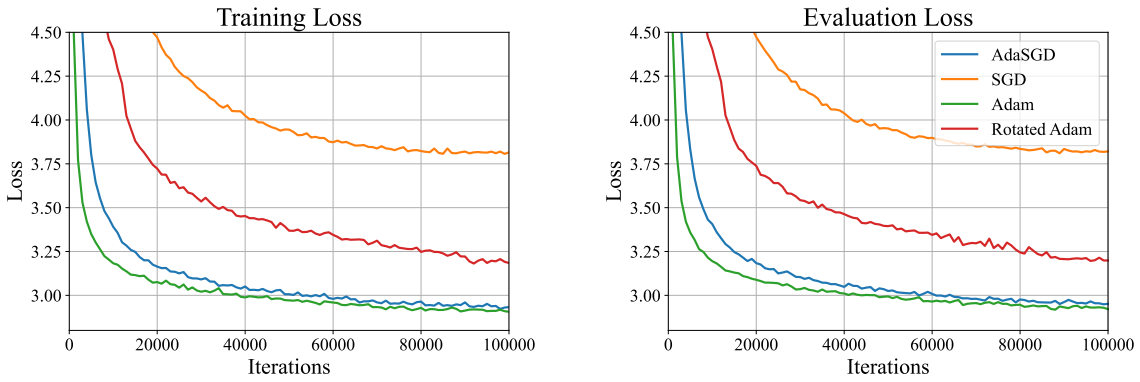

Figure 1: We plot the training losses of `Adam`, `AdaSGD` and `SGD`. `rotated Adam` means running `Adam` on a rotated loss. `Adam` on the original loss converges the fastest as expected. But `Adam` on a rotated loss convergence is much slower and is even worse than `AdaSGD`.

2. Main Results: Convergence Rates of `Adam`

In this section, we present our main theoretical results, which is the convergence rate of `Adam` for stochastic smooth loss with bounded gradient noise. We allow non-convex losses and thus the convergence is measured by the 1-norm of the loss gradient. For deterministic loss, the best convergence rate (Theorem 2.2) is achieved by `SignGD` (`Adam` with $\beta_1 = \beta_2 = 0$). For stochstic loss with bounded gradient noise, the best rate (Theorem 2.6) is achieved by `RMSProp` (`Adam` with $\beta_1 = 0$ and $\beta_2 \in [0, 1]$).

Similar to previous works [6], our analysis could be extended to the most general case of `Adam`, where both β_1,β_2 are non-zero, but the rate becomes strictly worse than the `RMSProp` (the case of $\beta_1 = 0$), as there will be some extra polynomials of $\frac{1}{1-\beta_1}$. We decide not to include result for the most general case, on one hand for ease of presentation, and on the other hand, because such result could explain the optimization benefit of momentum ($\beta_1 > 0$) in practice and does not add any insight on the benefit of `Adam`. We hypothesis that we are missing some important features of loss landscape of transformers in the theoretical assumptions and we leave this for future work.

2.1. Wamrup: `SignGD` ($\beta_1 = \beta_2 = 0$)

In this section, we use the convergence analysis for `SignGD` (`Adam` with $\beta_1 = \beta_2 = 0$) as a warm-up and illustrate how `Adam` could benefit from a non-rotational invariant property of the loss landscape, which in particular is the ℓ_∞ smoothness. The key observation here is that `SignGD` is the normalized steepest descent with respect to ℓ_∞ norm (see [24]), and thus it is more natural to analyze its convergence using ℓ_∞-norm-related geometry of the loss.

Definition 2.1 *Given a norm $\|\cdot\|$ on $\mathbb{R}^d$ and $\|\cdot\|_*$ as its dual norm, we say a function L is H-smooth w.r.t. $\|\cdot\|$ if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we have that $\|\nabla L(\boldsymbol{x}) - \nabla L(\boldsymbol{y})\|_* \le H \|\boldsymbol{x} - \boldsymbol{y}\|$.*

Theorem 2.2 *Let L be a H-smooth with respect to $\|\cdot\|_\infty$ and $\{\boldsymbol{x}_t\}_{t=1}^T$ be the iterates of `SignGD` (`Adam` with $\beta_1 = \beta_2 = 0$) on L with initialization $\boldsymbol{x}_0$ and learning rate η, it holds that*

$$\min_{1\le t\le T} \|\nabla L(\boldsymbol{x}_t)\|_1 \le \frac{L(\boldsymbol{x}_0) - \min L}{T\eta} + \frac{H\eta}{2}$$

if we choose $\eta = \sqrt{\frac{2(L(\boldsymbol{x}_0)-\min L)}{TH}}$, then $\min_{1\le t\le T} \|\nabla L(\boldsymbol{x}_t)\|_1 \le \sqrt{\frac{2H(L(\boldsymbol{x}_0)-\min L)}{T}}$.

2.2. Main Result: RMSProp ($\beta_1 = 0, \beta_2 \in [0, 1]$)

It is well-known that SignGD might not converge in the stochastic case as the expectation of descent direction for mini-batch loss may not be a descent direction, and RMSProp is proposed to address this issue by using a moving average of the squared gradient per coordinate to reduce the coorleation between the denominator and the numerator, thus making the expected update direction less biased [9]. In this subsection we formalize the above intuition and show indeed a positive β_2 in Adam helps convergencein in the stochastic case. The main challenges here are from the both lower bounding the first-order term and upper bounding the second-order term in the modified descent lemma (the counterpart of Equation 1 for RMSProp). To circumvent these difficulties, we introduce a slightly stronger assumption than Definition 2.1, which is Definition 2.3. By definition, $\mathbf{H}$-smooth coordinate-wisely w.r.t. ℓ_∞ norm implies $\sum_{i=1}^d H_i$ smooth w.r.t. ℓ_∞ norm. Due to the limitation of space, we only present the main result here. The sketch of the proof is presented in Appendix E.1 while the full proof is in Appendix E with technical lemmas in Appendix D.

Definition 2.3 *For any $\mathbf{H} = (H_1, \dots, H_d) \in \mathbb{R}^d$, we say a function L is* $\mathbf{H}$-smooth coordinate-wisely w.r.t. ℓ_∞ norm , *iff for any $i \in [d]$, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $|\nabla_i L(\boldsymbol{x}) - \nabla_i L(\boldsymbol{y})| \le H_i \|\boldsymbol{x} - \boldsymbol{y}\|_\infty$.*

Assumption 2.4 (Bounded coordinate-wise noise) *There exist constants σ_i such that $|\nabla_i L_t(\boldsymbol{x}) - \nabla_i L(\boldsymbol{x})| \le \sigma_i$ for any $i \in [d]$, $t \in \mathbb{N}$ and $\boldsymbol{x} \in \mathbb{R}^d$.*

Assumption 2.5 (Bounded coordinate-wise stochastic gradient) *For any $i \in [d]$, $t \in \mathbb{N}$ and $\boldsymbol{x} \in \mathbb{R}^d$, $|\nabla_i L_t(\boldsymbol{x})| \le G$.*

Theorem 2.6 (Main) *Let $\{L_t\}_{t=1}^T$ be independent stochastic losses satisfying Assumptions 2.4 and 2.5 and that their expectation L is $\boldsymbol{H}$-coordinate-wisely smooth w.r.t. ℓ_∞ norm. For $\beta_1 = 0$, we have that*

$$\min_{1\le t\le T} \mathbb{E}\,\|\nabla L(\boldsymbol{x}_t)\|_1 \le E + \sqrt{E \cdot \sum_{i=1}^d \sigma_i} + \sqrt{dE}\epsilon^{1/4}$$

with $C = \ln \frac{G^2 + \max_i \sigma_i^2 + \epsilon}{\epsilon}$ *and* $E = O\left(\frac{1}{\sqrt{T}}\sqrt{\sqrt{C}\,(L(\boldsymbol{x}_0) - \min L)\sum_{i=1}^d H_i} + \sqrt{\frac{(\log T + C)}{T}}\sum_{i=1}^d \sigma_i\right)$

if we choose $1 - \beta_2 = \Omega(\frac{\log T + C}{T})$ *and* $\eta = \Theta\left(\sqrt{\frac{L(\boldsymbol{x}_0) - \min L}{\sqrt{C}\sum_{i=1}^d H_i T}}\right)$

We will interpret the results in the deterministic and stochastic settings respectively and we will see the match the standard rates but with the usual ℓ_2 smoothness replaced by ℓ_∞ related smoothness, which is much smaller for language models such as GPT-2 empirically Table 1. In the deterministic setting, $\sigma_i = 0$ and thus the RHS becomes $O(E) = O(\sqrt{\frac{\sum_{i=1}^d H_i (L(\boldsymbol{x}_0) - \min_L)}{T}})$, which is much faster than the $O(T^{-\frac{1}{4}})$ rate by Défossez et al. [6]. In the stochastic case, the RHS becomes

$$O\left(\sqrt{E \cdot \sum_{i=1}^n \sigma_i}\right) = O\left(\left(\frac{\log T}{T}\right)^{1/4} \sum_{i=1}^d \sigma_i) + O((\frac{(\sum_{i=1}^d H_i (L(\boldsymbol{x}_0) - \min_L))^{1/4} (\sum_{i=1}^d \sigma_i)^{1/2}}{T^{1/4}}\right).$$

While many previous works rely on the relatively large magnitude of ϵ compared to $\boldsymbol{v}_t$ and give a bound in the regime of SGD when the adaptive effect is dominated by the constant ϵ [5, 25], our result has a much milder dependency on ϵ, *i.e.*, $\log \epsilon$ and thus we could ignore the last addtive term by picking ϵ to be inverse of some high-degree polynomial of T, or machine precision. Like previous works[6, 7], we also assume the stochastic gradients are almost surely bounded. However,our convergence rates depends on $\log G/\epsilon$ instead of G itself. In fact, we can even get $C = \log \frac{T \max_i H_i \eta}{\epsilon}$ without any assumption on the gradient magnitude, where η is learning rate.

(1,1)-norm as a surrogate complexity measure. H_i is determined by $\sup_{\boldsymbol{x}} \sum_{j=1}^{d} \left|\nabla^2_{i,j} L(\boldsymbol{x})\right|$, which is difficult to compute because it requires taking supreme over the entire domain. Instead, we approximate $\sum_{i=1}^{d} H_i$ locally by the $(1,1)$-norm of Hessian of loss along the training trajectory, which can also efficiently be approximated by using hessian-vector product against random Cauchy vectors. Definition 2.3 is not rotation-invariant in the sense that the $(1,1)$-norm of Hessian matrix can vary a lot when a rotation is performed on the loss. In comparison, previous works often assume $\|\nabla L(\boldsymbol{x}) - \nabla L(\boldsymbol{y})\|_2 \le H \|\boldsymbol{x} - \boldsymbol{y}\|_2$ and H can be interpreted as the top singular value of Hessian matrix which won't be changed after rotation.

3. Experiments

In order to empirically investigate and confirm the implications of our propsosed theory, we compare the training performance of Adam with AdaSGD, SGD and rotated Adam on a language modeling task with a transformer-based architecture. The details can be found in Appendix F.

3.1. Analysis of results

Since we propose the $(1,1)$-norm of Hessian as a non rotation-invariant metric that can affect the convergence rate of Adam, we also measure it for original loss function L and rotated loss function $\tilde{L}$ on checkpoints trained with different losses. The results are presented in Table 1.

For the same checkpoint, no matter if the Adam training is done on the original loss or rotated loss, the $(1,1)$-norm of rotated loss $\tilde{L}$ is always larger than that of original loss L and comparable to larger than d times the top singular value. It suggests that the loss exhibits good ℓ_∞ geometry property in the original space, which no longer holds after random permutation. Moreover, comparing the $(1,1)$-norm of Hessian of the rotated loss evaluated at the checkpoint trained with rotated loss, the $(1,1)$-norm of the original loss evaluated at the checkpoint trained with the original loss is much smaller (more than 20 times, 57.48-¿2.70). This together with our Theorem 2.6 explains why in Figure 1, rotated Adam performs worse than Adam.

	Trained with L		Trained with rotated loss $\tilde{L}$	
Measured Loss	L	$\tilde{L}$	L	$\tilde{L}$
$\|\cdot\|_{1,1}/d$	2.70	121.05	5.41	57.48
$\|\cdot\|_2$	65.02		63	

Table 1: $(1,1)$-norm and ℓ_2 norm of different losses on different checkpoints.

References

[1] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023.

[2] Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, 2018. URL https://arxiv.org/pdf/1705.07774.pdf.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/pdf/1808.02941.pdf.

[5] Soham De, Anirbit Mukherjee, and Enayat Ullah. Convergence guarantees for rmsprop and adam in non-convex optimization and an empirical comparison to nesterov acceleration. *arXiv preprint arXiv:1807.06766*, 2018.

[6] Alexandre Défossez, Leon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *Transactions on Machine Learning Research*, 2022. URL https://arxiv.org/pdf/2003.02395.pdf.

[7] Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. A novel convergence analysis for algorithms of the adam family. *arXiv preprint arXiv:2112.03459*, 2021. URL https://arxiv.org/pdf/2112.03459.pdf.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. URL https://arxiv.org/pdf/1512.03385.pdf.

[9] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012.

[10] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. URL https://arxiv.org/pdf/1412.6980.pdf.

[12] Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between sgd and adam on transformers, but sign descent might be. In *The Eleventh International Conference on Learning Representations*, 2022. URL https://arxiv.org/pdf/2304.13960.pdf.

[13] Frederik Kunstner, Robin Yadav, Alan Milligan, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why adam outperforms gradient descent on language models. *arXiv preprint arXiv:2402.19449*, 2024. URL https://arxiv.org/pdf/2402.19449.

[14] Kfir Levy, Ali Kavis, and Volkan Cevher. Storm+: Fully adaptive sgd with recursive momentum for nonconvex optimization. *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper/2021/file/ac10ff1941c540cd87c107330996f4f6-Paper.pdf.

[15] Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a3cc50126338b175e56bb3cad134db0b-Paper-Conference.pdf.

[16] OpenAI. Gpt-4 technical report. *arXiv*, 2023.

[17] Yan Pan and Yuanzhi Li. Toward understanding why adam converges faster than sgd for transformers. *arXiv preprint arXiv:2306.00204*, 2023. URL https://arxiv.org/pdf/2306.00204.

[18] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[19] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[20] Naichen Shi and Dawei Li. Rmsprop converges with proper hyperparameter. In *International conference on learning representation*, 2021. URL https://openreview.net/pdf?id=3UDSdyIcBDA.

[21] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[22] Jiaxuan Wang and Jenna Wiens. Adasgd: Bridging the gap between sgd and adam. *arXiv preprint arXiv:2006.16541*, 2020. URL https://arxiv.org/pdf/2006.16541.

[23] Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 2020.

[24] Shuo Xie and Zhiyuan Li. Implicit bias of adamw: ℓ_∞ norm constrained optimization. *arXiv preprint arXiv:2404.04454*, 2024. URL https://arxiv.org/pdf/2404.04454.

[25] Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. *Advances in neural information processing systems*, 2018. URL https://proceedings.neurips.cc/paper/2018/file/90365351ccc7437a1309dc64e4db32a3-Paper.pdf.

[26] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

[27] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/pdf/1912.03194.pdf.

[28] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[29] Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/pdf/2208.09632.pdf.

[30] Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need adam: A hessian perspective. *arXiv preprint arXiv:2402.16788*, 2024. URL https://arxiv.org/pdf/2402.16788v1.

[31] Dongruo Zhou, Jinghui Chen, Yuan Cao, Yiqi Tang, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018. URL https://arxiv.org/pdf/1808.05671.

[32] Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2019. URL https://arxiv.org/pdf/1811.09358.pdf.

Appendix A. Related Works

Comparison between `Adam` and `SGD` Previous work tries to analyze the difference between `Adam` and `SGD` from different perspectives. Zhou et al. [31] proves a faster convergence rate of `Adam` than `SGD` when the stochastic gradients are sparse. Zhang et al. [27] suggests that `SGD` suffers more from heavy-tailed noise than `Adam`. Pan and Li [17] claims that `Adam` has lower directional sharpness because of the effect of coordinate-wise clipping. Other works also consider the coordinate-wise normalization of `Adam` [2, 12]. Kunstner et al. [13] shows that the heavy-tailed class imbalance in language modeling tasks will cause `SGD` to converge slower when it can only optimize majority class well. Zhang et al. [30] finds that `Adam` is better at handling the block heterogeneity of Hessian matrix, which is a specific phenomenon in transformers. When viewing `Adam` as an adaptive method, there are works showing that adaptive methods have an advantage of achieving optimal convergence rate without relying on problem-dependent constant [14, 23].

Convergence rate of Adam There are many works showing convergence rate for `Adam` Chen et al. [4], Défossez et al. [6], Guo et al. [7], Shi and Li [20], Zhang et al. [29], Zhou et al. [31], Zou et al. [32]. Most of them rely on the smoothness of the loss function, which is measured w.r.t. ℓ_2 norm. Zhang et al. [26] proposes the (L_0, L_1) smoothness condition should be more reasonable than globally bounded smoothness. Li et al. [15] further generalizes the (L_0, L_1) smoothness condition. However, they still focus on the default ℓ_2 norm which is rotation-invariant. To the best of our knowledge, we are the first to assume gradient Lipschitzness under ℓ_∞ norm.

Appendix B. Convergence rate of `SignGD` for deterministic loss

Proof [Proof of Theorem 2.2] We will directly prove a more general verion of Theorem 2.2. Because L is H-smooth with respect to $\|\cdot\|_\infty$, we have that

$$\begin{aligned} L(\boldsymbol{x}_{t+1}) - L(\boldsymbol{x}_t) &\leq -\nabla L(\boldsymbol{x}_t)^\top (\boldsymbol{x}_t - \boldsymbol{x}_{t+1}) + \frac{H}{2}\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2 \\ &\leq -\eta \|\nabla L(\boldsymbol{x}_t)\|_* + \frac{\eta^2 H}{2}\eta^2 \end{aligned} \tag{1}$$

This implies that

$$\min_{1\leq t\leq T} \|\nabla L(\boldsymbol{x}_t)\|_* \leq \frac{1}{T}\sum_{t=1}^{T} \|\nabla L(\boldsymbol{x}_t)\|_* \leq \frac{L(\boldsymbol{x}_0) - L(\boldsymbol{x}_T)}{T\eta} + \frac{H\eta}{2},$$

which completes the proof. ■

Appendix C. Invariance property of Adam and SGD

In this section, we will show the different property between `Adam` and `SGD` in the following Theorem C.2.

Algorithm 1 `Adam`

Hyperparam: $\beta_1, \beta_2, \epsilon \geq 0$, total steps T, learning rate schedule $\{\eta_t\}_{t=1}^T$, ϵ
Input: initialization $\boldsymbol{x}_0$, stochastic loss functions $\{L_t\}_{t=1}^T$
$\boldsymbol{m}_0 \leftarrow \boldsymbol{g}_1, \boldsymbol{v}_0 \leftarrow \boldsymbol{g}_1^2$
for $t = 1, 2, \cdots, T$:
$\boldsymbol{g}_t \leftarrow \nabla L_t(\boldsymbol{x}_{t-1})$
$\boldsymbol{m}_t \leftarrow \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t$
$\boldsymbol{v}_t \leftarrow \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)\boldsymbol{g}_t^2$
$\boldsymbol{x}_t \leftarrow \boldsymbol{x}_{t-1} - \eta_t \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t + \epsilon}}$
return $\boldsymbol{x}_T$

Algorithm 2 `AdaSGD`

Hyperparam: $\beta_1, \beta_2 > 0$, total steps T, learning rate schedule $\{\eta_t\}_{t=1}^T$
Input: initialization $\boldsymbol{x}_0$, stochastic loss functions $\{L_t\}_{t=1}^T$
$\boldsymbol{m}_0 \leftarrow \boldsymbol{g}_1, v_0 \leftarrow \|\boldsymbol{g}_1\|_2^2 / d$
for $t = 1, 2, \cdots, T$:
$\boldsymbol{g}_t \leftarrow \nabla L_t(\boldsymbol{x}_{t-1})$
$\boldsymbol{m}_t \leftarrow \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t$
$v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2)(\|\boldsymbol{g}_t\|_2^2 / d)$
$\boldsymbol{x}_t \leftarrow \boldsymbol{x}_{t-1} - \frac{\eta_t}{\sqrt{v_t}} \boldsymbol{m}_t$
return $\boldsymbol{x}_T$

Rotation. For an invertible function $\mathcal{T} : \mathbb{R}^d \to \mathbb{R}^d$, $\mathcal{T}$ is a rotating transformation if there exists an orthogonal matrix $\boldsymbol{T} \in \mathbb{R}^{d \times d}$ such that $\mathcal{T}(\boldsymbol{x}) = \boldsymbol{T}\boldsymbol{x}$. $\mathcal{T}$ is a permutating transformation if there exists a permutation $\pi : [d] \to [d]$ such that $\mathcal{T}(\boldsymbol{x}) = [x_{\pi(1)}, \ldots, x_{\pi(d)}]^\top$. A permutating transformation is always a rotating transformation. We will use $\mathcal{R}$ to denote a rotating transformation.

Definition C.1 *For initialization $\boldsymbol{x}_0$ and stochastic losses $\{L_t\}_{t=1}^T$, we can get $\boldsymbol{x}_t$ when running algorithm A on $(\boldsymbol{x}_0, \{L_t\}_{t=1}^T)$. For a transformation $\mathcal{T}$, we can also get $\tilde{\boldsymbol{x}}_t$ when running A with the same hyperparameters on $(\tilde{\boldsymbol{x}}_0, \{\tilde{L}_t\}_{t=1}^T)$ with $\tilde{\boldsymbol{x}}_0 = \mathcal{T}^{-1}(\boldsymbol{x}_0)$ and $\tilde{L}_t = L_t \circ \mathcal{T}$.*

An algorithm A is invariant w.r.t. $\mathcal{T}$ if it always holds that $\tilde{\boldsymbol{x}}_t = \mathcal{T}^{-1}(\boldsymbol{x}_t)$ for any hyperparameters, initialization and stochastic losses. An algorithm A is rotation invariant if it is invariant w.r.t. any rotating transformation $\mathcal{R}$. And A is permutation invariant if it is invariant w.r.t. any permutating transformation.

Theorem C.2 `SGD` *and* `AdaSGD` *are rotation-invariant.* `Adam` *and* `SignGD` *are permutation-invariant.*

Proof For `SGD` and `AdaSGD`, we will show they are rotation-invariant by induction. For any rotating transformation $\mathcal{R}(\boldsymbol{x}) = \boldsymbol{R}\boldsymbol{x}$, suppose $\tilde{\boldsymbol{x}}_s = \mathcal{R}^{-1}(\boldsymbol{x}_s) = \boldsymbol{R}^\top \boldsymbol{x}_s$ holds for $s \leq t-1$. Then we have that $\tilde{\boldsymbol{g}}_t = \nabla_{\tilde{\boldsymbol{x}}} \tilde{L}_t(\tilde{\boldsymbol{x}}_t) = \boldsymbol{R}^\top \nabla_{\boldsymbol{x}} L(\boldsymbol{R}^{-1}\tilde{\boldsymbol{x}}_{t-1}) = \boldsymbol{R}^\top \nabla_{\boldsymbol{x}} L(\boldsymbol{x}_{t-1}) = \boldsymbol{R}^\top \boldsymbol{g}_t$ and $\tilde{\boldsymbol{m}}_t = \boldsymbol{R}^\top \boldsymbol{m}_t$. From the update rule of `SGD`, we have that $\tilde{\boldsymbol{x}}_t = \tilde{\boldsymbol{x}}_{t-1} - \eta_t \tilde{\boldsymbol{m}}_t = \boldsymbol{R}^\top \boldsymbol{x}_{t-1} - \eta_t \boldsymbol{R}^\top \boldsymbol{m}_t = \boldsymbol{R}^\top (\boldsymbol{x}_{t-1} - \eta_t \boldsymbol{m}_t) = \boldsymbol{R}^\top \boldsymbol{x}_t$. For the update rule of `AdaSGD`, we further have that $\|\tilde{\boldsymbol{g}}_t\|_2^2 = \|\boldsymbol{g}_t\|_2^2$ because $\boldsymbol{R}$ is an orthogonal matrix. Then $\tilde{v}_t = v_t$ and the derivation is similar.

For `Adam` and `SignGD`, it is easy to show by induction they are invariant w.r.t. any permutating transformation because the operation on gradient is performed on each coordinate separately. We only need to show they are not invariant w.r.t. a rotating transformation. We choose $\boldsymbol{R} = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}; \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}]$, $L_t(\boldsymbol{x}) = L(\boldsymbol{x}) = 2x_1^2 + x_2^2$. Due to the update rule of `SignGD`, it can only update $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ in the direction of $[1, 1]$ and $[1, -1]$. But when rotating the update direction on $\tilde{\boldsymbol{x}}$ back to the space of $\boldsymbol{x}$. The update direction can only be $[1, 0]$ or $[0, 1]$ that are different from the update direction in the original space. Because the first step in `Adam` takes the same direction in `SignGD`, we simultaneously show that both `SignGD` and `Adam` are not rotation-invariant. ■

Appendix D. Technical Lemmas

Lemma D.1 *Given any $\beta_1 \le \beta_2 < 1$, suppose scalar sequences $\{v_t\}_{t=0}^{\infty}$ and $\{g_t\}_{t=1}^{\infty}$ satisfy that $v_0 \ge 0, v_1 > 0$ and $v_t - \beta_2 v_{t-1} \ge (1-\beta_2) g_t^2$ for $t \ge 1$. Given initial value $m_0 \le v_0$, define $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t^2$ for $t \ge 1$. For any coefficients $\{\eta_t\}_{t=1}^{\infty}$ and $0 \le T_1 < T_2$, it always holds that*

$$\sum_{t=T_1+1}^{T_2} \eta_t \frac{m_t}{v_t} \le \sum_{t=T_1+1}^{T_2} \eta_t + \frac{\beta_2 - \beta_1}{(1-\beta_2)} \sum_{t=0}^{T_2} \alpha_t \ln v_t$$

with

$$\alpha_t = \begin{cases} \eta_t - \sum_{i=1}^{T_2-t} (1-\beta_1)\beta_1^{i-1}\eta_{t+i} & T_1 + 1 \le t \le T_2 \\ -\sum_{i=T_1+1}^{T_2} (1-\beta_1)\beta_1^{i-t-1}\eta_i & 1 \le t \le T_1 \\ -\sum_{i=T_1+1}^{T_2} \eta_i \beta_1^{i-1} & t = 0 \end{cases}$$

Specifically, when all the η_t are the same η and there exists constant C such that $\ln \frac{v_t}{v_s} \le C_0$ for any $s < t$, we can have

$$\sum_{t=T_1+1}^{T_2} \eta \frac{m_t}{v_t} \le \eta(T_2 - T_1) + \frac{(\beta_2 - \beta_1)\eta C_0}{(1-\beta_2)(1-\beta_1)}.$$

Proof [Proof of Lemma D.1] Notice that $1-x \le \ln \frac{1}{x}$ for any positive x. We can have that

$$
\begin{aligned}
&\sum_{t=T_1+1}^{T_2} \eta_t \frac{m_t}{v_t} \\
&= \sum_{t=T_1+1}^{T_2} \eta_t \left(\frac{\beta_1^t m_0}{v_t} + \sum_{i=0}^{t-1} (1-\beta_1)\beta_1^i \frac{g_{t-i}^2}{v_t} \right) \\
&\le \sum_{t=T_1+1}^{T_2} \eta_t \frac{\beta_1^t v_0}{v_t} + \sum_{t=T_1+1}^{T_2} \eta_t \sum_{i=0}^{t-1} (1-\beta_1)\beta_1^i \frac{v_{t-i} - \beta_2 v_{t-i-1}}{(1-\beta_2)v_t} \\
&= \sum_{t=T_1+1}^{T_2} \eta_t \frac{\beta_1^t v_0}{v_t} + \sum_{t=T_1+1}^{T_2} \eta_t \frac{1-\beta_1}{(1-\beta_2)v_t} \left(v_t - \beta_1^{t-1}\beta_2 v_0 + \sum_{i=1}^{t-1} (\beta_1^i - \beta_2\beta_1^{i-1}) v_{t-i} \right) \\
&= \sum_{t=T_1+1}^{T_2} \eta_t + \frac{\beta_2-\beta_1}{1-\beta_2} \sum_{t=T_1+1}^{T_2} \eta_t \left(1 - \beta_1^{t-1}\frac{v_0}{v_t} - (1-\beta_1)\sum_{i=1}^{t-1} \beta_1^{i-1}\frac{v_{t-i}}{v_t} \right) \\
&\le \sum_{t=T_1+1}^{T_2} \eta_t + \frac{\beta_2-\beta_1}{1-\beta_2} \sum_{t=T_1+1}^{T_2} \eta_t \left(\beta_1^{t-1}\left(1 - \frac{v_0}{v_t}\right) + (1-\beta_1)\sum_{i=1}^{t-1} \beta_1^{i-1}\left(1 - \frac{v_{t-i}}{v_t}\right) \right) \\
&\le \sum_{t=T_1+1}^{T_2} \eta_t + \frac{\beta_2-\beta_1}{1-\beta_2} \sum_{t=T_1+1}^{T_2} \eta_t \left(\beta_1^{t-1} \ln\left(\frac{v_t}{v_0}\right) + (1-\beta_1)\sum_{i=1}^{t-1} \beta_1^{i-1} \ln\left(\frac{v_t}{v_{t-i}}\right) \right) \\
&= \sum_{t=T_1+1}^{T_2} \eta_t + \frac{\beta_2-\beta_1}{1-\beta_2} \sum_{t=T_1+1}^{T_2} \ln v_t \left(\eta_t - \sum_{i=1}^{T_2-t} (1-\beta_1)\beta_1^{i-1}\eta_{t+i} \right) \\
&- \frac{\beta_2-\beta_1}{1-\beta_2} \sum_{t=1}^{T_1} \ln v_t \left(\sum_{i=T_1+1}^{T_2} (1-\beta_1)\beta_1^{i-t-1}\eta_i \right) - \frac{\beta_2-\beta_1}{1-\beta_2} \ln v_0 \left(\sum_{i=T_1+1}^{T_2} \eta_i \beta_1^{i-1} \right) \\
&= \sum_{t=T_1+1}^{T_2} \eta_t + \frac{\beta_2-\beta_1}{1-\beta_2} \sum_{t=0}^{T_2} \alpha_t \ln v_t, \qquad (2)
\end{aligned}
$$

which finishes the proof for general η_t.

When all the $\eta_t = \eta$, we can determine the sign of coefficient α_t for each $\ln v_t$. When $T_1+1 \le t \le T_2$, $\alpha_t = \eta\beta_1^{T_2-t}$ and is positive. When $1 \le t \le T_1$, $\alpha_t = \eta(\beta_1^{T_2-t} - \beta_1^{T_1-t})$ and is negative. For $t=0$, $\alpha_0 = \eta\frac{\beta_1^{T_2}-\beta_1^{T_1}}{1-\beta_1}$ and is negative. It is easy to verify that $\sum_{t=0}^{T_2}\alpha_t = 0$. For any $0 \le t_1 \le T_1$ and $T_1+1 \le t_2 \le T_2$, we can find non-negative α_{t_1,t_2} such that $\sum_{t=T_1+1}^{T_2}\alpha_{t_1,t} = -\alpha_{t_1}$ holds for any $0 \le t_1 \le T_1$ and $\sum_{t=0}^{T_1}\alpha_{t,t_2} = \alpha_{t_2}$ holds for any $T_1+1 \le t_2 \le T_2$. Then we can rewrite

$\sum_{t=0}^{T_2} \alpha_t \ln v_t$ as

$$\begin{aligned}
\sum_{t=0}^{T_2} \alpha_t \ln v_t &= \sum_{t_1=0}^{T_1} \sum_{t_2=T_1+1}^{T_2} (-\alpha_{t_1,t_2}) \ln v_{t_1} + \sum_{t_2=T_1+1}^{T_2} \sum_{t_1=0}^{T_1} \alpha_{t_1,t_2} \ln v_{t_2} \\
&= \sum_{t_1=0}^{T_1} \sum_{t_2=T_1+1}^{T_2} \alpha_{t_1,t_2} \ln \left(\frac{v_{t_2}}{v_{t_1}} \right) \\
&\leq \sum_{t_1=0}^{T_1} \sum_{t_2=T_1+1}^{T_2} \alpha_{t_1,t_2} C \\
&= C \sum_{t_2=T_1+1}^{T_2} \alpha_{t_2} = C \sum_{t_2=T_1+1}^{T_2} \eta \beta_1^{T_2-t_2} = \eta C \frac{1-\beta_1^{T_2-T_1}}{1-\beta_1}.
\end{aligned}$$

We can plug it into Equation 2 and get that

$$\begin{aligned}
\sum_{t=T_1+1}^{T_2} \eta_t \frac{m_t}{v_t} &\leq (T_2 - T_1)\eta + \frac{\beta_2 - \beta_1}{1-\beta_2} \eta C \frac{1-\beta_1^{T_2-T_1}}{1-\beta_1} \\
&= \eta(T_2 - T_1) + \frac{(\beta_2 - \beta_1)\eta C_0}{(1-\beta_2)(1-\beta_1)}.
\end{aligned}$$

■

Starting from here till the end of this section, $\{g_t\}_{t=1}^T$ is any scalar sequence and $\{v_t\}_{t=0}^T$ is defined by $v_0 = g_1^2$ and $v_t = (1-\beta_2)g_t^2 + \beta_2 v_{t-1}$ for $t \geq 1$.

Lemma D.2 *Suppose there exists C_1, C_2 such that $|g_t - g_s| \leq C_1(t-s)\sqrt{1 + \frac{\beta_2 C_2}{(1-\beta_2)(t-s)}}$ for any $1 \leq s < t$. Then it holds that*

$$\left| \sqrt{v_t + \epsilon} - \sqrt{v_{t-1} + \epsilon} \right| \leq \left| \sqrt{v_t} - \sqrt{v_{t-1}} \right| \leq \frac{12(1+C_2)}{\sqrt{C_2}} C_1 + (1 - \sqrt{\beta_2}) \beta_2^{\frac{t-1}{2}} |g_1|.$$

Proof When $|g_t| \leq \sqrt{v_{t-1}}$, we have $\sqrt{v_t} - \sqrt{v_{t-1}} \leq 0$. Then we have

$$\begin{aligned}
&\sqrt{v_{t-1}} - \sqrt{v_t} \\
=&\frac{v_{t-1} - v_t}{\sqrt{v_{t-1}} + \sqrt{v_t}} \\
=&\frac{(1-\beta_2)(v_{t-1} - g_t^2)}{\sqrt{v_{t-1}} + \sqrt{v_t}} \\
\leq&(1-\beta_2)\frac{\beta_2^{t-1}(v_0 - g_t^2) + (1-\beta_2)\sum_{i=0}^{t-2} \beta_2^i (g_{t-1-i}^2 - g_t^2)}{\sqrt{v_{t-1}} + \sqrt{v_t}} \\
\leq&\frac{(1-\beta_2)^2 \sum_{i=0}^{t-2} \beta_2^i |g_{t-1-i}| \left(|g_{t-1-i}| - |g_t| \right)}{\sqrt{v_{t-1}} + \sqrt{v_t}} + \frac{(1-\beta_2)^2 \sum_{i=0}^{t-2} \beta_2^i |g_t| \left(|g_{t-1-i}| - |g_t| \right)}{\sqrt{v_{t-1}} + \sqrt{v_t}} + \frac{(1-\beta_2)\beta_2^{t-1} v_0}{\sqrt{v_{t-1}} + \sqrt{v_t}}.
\end{aligned}$$

We deal with the first term

$$\begin{aligned}
(1-\beta_2)^2\sum_{i=0}^{t-2}\beta_2^i\left|g_{t-1-i}\right|\left(\left|g_{t-1-i}\right|-\left|g_t\right|\right) &\le (1-\beta_2)^2\left(\sum_{i=0}^{t-2}\beta_2^i g_{t-1-i}^2\right)^{\frac{1}{2}}\left(\sum_{i=0}^{t-2}\beta_2^i(\left|g_{t-1-i}\right|-\left|g_t\right|)^2\right)^{\frac{1}{2}}\\
&\le (1-\beta_2)\sqrt{v_{t-1}}\left((1-\beta_2)\sum_{i=0}^{t-2}\beta_2^i\left|g_{t-1-i}-g_t\right|^2\right)^{\frac{1}{2}}\\
&\le C_1(1-\beta_2)\sqrt{v_{t-1}}\left((1-\beta_2)\sum_{i=1}^{t-1}\beta_2^i\left(i^2+\frac{\beta_2 C_2 i}{1-\beta_2}\right)\right)^{\frac{1}{2}}\\
&\le C_1(1-\beta_2)\sqrt{v_{t-1}}\left((1-\beta_2)\sum_{i=1}^{\infty}\beta_2^i\left(i^2+\frac{\beta_2 C_2 i}{1-\beta_2}\right)\right)^{\frac{1}{2}}\\
&= C_1(1-\beta_2)\sqrt{v_{t-1}}\frac{\sqrt{\beta_2+(1+C_2)\beta_2^2}}{1-\beta_2}\\
&= C_1\sqrt{v_{t-1}}\sqrt{\beta_2+(1+C_2)\beta_2^2}.
\end{aligned}$$

Then we deal with the second term

$$\begin{aligned}
(1-\beta_2)^2\sum_{i=0}^{t-2}\beta_2^i\left|g_t\right|\left(\left|g_{t-1-i}\right|-\left|g_t\right|\right) &\le (1-\beta_2)^2\sum_{i=0}^{t-2}\beta_2^i\sqrt{v_{t-1}}(\left|g_{t-1-i}\right|-\left|g_t\right|)\\
&\le (1-\beta_2)\sqrt{v_{t-1}}\left((1-\beta_2)\sum_{i=0}^{t-2}\beta_2^i\left|g_{t-1-i}-g_t\right|^2\right)^{\frac{1}{2}}\\
&\le C_1\sqrt{v_{t-1}}\sqrt{\beta_2+(1+C_2)\beta_2^2}.
\end{aligned}$$

For the third term, we have that

$$\frac{(1-\beta_2)\beta_2^{t-1}v_0}{\sqrt{v_{t-1}}+\sqrt{v_t}} \le \frac{(1-\beta_2)\beta_2^{t-1}v_0}{\sqrt{\beta_2^{t-1}v_0}+\sqrt{\beta_2^t v_0}} = (1-\sqrt{\beta_2})\beta_2^{\frac{t-1}{2}}\sqrt{v_0}.$$

So we have that

$$\begin{aligned}
\sqrt{v_{t-1}}-\sqrt{v_t} &\le \frac{2C_1\sqrt{v_{t-1}}\sqrt{\beta_2+(1+C_2)\beta_2^2}}{\sqrt{v_{t-1}}+\sqrt{v_t}} + (1-\sqrt{\beta_2})\beta_2^{\frac{t-1}{2}}\sqrt{v_0}\\
&\le 2\sqrt{\beta_2+(1+C_2)\beta_2^2}C_1 + (1-\sqrt{\beta_2})\beta_2^{\frac{t-1}{2}}\sqrt{v_0}\\
&\le 2\sqrt{\beta_2+(1+C_2)\beta_2^2}C_1 + (1-\sqrt{\beta_2})\beta_2^{\frac{t-1}{2}}\left|g_1\right|.
\end{aligned}$$

When $|g_t| \ge \sqrt{v_{t-1}}$, we define $f(u) = \sqrt{\beta_2 v_{t-1}+(1-\beta_2)u^2}$. Then $\sqrt{v_t} = f(|g_t|)$ and $\sqrt{v_{t-1}} = f(\sqrt{v_{t-1}})$. Since $f(u)$ is a convex function w.r.t. u, when $|g_t| \ge \sqrt{v_{t-1}}$ we have

$$0 \le \sqrt{v_t}-\sqrt{v_{t-1}} \le f'(|g_t|)(|g_t|-\sqrt{v_{t-1}}) = \frac{(1-\beta_2)\left|g_t\right|}{\sqrt{\beta_2 v_{t-1}+(1-\beta_2)\left|g_t\right|^2}}(|g_t|-\sqrt{v_{t-1}}). \quad (3)$$

We further define $e_{t-1} = \beta_2^{t-1}\sqrt{v_0} + (1-\beta_2)\sum_{i=0}^{t-2}\beta_2^i\,|g_{t-1-i}|$. We can verify that

$$\begin{aligned}
&\beta_2^{t-1}(\sqrt{v_0} - e_{t-1})^2 + \sum_{i=0}^{t-2}(1-\beta_2)\beta_2^i\left(|g_{t-1-i}| - e_{t-1}\right)^2 \\
=&\beta_2^{t-1}v_0 + (1-\beta_2)\sum_{i=0}^{t-2}\beta_2^i g_{t-1-i}^2 + \left(\beta_2^{t-1} + (1-\beta_2)\left(\sum_{i=0}^{t-1}\beta_2^i\right)\right)e_{t-1}^2 \\
-&2e_{t-1}\left(\beta_2^{t-1}\sqrt{v_0} + (1-\beta_2)\sum_{i=0}^{t-2}\beta_2^i\,|g_{t-1-i}|\right) \\
=&v_{t-1} + e_{t-1}^2 - 2e_{t-1}^2 = v_{t-1} - e_{t-1}^2.
\end{aligned} \tag{4}$$

Then we define the gap between $|g_t|$ and e_{t-1} as $\Delta_t := |g_t| - e_{t-1}$. From the assumption, we have that

$$\begin{aligned}
\Delta_t &\le \beta_2^{t-1}\,|g_t - \sqrt{v_0}| + (1-\beta_2)\sum_{i=0}^{t-2}\beta_2^i\,|g_t - g_{t-i-1}| \\
&\le \beta_2^{t-1}\,|g_t - \sqrt{v_0}| + (1-\beta_2)\sum_{i=1}^{t-1}\beta_2^i C_1 i\sqrt{1 + \frac{\beta_2 C_2}{(1-\beta_2)i}} \\
&\le \beta_2^{t-1}\,|g_t - \sqrt{v_0}| + C_1(1-\beta_2)\sum_{i=1}^{t-1}\beta_2^i i\left(1 + \frac{\beta_2 C_2}{(1-\beta_2)i}\right) \\
&= \beta_2^{t-1}\,|g_t - \sqrt{v_0}| + \frac{C_1\beta_2(1+C_2)}{1-\beta_2}
\end{aligned}$$

Define $k = \min\left\{[\frac{\beta_2 C_2}{1-\beta_2}], [\frac{(1-\beta_2)\Delta_t^2}{8C_1^2\beta_2 C_2}]\right\}$. If such k is 0, then $\frac{(1-\beta_2)\Delta_t^2}{8C_1^2\beta_2 C_2} < 1$.[2] We can get a stronger bound $\Delta_t \le \frac{\sqrt{8\beta_2 C_2}C_1}{\sqrt{1-\beta_2}}$. Then we have

$$\begin{aligned}
\frac{(1-\beta_2)\,|g_t|}{\sqrt{\beta_2 v_{t-1} + (1-\beta_2)\,|g_t|^2}}(|g_t| - \sqrt{v_{t-1}}) &\le \frac{(1-\beta_2)\,|g_t|}{\sqrt{(1-\beta_2)\,|g_t|^2}}(|g_t| - \sqrt{v_{t-1}}) \\
&= \sqrt{1-\beta_2}(|g_t| - \sqrt{v_{t-1}}) \\
&\le \sqrt{1-\beta_2}\Delta_t \\
&\le \sqrt{8\beta_2 C_2}C_1.
\end{aligned}$$

2. We would assume $1-\beta_2$ very small and hence $\frac{\beta_2 C_2}{1-\beta_2}$ must be greater than 1.

When such k is a positive integer, we have that $1 \leq \frac{\beta_2 C_2}{(1-\beta_2)i}$ and $C_1\sqrt{\frac{2\beta_2 C_2 i}{1-\beta_2}} \leq \frac{\Delta_t}{2}$ when $i \leq k$. Then we have

$$\begin{aligned}
|g_{t-i}| &\geq |g_t| - \eta H i\sqrt{1+\frac{\beta_2 C_2}{(1-\beta_2)i}} \\
&\geq |g_t| - C_1\sqrt{\frac{2\beta_2 C_2 i}{1-\beta_2}} \\
&\geq e_t + \frac{\Delta_t}{2}.
\end{aligned}$$

We only consider the first k terms in Equation 4,

$$\begin{aligned}
v_{t-1} - e_{t-1}^2 &\geq \sum_{i=0}^{k-1}(1-\beta_2)\beta_2^i\left(|g_{t-1-i}| - e_{t-1}\right)^2 \\
&\geq \sum_{i=0}^{k-1}(1-\beta_2)\beta_2^i\frac{\Delta_t^2}{4} \\
&= (1-\beta_2^k)\frac{\Delta_t^2}{4}.
\end{aligned}$$

Then we have

$$\begin{aligned}
v_{t-1} &\geq e_{t-1}^2 + \frac{1-\beta_2^k}{4}\Delta_t^2 \\
&= (g_t - \Delta_t)^2 + \frac{1-\beta_2^k}{4}\Delta_t^2 \\
&\geq \frac{1-\beta_2^k}{4}\left((|g_t| - \Delta_t)^2 + \Delta_t^2\right) \\
&= \frac{1-\beta_2^k}{4}\left(\frac{g_t^2}{2} + \left(\frac{|g_t|}{\sqrt{2}} - \sqrt{2}\Delta_t\right)^2\right) \\
&\geq \frac{1-\beta_2^k}{8}g_t^2 \geq \frac{k(1-\beta_2)}{8}g_t^2
\end{aligned}$$

Back to Equation 3, we have that

$$\begin{aligned}
\frac{(1-\beta_2)|g_t|}{\sqrt{\beta_2 v_{t-1} + (1-\beta_2)g_t^2}}(|g_t| - \sqrt{v_{t-1}}) &\leq \frac{\sqrt{8}(1-\beta_2)|g_t|}{\sqrt{\beta_2 k(1-\beta_2)}|g_t|}(|g_t| - \sqrt{v_{t-1}}) \\
&\leq \frac{\sqrt{8(1-\beta_2)}}{\sqrt{\beta_2 k}}(|g_t| - e_t) \\
&= \frac{\sqrt{8(1-\beta_2)}}{\sqrt{\beta_2 k}}\Delta_t.
\end{aligned}$$

When $\frac{\beta_2 C_2}{1-\beta_2} \geq \frac{(1-\beta_2)\Delta_t^2}{8C_1^2\beta_2 C_2}$, $k = [\frac{(1-\beta_2)\Delta_t^2}{8C_1^2\beta_2 C_2}] \geq \frac{(1-\beta_2)\Delta_t^2}{16C_1^2\beta_2 C_2}$. Then $\frac{\sqrt{8(1-\beta_2)}}{\sqrt{\beta_2 k}}\Delta_t \leq 8\sqrt{2C_2}C_1$. When $\frac{\beta_2 C_2}{1-\beta_2} \leq \frac{(1-\beta_2)\Delta_t^2}{8C_1^2\beta_2 C_2}$, $k = [\frac{\beta_2 C_2}{1-\beta_2}] \geq \frac{\beta_2 C_2}{2(1-\beta_2)}$. Then we have that

$$\begin{aligned}
\frac{\sqrt{8(1-\beta_2)}}{\sqrt{\beta_2 k}}\Delta_t &\leq \frac{4(1-\beta_2)}{\beta_2\sqrt{C_2}}\Delta_t \leq \frac{4(1+C_2)}{\sqrt{C_2}}C_1 + \frac{4(1-\beta_2)\beta_2^{t-2}}{\sqrt{C_2}}\left|g_t - \sqrt{v_0}\right| \\
&= \frac{4(1+C_2)}{\sqrt{C_2}}C_1 + \frac{4(1-\beta_2)\beta_2^{t-2}}{\sqrt{C_2}}\left|g_t - g_1\right| \\
&\leq \frac{4(1+C_2)}{\sqrt{C_2}}C_1 + \frac{4(1-\beta_2)\beta_2^{t-2}}{\sqrt{C_2}}C_1(t-1)\sqrt{1 + \frac{\beta_2 C_2}{(1-\beta_2)(t-1)}} \\
&\leq \frac{4(1+C_2)}{\sqrt{C_2}}C_1 + \frac{4(1-\beta_2)\beta_2^{t-2}}{\sqrt{C_2}}C_1(t-1)\left(1 + \frac{\beta_2 C_2}{(1-\beta_2)(t-1)}\right) \\
&\leq \frac{4(1+C_2)}{\sqrt{C_2}}C_1 + 8\frac{C_1}{\sqrt{C_2}} + 4\sqrt{C_2}C_1 = \frac{4C_1(3+2C_2)}{\sqrt{C_2}}.
\end{aligned}$$

The last inequality is because $\max_t \beta_2^{t-2}(t-1) \leq \frac{2}{1-\beta_2}$. ■

Lemma D.3 *If we assume there exists C_3 such that $\ln \frac{v_t+\epsilon}{v_s+\epsilon} \leq C_3$ for all $1 \leq s, t \leq T$, then we have*

$$\sum_{t=1}^{T}\left|1 - \frac{\sqrt{v_{t-1}+\epsilon}}{\sqrt{v_t+\epsilon}}\right| \leq \frac{3}{2}\frac{1-\beta_2}{\beta_2}T + \frac{C_3}{2}.$$

Proof [Proof of Lemma D.3] When $\sqrt{v_t+\epsilon} \leq \sqrt{v_{t-1}+\epsilon}$, we have $\sqrt{v_t+\epsilon} = \sqrt{(1-\beta_2)g_t^2 + \beta_2 v_{t-1} + \epsilon} \geq \sqrt{\beta_2}\sqrt{v_{t-1}+\epsilon}$. Then

$$0 \geq 1 - \frac{\sqrt{v_{t-1}+\epsilon}}{\sqrt{v_t+\epsilon}} \geq 1 - \frac{1}{\sqrt{\beta_2}} \geq 1 - \frac{1}{\beta_2} = \frac{\beta_2 - 1}{\beta_2}.$$

We have $\left|1 - \frac{\sqrt{v_{t-1}+\epsilon}}{\sqrt{v_t+\epsilon}}\right| \leq \frac{1-\beta_2}{\beta_2}$.

When $\sqrt{v_t+\epsilon} > \sqrt{v_{t-1}+\epsilon}$, $\left|1 - \frac{\sqrt{v_{t-1}+\epsilon}}{\sqrt{v_t+\epsilon}}\right| = 1 - \frac{\sqrt{v_{t-1}+\epsilon}}{\sqrt{v_t+\epsilon}} \leq \ln \frac{\sqrt{v_t+\epsilon}}{\sqrt{v_{t-1}+\epsilon}}$.

When combining both parts, we have

$$\begin{aligned}
\sum_{t=1}^{T}\left|1-\frac{\sqrt{v_{t-1}+\epsilon}}{\sqrt{v_t+\epsilon}}\right| &= \sum_{\sqrt{v_t+\epsilon}\le\sqrt{v_{t-1}+\epsilon}}\left|1-\frac{\sqrt{v_{t-1}+\epsilon}}{\sqrt{v_t+\epsilon}}\right| + \sum_{\sqrt{v_t+\epsilon}>\sqrt{v_{t-1}+\epsilon}}\left|1-\frac{\sqrt{v_{t-1}+\epsilon}}{\sqrt{v_t+\epsilon}}\right| \\
&\le \frac{1-\beta_2}{\beta_2}T + \sum_{\sqrt{v_t+\epsilon}>\sqrt{v_{t-1}+\epsilon}}\ln\frac{\sqrt{v_t+\epsilon}}{\sqrt{v_{t-1}+\epsilon}} \\
&= \frac{1-\beta_2}{\beta_2}T + \ln\frac{\sqrt{v_T+\epsilon}}{\sqrt{v_0+\epsilon}} - \sum_{\sqrt{v_t+\epsilon}\le\sqrt{v_{t-1}+\epsilon}}\ln\frac{\sqrt{v_t+\epsilon}}{\sqrt{v_{t-1}+\epsilon}} \\
&\le \frac{1-\beta_2}{\beta_2}T + \ln\frac{\sqrt{v_T+\epsilon}}{\sqrt{v_0+\epsilon}} - \sum_{\sqrt{v_t+\epsilon}\le\sqrt{v_{t-1}+\epsilon}}\ln\sqrt{\beta_2} \\
&\le \frac{1-\beta_2}{\beta_2}T + \ln\frac{\sqrt{v_T+\epsilon}}{\sqrt{v_0+\epsilon}} + \frac{T}{2}\left(-\ln\beta_2\right) \\
&\le \frac{1-\beta_2}{\beta_2}T + \ln\frac{\sqrt{v_T+\epsilon}}{\sqrt{v_0+\epsilon}} + \frac{T}{2}\left(1-\beta_2\right) \\
&\le \frac{3}{2}\frac{1-\beta_2}{\beta_2}T + \frac{C_3}{2}
\end{aligned}$$

■

Lemma D.4 *Suppose there exists C_1, C_2, C_3 such that $|g_t - g_s| \le C_1(t-s)\sqrt{1+\frac{\beta_2 C_2}{(1-\beta_2)(t-s)}}$ for any $1 \le s < t \le T$ and $\ln\frac{v_t+\epsilon}{v_s+\epsilon} \le C_3$ for any $1 \le s, t \le T$. Then it holds that*

$$\begin{aligned}
\sum_{t=1}^{T}\frac{g_t^2}{\sqrt{v_t+\epsilon}} \ge \sum_{t=1}^{T}\frac{v_t-\beta_2^t v_0}{2\sqrt{v_t+\epsilon}} - \frac{18(1+C_2)C_1}{\sqrt{C_2}}T \\
- 6\frac{(1+C_2)C_1C_3\beta_2}{\sqrt{C_2}(1-\beta_2)} - \frac{3(1-\sqrt{\beta_2})}{2}T\left|g_1\right| - \frac{\beta_2 C_3}{2(1+\sqrt{\beta_2})}\left|g_1\right|
\end{aligned}$$

Proof [Proof of Lemma D.4]

$$
\begin{aligned}
&\sum_{t=1}^{T} \frac{g_t^2}{\sqrt{v_t+\epsilon}} \\
&= \sum_{t=1}^{T} \frac{(v_t+\epsilon) - \beta_2(v_{t-1}+\epsilon) - (1-\beta_2)\epsilon}{(1-\beta_2)\sqrt{v_t+\epsilon}} \\
&= \sum_{t=1}^{T} \frac{1+\beta_2}{1-\beta_2}\sqrt{v_t+\epsilon} - \frac{2\beta_2}{1-\beta_2}\sqrt{v_{t-1}+\epsilon} - \frac{\beta_2}{1-\beta_2}\frac{(\sqrt{v_t+\epsilon}-\sqrt{v_{t-1}+\epsilon})^2}{\sqrt{v_t+\epsilon}} - \frac{\epsilon}{\sqrt{v_t+\epsilon}} \\
&= \frac{2\beta_2}{1-\beta_2}\sqrt{v_T+\epsilon} - \frac{2\beta_2}{1-\beta_2}\sqrt{v_0+\epsilon} + \sum_{t=1}^{T}\left(\sqrt{v_t+\epsilon} - \frac{\epsilon}{\sqrt{v_t+\epsilon}}\right) - \frac{\beta_2}{1-\beta_2}\sum_{t=1}^{T}\frac{(\sqrt{v_t+\epsilon}-\sqrt{v_{t-1}+\epsilon})^2}{\sqrt{v_t+\epsilon}} \\
&= \frac{2\beta_2}{1-\beta_2}\sqrt{v_T+\epsilon} - \frac{2\beta_2}{1-\beta_2}\sqrt{v_0+\epsilon} + \sum_{t=1}^{T}\frac{v_t}{\sqrt{v_t+\epsilon}} - \frac{\beta_2}{1-\beta_2}\sum_{t=1}^{T}\frac{(\sqrt{v_t+\epsilon}-\sqrt{v_{t-1}+\epsilon})^2}{\sqrt{v_t+\epsilon}} \\
&\geq \frac{2\beta_2}{1-\beta_2}(\sqrt{v_T+\epsilon} - \sqrt{v_0+\epsilon}) + \sum_{t=1}^{T}\frac{v_t}{\sqrt{v_t+\epsilon}} - \frac{\beta_2}{1-\beta_2}\sum_{t=1}^{T}\frac{(\sqrt{v_t+\epsilon}-\sqrt{v_{t-1}+\epsilon})^2}{\sqrt{v_t+\epsilon}} \\
&\geq \frac{2\beta_2}{1-\beta_2}(\sqrt{\beta_2^T v_0+\epsilon} - \sqrt{v_0+\epsilon}) + \sum_{t=1}^{T}\frac{\beta_2^t v_0}{\sqrt{\beta_2^t v_0+\epsilon}} + \sum_{t=1}^{T}\left(\frac{v_t}{\sqrt{v_t+\epsilon}} - \frac{\beta_2^t v_0}{\sqrt{\beta_2^t v_0+\epsilon}}\right) \\
&- \frac{\beta_2}{1-\beta_2}\sum_{t=1}^{T}\frac{(\sqrt{v_t+\epsilon}-\sqrt{v_{t-1}+\epsilon})^2}{\sqrt{v_t+\epsilon}}
\end{aligned}
$$

We prove that $\frac{\beta_2 u}{\sqrt{\beta_2 u+\epsilon}} \geq \frac{2\beta_2}{1-\beta_2}(\sqrt{u+\epsilon} - \sqrt{\beta_2 u+\epsilon})$ for any positive value. Actually the right hand side equals to $\frac{2\beta_2}{1-\beta_2}\frac{(1-\beta_2)u}{\sqrt{u+\epsilon}+\sqrt{\beta_2 u+\epsilon}} = \frac{2\beta_2 u}{\sqrt{u+\epsilon}+\sqrt{\beta_2 u+\epsilon}}$ and the inequality comes from $\sqrt{u+\epsilon} \geq \sqrt{\beta_2 u+\epsilon}$. Therefore, we can get that

$$
\sum_{t=1}^{T}\frac{\beta_2^t v_0}{\sqrt{\beta_2^t v_0+\epsilon}} \geq \sum_{t=1}^{T}\frac{2\beta_2}{1-\beta_2}\left(\sqrt{\beta_2^{t-1}v_0+\epsilon} - \sqrt{\beta_2^t v_0+\epsilon}\right) = \frac{2\beta_2}{1-\beta_2}\left(\sqrt{v_0+\epsilon} - \sqrt{\beta_2^T v_0+\epsilon}\right)
$$

and the first two terms are canceled out.

For the third term, define $f(x) = \frac{x}{\sqrt{x+\epsilon}}$. $f(x)$ is a concave function and $v_t \geq \beta_2^t v_0$. Then we have that

$$
\begin{aligned}
\frac{v_t}{\sqrt{v_t+\epsilon}} - \frac{\beta_2^t v_0}{\sqrt{\beta_2^t v_0+\epsilon}} &= f(v_t) - f(\beta_2^t v_0) \geq f'(v_t)(v_t - \beta_2^t v_0) \\
&= \frac{v_t+2\epsilon}{2(v_t+\epsilon)^{1.5}}(v_t - \beta_2^t v_0) \\
&\geq \frac{v_t - \beta_2^t v_0}{2\sqrt{v_t+\epsilon}}.
\end{aligned}
$$

From Lemma D.2 and Lemma D.3, we have that

$$\begin{aligned}
&\sum_{t=2}^{T} \frac{(\sqrt{v_t+\epsilon}-\sqrt{v_{t-1}+\epsilon})^2}{\sqrt{v_t+\epsilon}} \\
\leq& \left(\frac{12(1+C_2)}{\sqrt{C_2}}C_1 + (1-\sqrt{\beta_2})\left|g_1\right|\right)\sum_{t=2}^{T}\left|1-\frac{\sqrt{v_{t-1}+\epsilon}}{\sqrt{v_t+\epsilon}}\right| \\
\leq& \left(\frac{12(1+C_2)}{\sqrt{C_2}}C_1 + (1-\sqrt{\beta_2})\left|g_1\right|\right)\left(\frac{3}{2}\frac{1-\beta_2}{\beta_2}T+\frac{C_3}{2}\right) \\
\leq& \frac{18(1+C_2)C_1}{\sqrt{C_2}\beta_2}(1-\beta_2)T + 6\frac{(1+C_2)C_1C_3}{\sqrt{C_2}} + \frac{3(1-\sqrt{\beta_2})}{2\beta_2}(1-\beta_2)T\left|g_1\right| + \frac{1-\sqrt{\beta_2}}{2}C_3\left|g_1\right|,
\end{aligned}$$

which finishes the proof. ■

Appendix E. Convergence rate of RMSProp for stochastic loss

Additional Notations. We define $\bar{\boldsymbol{v}}_t = \beta_2\bar{\boldsymbol{v}}_{t-1}+(1-\beta_2)\bar{\boldsymbol{g}}_t^2$ as the exponential moving average of squared deterministic gradient and $\hat{\boldsymbol{v}}_t = \mathbb{E}[\boldsymbol{v}_t|L_{<t}] = (1-\beta_2)\mathbb{E}[\boldsymbol{g}_t^2|L_{<t}]+\beta_2\boldsymbol{v}_{t-1}$ as the conditional expectation of $\boldsymbol{v}_t$ conditional on steps before time t.

E.1. Proof sketch

Before presenting the complete proof, we first present the key steps that perform the analysis in a coordinate-wise manner and don't rely on the boundness of gradient.

In a single step, we first focus on the second order term which will be bounded with the following Lemma E.6 when assuming the gradient coordinate is Lipschitz w.r.t. ℓ_∞ norm. It allows us to have different H_i for each coordinates while the coefficient will be the same H when assuming ℓ_2 norm smoothness. The dependence on ℓ_2 smoothness coefficient will be d times worse.

Lemma E.1 (second order term) *Under Definition 2.3, we have for any $\boldsymbol{x}$ and any $\boldsymbol{\Delta}\in\mathbb{R}^d$*

$$\boldsymbol{\Delta}^\top\nabla^2 L(\boldsymbol{x})\boldsymbol{\Delta} \leq \sum_{i=1}^{d} H_i\boldsymbol{\Delta}_i^2.$$

The first-order term is often a big challenge because it is hard to take expectation when the denominator $\boldsymbol{v}_t$ has $\boldsymbol{g}_t$. Previous works will replace the stochastic $\boldsymbol{v}_t$ with $\hat{\boldsymbol{v}}_t$ which is fixed conditional on steps before time t and show the gap induced by the replacement is small because of boundness of gradients. Instead, in Lemma E.5 we will replace $\mathbb{E}\frac{g_{t,i}\bar{g}_{t,i}}{\sqrt{v_{t,i}+\epsilon}}$ by $\mathbb{E}\frac{\bar{g}_{t,i}^2}{\sqrt{\bar{v}_{t,i}+\sigma_i^2+\epsilon/2}}$ and show the gap can be bounded by noise σ_i while sacrificing some constant.

When dealing with $\mathbb{E}\frac{\bar{g}_{t,i}^2}{\sqrt{\bar{v}_{t,i}+\sigma_i^2+\epsilon/2}}$, we also don't rely on the upper bound of $\bar{v}_{t,i}$ to relax it into $O(\bar{g}_{t,i}^2)$. In Lemma D.4, we employ some nontrivial telescoping sum technique to connect $\mathbb{E}\sum_{i=1}^{T}\frac{\bar{g}_{t,i}^2}{\sqrt{\bar{v}_{t,i}+\sigma_i^2+\epsilon/2}}$ with $\mathbb{E}\sum_{i=1}^{T}\frac{\bar{v}_{t,i}-\beta_2^t\bar{v}_{0,i}}{\sqrt{\bar{v}_{t,i}+\sigma_i^2+\epsilon/2}}$ while the gap can be handled when $\bar{g}_{t,i}$ doesn't move too fast. Actually we control the moving speed of $\bar{g}_{t,i}$ in with Lipschitz assumption and only

need the moving speed of $\boldsymbol{x}_t$ is not too fast, which is shown in Lemma D.1 as an extension from Lemma 4.2 in Xie and Li [24].

Then we employ Jensen's inequality to extract many $\mathbb{E}\frac{\bar{g}_{t,i}^2}{\sqrt{\bar{g}_{t,i}^2+\sigma_i^2+\epsilon/2}}$ from $\mathbb{E}\sum_{i=1}^T \frac{\bar{v}_{t,i}-\beta_2^t\bar{v}_{0,i}}{\sqrt{\bar{v}_{t,i}+\sigma_i^2+\epsilon/2}}$. It can infer that the minimum value of $\frac{\|\bar{\boldsymbol{g}}_t\|_1^2}{\|\bar{\boldsymbol{g}}_t\|_1+\sum_{i=1}^d\sigma_i+d\sqrt{\epsilon/2}}$ is smaller than the E defined in Theorem 2.6 and we solve the quadratic function to get an upper bound for $\|\boldsymbol{g}_t\|_1$.

E.2. Complete Proof

We first recall the definitions of some notations. $\boldsymbol{g}_t$ denotes the gradient of mini-batch $L_t(\boldsymbol{x}_{t-1})$ at step t. And $\mathbb{E}\left[\boldsymbol{g}_t|\boldsymbol{x}_{t-1}\right]=\nabla L(\boldsymbol{x}_{t-1})$ because $\mathbb{E}L_t=L$. The full-batch gradient is $\bar{\boldsymbol{g}}_t=\nabla L(\boldsymbol{x}_{t-1})$. Different kinds of second-order momentum are defined in the following way.

$$
\begin{aligned}
v_{t,i}&=\beta_2^t g_{1,i}^2+(1-\beta_2)\sum_{j=0}^{t-1}\beta_2^j g_{t-j,i}^2\\
\hat{v}_{t,i}&=(1-\beta_2)\mathbb{E}\left[g_{t,i}^2|\boldsymbol{x}_{t-1}\right]+\beta_2 v_{t-1,i}=\mathbb{E}\left[v_{t,i}|\boldsymbol{x}_{t-1}\right]\\
\bar{v}_{t,i}&=\beta_2^t \bar{g}_{1,i}^2+(1-\beta_2)\sum_{j=0}^{t-1}\beta_2^j \bar{g}_{t-j,i}^2
\end{aligned}
$$

Here are the assumptions.

Definition 2.3 *For any $\mathbf{H}=(H_1,\dots,H_d)\in\mathbb{R}^d$, we say a function L is* $\mathbf{H}$-smooth coordinate-wisely w.r.t. ℓ_∞ norm , *iff for any $i\in[d]$, $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d$, $|\nabla_i L(\boldsymbol{x})-\nabla_i L(\boldsymbol{y})|\le H_i\left\|\boldsymbol{x}-\boldsymbol{y}\right\|_\infty$.*

Assumption 2.4 (Bounded coordinate-wise noise) *There exist constants σ_i such that $|\nabla_i L_t(\boldsymbol{x})-\nabla_i L(\boldsymbol{x})|\le\sigma_i$ for any $i\in[d]$, $t\in\mathbb{N}$ and $\boldsymbol{x}\in\mathbb{R}^d$.*

Assumption 2.5 (Bounded coordinate-wise stochastic gradient) *For any $i\in[d]$, $t\in\mathbb{N}$ and $\boldsymbol{x}\in\mathbb{R}^d$, $|\nabla_i L_t(\boldsymbol{x})|\le G$.*

Lemma E.2 *Let $C=\ln\frac{G^2+\max_i\sigma_i^2+\epsilon}{\epsilon}$. It holds that $\ln\frac{v_{t,i}+\epsilon}{v_{s,i}+\epsilon}\le C$ and $\ln\frac{\bar{v}_{t,i}+\epsilon}{\bar{v}_{s,i}+\epsilon}\le C$ for any $i\in[d]$ and any $t\ge s\ge 1$.*

First we give an upper bound for $v_{t,i}$ and $\hat{v}_{t,i}$ with $\bar{v}_{t,i}$.

Lemma E.3

$$\hat{v}_{t,i}\le 2\bar{v}_{t,i}+2\sigma_i^2$$

Proof

$$v_{t,i}=(1-\beta_2)\sum_{j=0}^{t-1}\beta_2^j g_{t-j,i}^2\le(1-\beta_2)\sum_{j=0}^{t-1}\beta_2^j(\bar{g}_{t-j,i}+\sigma_i)^2\le(1-\beta_2)\sum_{j=0}^{t-1}\beta_2^j\cdot 2(\bar{g}_{t-j,i}^2+\sigma_i^2)=2\bar{v}_t+2\sigma_i^2.$$

$$\hat{v}_{t,i}=\mathbb{E}\left[v_{t,i}|\boldsymbol{x}_{t-1}\right]\le 2\bar{v}_{t.i}+2\sigma_i^2.$$

■

Lemma E.4 (Distance between gradients) *Let $g_{t,i}$, $m_{t,i}$ and $v_{t,i}$ be defined in* `Adam`*. Then we have that*

$$|\bar{g}_{t,i} - \bar{g}_{t-k,i}| \le \eta H_i k \sqrt{1 + \frac{\beta_2 - \beta_1}{(1-\beta_2)(1-\beta_1)k} C}.$$

Specifically, when $\beta_1 = 0$, we have that

$$|\bar{g}_{t,i} - \bar{g}_{t-k,i}| \le \eta H_i k \sqrt{1 + \frac{\beta_2}{(1-\beta_2)k} C}.$$

Proof [Proof of Lemma E.4] From the lipschitzness we can have that

$$|\bar{g}_{t,i} - \bar{g}_{t-k,i}| \le H_i \left\| \boldsymbol{x}_t - \boldsymbol{x}_{t-k} \right\|_\infty.$$

For each coordinate $j \in [d]$, we can have

$$\begin{aligned}
|x_{t,j} - x_{t-k,j}| &= \left| \sum_{l=t-k+1}^{t} \eta \frac{m_{l,j}}{\sqrt{v_{l,j} + \epsilon}} \right| \\
&= \left| \sum_{l=t-k+1}^{t} \eta \frac{\beta_1^l m_{0,j} + (1-\beta_1) \sum_{n=0}^{l-1} \beta_1^n \bar{g}_{l-n,j}}{\sqrt{v_{l,j} + \epsilon}} \right| \\
&\le \left(\sum_{l=t-k+1}^{t} \eta \frac{\beta_1^l m_{0,j}^2 + (1-\beta_1) \sum_{n=0}^{l-1} \beta_1^n \bar{g}_{l-n,j}^2}{v_{l,j} + \epsilon} \right)^{\frac{1}{2}} \left(\sum_{l=t-k+1}^{t} \eta \left(\beta_1^l + (1-\beta_1) \sum_{n=0}^{l-1} \beta_1^n \right) \right)^{\frac{1}{2}}
\end{aligned}$$

If we define $x_t = \beta_1^t m_{0,j}^2 + (1-\beta_1) \sum_{n=0}^{t-1} \beta_1^n \bar{g}_{t-n,j}^2$ and $y_t = v_{t,j} + \epsilon$, then $\{x_t\}$ and $\{y_t\}$ satisfy the condition in Lemma D.1 and we can have

$$\sum_{l=t-k+1}^{t} \eta \frac{\beta_1^l m_{0,j}^2 + (1-\beta_1) \sum_{n=0}^{l-1} \beta_1^n \bar{g}_{l-n,j}^2}{v_{l,j} + \epsilon} \le \eta k + \frac{(\beta_2 - \beta_1)\eta C}{(1-\beta_2)(1-\beta_1)}.$$

And we know that

$$\sum_{l=t-k+1}^{t} \eta \left(\beta_1^l + (1-\beta_1) \sum_{n=0}^{l-1} \beta_1^n \right) = \sum_{l=t-k+1}^{t} \eta = \eta k.$$

Therefore, we have that

$$|x_{t,j} - x_{t-k,j}| \le \eta k \sqrt{1 + \frac{\beta_2 - \beta_1}{(1-\beta_2)(1-\beta_1)k} C}$$

and it is the same upper bound for $\left\| \boldsymbol{x}_t - \boldsymbol{x}_{t-k} \right\|_\infty$. ■

Lemma E.5 (first-order approximation, no momentum)

$$\begin{aligned}\mathbb{E}\sum_{t=1}^{T}\frac{g_{t,i}\bar{g}_{t,i}}{\sqrt{v_{t,i}+\epsilon}} \geq \frac{1}{4\sqrt{2}}\mathbb{E}\sum_{t=1}^{T}\frac{\bar{v}_{t,i}-\beta_2^t\bar{v}_{0,i}}{\sqrt{\bar{v}_{t,i}+\sigma_i^2+\epsilon/2}} - \frac{9(1+C)\eta H_i}{\sqrt{2}\sqrt{C}}T \\ - \frac{3\sqrt{C}(1+C)\beta_2}{\sqrt{2}(1-\beta_2)}\eta H_i - \frac{3(1-\sqrt{\beta_2})}{4\sqrt{2}}T\mathbb{E}\left|\bar{g}_{1,i}\right| - \frac{C}{4\sqrt{2}(1+\sqrt{\beta_2})}\mathbb{E}\left|\bar{g}_{1,i}\right| \\ - 4\sqrt{1-\beta_2}T\sigma_i - \frac{4\sigma_i C}{\sqrt{1-\beta_2}}\end{aligned}$$

Proof [Proof of Lemma E.5] The first order change can decomposed into two terms.[3]

$$\begin{aligned}\mathbb{E}\sum_{t=1}^{T}\frac{g_t\bar{g}_t}{\sqrt{v_t+\epsilon}} &= \mathbb{E}\sum_{t=1}^{T}\frac{g_t\bar{g}_t}{\sqrt{\hat{v}_t+\epsilon}} + \mathbb{E}\left[\sum_{t=1}^{T}\frac{g_t\bar{g}_t}{\sqrt{v_t+\epsilon}} - \frac{g_t\bar{g}_t}{\sqrt{\hat{v}_t+\epsilon}}\right] \\ &= \mathbb{E}\sum_{t=1}^{T}\mathbb{E}\left[\frac{g_t\bar{g}_t}{\sqrt{\hat{v}_t+\epsilon}}\middle|\boldsymbol{x}_{t-1}\right] + \mathbb{E}\left[\sum_{t=1}^{T}\frac{g_t\bar{g}_t}{\sqrt{v_t+\epsilon}} - \frac{g_t\bar{g}_t}{\sqrt{\hat{v}_t+\epsilon}}\right] \\ &= \mathbb{E}\sum_{t=1}^{T}\frac{\bar{g}_t^2}{\sqrt{\hat{v}_t+\epsilon}} + \mathbb{E}\left[\sum_{t=1}^{T}\frac{g_t\bar{g}_t}{\sqrt{v_t+\epsilon}} - \frac{g_t\bar{g}_t}{\sqrt{\hat{v}_t+\epsilon}}\right]\end{aligned} \tag{5}$$

For the second term, we have that

$$\begin{aligned}\left|g_t\bar{g}_t\left(\frac{1}{\sqrt{v_t+\epsilon}} - \frac{1}{\sqrt{\hat{v}_t+\epsilon}}\right)\right| &= \frac{\left|g_t\bar{g}_t\left(\hat{v}_t - v_t\right)\right|}{\sqrt{v_t+\epsilon}\sqrt{\hat{v}_t+\epsilon}\left(\sqrt{v_t+\epsilon}+\sqrt{\hat{v}_t+\epsilon}\right)} \\ &= \frac{\left|g_t\bar{g}_t(1-\beta_2)\left(\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right] - g_t^2\right)\right|}{\sqrt{v_t+\epsilon}\sqrt{\hat{v}_t+\epsilon}\left(\sqrt{v_t+\epsilon}+\sqrt{\hat{v}_t+\epsilon}\right)} \\ &= \frac{\left|g_t\bar{g}_t(1-\beta_2)\left(\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}+g_t\right)\left(\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}-g_t\right)\right|}{\sqrt{v_t+\epsilon}\sqrt{\hat{v}_t+\epsilon}\left(\sqrt{v_t+\epsilon}+\sqrt{\hat{v}_t+\epsilon}\right)} \\ &\leq \frac{1}{2}\frac{\bar{g}_t^2}{\sqrt{\hat{v}_t+\epsilon}}\frac{\left(\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}+g_t\right)^2}{\mathbb{E}[\left(\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}+g_t\right)^2|\boldsymbol{x}_{t-1}]} \\ &\quad + \frac{1}{2}\frac{(1-\beta_2)^2g_t^2\mathbb{E}[\left(\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}+g_t\right)^2|\boldsymbol{x}_{t-1}]\left(\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}-g_t\right)^2}{(v_t+\epsilon)\sqrt{\hat{v}_t+\epsilon}\left(\sqrt{v_t+\epsilon}+\sqrt{\hat{v}_t+\epsilon}\right)^2}\end{aligned}$$

We know that

$$\mathbb{E}\left[\frac{\bar{g}_t^2}{\sqrt{\hat{v}_t+\epsilon}}\frac{\left(\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}+g_t\right)^2}{\mathbb{E}[\left(\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}+g_t\right)^2|\boldsymbol{x}_{t-1}]}\middle|\boldsymbol{x}_{t-1}\right] = \frac{\bar{g}_t^2}{\sqrt{\hat{v}_t+\epsilon}}\frac{\mathbb{E}[\left(\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}+g_t\right)^2|\boldsymbol{x}_{t-1}]}{\mathbb{E}[\left(\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}+g_t\right)^2|\boldsymbol{x}_{t-1}]} = \frac{\bar{g}_t^2}{\sqrt{\hat{v}_t+\epsilon}}$$

3. We omit subscript i for simplicity.

and

$$
\begin{aligned}
&\frac{(1-\beta_2)^2 g_t^2 \mathbb{E}[\left(\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}+g_t\right)^2|\boldsymbol{x}_{t-1}]\left(\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}-g_t\right)^2}{(v_t+\epsilon)\sqrt{\hat{v}_t+\epsilon}\left(\sqrt{v_t+\epsilon}+\sqrt{\hat{v}_t+\epsilon}\right)^2} \\
\leq&\frac{(1-\beta_2)^2 g_t^2 \mathbb{E}[\left(2\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]+2g_t^2\right)|\boldsymbol{x}_{t-1}]\left(\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}-g_t\right)^2}{(v_t+\epsilon)\sqrt{\hat{v}_t+\epsilon}\left(\sqrt{v_t+\epsilon}+\sqrt{\hat{v}_t+\epsilon}\right)^2} \\
=&\frac{(1-\beta_2)^2 g_t^2 4\mathbb{E}[g_t^2|\boldsymbol{x}_{t-1}]\left(\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}-g_t\right)^2}{(v_t+\epsilon)\sqrt{\hat{v}_t+\epsilon}\left(\sqrt{v_t+\epsilon}+\sqrt{\hat{v}_t+\epsilon}\right)^2} \\
\leq&4(1-\beta_2)^2\frac{g_t^2}{v_t+\epsilon}\frac{\mathbb{E}[g_t^2|\boldsymbol{x}_{t-1}]}{\hat{v}_t+\epsilon}\frac{\left(\left|\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}\right|+|g_t|\right)}{\sqrt{v_t+\epsilon}+\sqrt{\hat{v}_t+\epsilon}}\left|\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}-g_t\right| \\
\leq&4(1-\beta_2)^2\frac{g_t^2}{v_t+\epsilon}\frac{1}{1-\beta_2}\frac{1}{\sqrt{1-\beta_2}}\left|\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}-g_t\right| \\
=&4\sqrt{1-\beta_2}\frac{g_t^2}{v_t+\epsilon}\left|\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}-g_t\right| \\
\leq&4\sqrt{1-\beta_2}\frac{g_t^2}{v_t+\epsilon}\left(\left|\sqrt{\mathbb{E}\left[g_t^2|\boldsymbol{x}_{t-1}\right]}-\bar{g}_t\right|+|\bar{g}_t-g_t|\right) \\
\leq&8\sqrt{1-\beta_2}\frac{g_t^2}{v_t+\epsilon}\sigma.
\end{aligned}
$$

Then back to Equation 5, we have that

$$
\begin{aligned}
\mathbb{E}\sum_{t=1}^T\frac{g_t\bar{g}_t}{\sqrt{v_t+\epsilon}}&=\mathbb{E}\sum_{t=1}^T\frac{\bar{g}_t^2}{\sqrt{\hat{v}_t+\epsilon}}+\mathbb{E}\left[\sum_{t=1}^T\frac{g_t\bar{g}_t}{\sqrt{v_t+\epsilon}}-\frac{g_t\bar{g}_t}{\sqrt{\hat{v}_t+\epsilon}}\right] \\
&\geq\mathbb{E}\sum_{t=1}^T\frac{\bar{g}_t^2}{\sqrt{\hat{v}_t+\epsilon}}-\frac{1}{2}\mathbb{E}\sum_{t=1}^T\frac{\bar{g}_t^2}{\sqrt{\hat{v}_t+\epsilon}}-\frac{1}{2}8\sqrt{1-\beta_2}\sigma\mathbb{E}\sum_{t=1}^T\frac{g_t^2}{v_t+\epsilon} \\
&=\frac{1}{2}\mathbb{E}\sum_{t=1}^T\frac{\bar{g}_t^2}{\sqrt{\hat{v}_t+\epsilon}}-\frac{1}{2}8\sqrt{1-\beta_2}\sigma\mathbb{E}\sum_{t=1}^T\frac{g_t^2}{v_t+\epsilon} \\
&\geq\frac{1}{2}\mathbb{E}\frac{1}{\sqrt{2}}\sum_{t=1}^T\frac{\bar{g}_t^2}{\sqrt{\bar{v}_t+\sigma^2+\epsilon/2}}-\frac{1}{2}8\sqrt{1-\beta_2}\sigma\mathbb{E}\sum_{t=1}^T\frac{g_t^2}{v_t+\epsilon}
\end{aligned}
$$

For the first term, we can show that $\{\bar{g}_{t,i}\}$ and $\{\bar{v}_{t,i}\}$ satisfy the condition for Lemma D.4 with $C_1=\eta H_i, C_2=C, C_3=C$ and $\epsilon=\sigma_i^2+\epsilon/2$ based on Lemma E.2 and Lemma E.4. Then we can

have

$$\sum_{t=1}^{T}\frac{\bar{g}_{t,i}^2}{\sqrt{\bar{v}_{t,i}+\sigma_i^2+\epsilon/2}} \geq \sum_{t=1}^{T}\frac{\bar{v}_{t,i}-\beta_2^t v_0}{2\sqrt{\bar{v}_{t,i}+\sigma_i^2+\epsilon/2}} - \frac{18(1+C)\eta H_i}{\sqrt{C}}T \\ - 6\frac{\sqrt{C}(1+C)\beta_2\eta H_i}{1-\beta_2} - \frac{3(1-\sqrt{\beta_2})}{2}T\left|\bar{g}_{1,i}\right| - \frac{\beta_2 C}{2(1+\sqrt{\beta_2})}\left|\bar{g}_{1,i}\right|.$$

For the second term, we can apply Lemma D.1 with $\beta_1 = 0$ and get that

$$\sum_{t=1}^{T}\frac{g_t^2}{v_t+\epsilon} \leq T + \frac{C}{1-\beta_2}.$$

Combining these two terms, we can get that

$$\mathbb{E}\sum_{t=1}^{T}\frac{g_{t,i}\bar{g}_{t,i}}{\sqrt{v_{t,i}+\epsilon}} \geq \frac{1}{4\sqrt{2}}\mathbb{E}\sum_{t=1}^{T}\frac{\bar{v}_{t,i}-\beta_2^t\bar{v}_{0,i}}{\sqrt{\bar{v}_{t,i}+\sigma_i^2+\epsilon/2}} - \frac{9(1+C)\eta H_i}{\sqrt{2}\sqrt{C}}T \\ - \frac{3\sqrt{C}(1+C)\beta_2}{\sqrt{2}(1-\beta_2)}\eta H_i - \frac{3(1-\sqrt{\beta_2})}{4\sqrt{2}}T\mathbb{E}\left|\bar{g}_{1,i}\right| - \frac{C}{4\sqrt{2}(1+\sqrt{\beta_2})}\mathbb{E}\left|\bar{g}_{1,i}\right| \\ - 4\sqrt{1-\beta_2}T\sigma_i - \frac{4\sigma_i C}{\sqrt{1-\beta_2}}$$

■

Lemma E.6 (second order term) *Under Definition 2.3, we have for any $\boldsymbol{x}$ and any $\boldsymbol{\Delta} \in \mathbb{R}^d$*

$$\boldsymbol{\Delta}^\top\nabla^2 L(\boldsymbol{x})\boldsymbol{\Delta} \leq \sum_{i=1}^{d} H_i\boldsymbol{\Delta}_i^2.$$

Proof From Definition 2.3, we know that $\left|\nabla^2 L(\boldsymbol{x})_{i,:}\boldsymbol{\Delta}\right| \leq H_i\left\|\boldsymbol{\Delta}\right\|_\infty$ for any $\boldsymbol{\Delta} \in \mathbb{R}^d$. When $\boldsymbol{\Delta}$ is chosen as $\operatorname{sign}\left(\nabla^2 L(\boldsymbol{x})_{i,:}\right)$, we have $\sum_{j=1}^{d}\left|\nabla^2 L(\boldsymbol{x})_{ij}\right| \leq H_i$.

$$\begin{aligned}
2\sum_{i=1}^{d} H_i\Delta_i^2 &\geq 2\sum_{i=1}^{d}\sum_{j=1}^{d}\left|\nabla^2 L(\boldsymbol{x})_{ij}\right|\Delta_i^2 \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d}\left|\nabla^2 L(\boldsymbol{x})_{ij}\right|\Delta_i^2 + \sum_{i=1}^{d}\sum_{j=1}^{d}\left|\nabla^2 L(\boldsymbol{x})_{ji}\right|\Delta_i^2 \\
&\geq \sum_{i=1}^{d}\sum_{j=1}^{d}\left|\nabla^2 L(\boldsymbol{x})_{ij}\right|\Delta_i^2 + \sum_{i=1}^{d}\sum_{j=1}^{d}\left|\nabla^2 L(\boldsymbol{x})_{ij}\right|\Delta_j^2 \\
&\geq \sum_{i=1}^{d}\sum_{j=1}^{d} 2\nabla^2 L(\boldsymbol{x})_{ij}\Delta_i\Delta_j \\
&= 2\boldsymbol{\Delta}^\top\nabla^2 L(\boldsymbol{x})\boldsymbol{\Delta}.
\end{aligned}$$

■

Theorem E.7 *Under all the assumptions, we have that*

$$\mathbb{E}[\min_{1\le t\le T}\|\nabla L(\boldsymbol{x}_t)\|_1] \le E + \sqrt{E(\sum_{i=1}^d \sigma_i + d\sqrt{\epsilon/2})}$$

with $E = O\left(\frac{1}{\sqrt{T}}\sqrt{\sqrt{C}\,(L(\boldsymbol{x}_0) - L(\boldsymbol{x}_T))\sum_{i=1}^d H_i} + \frac{\sqrt{(\log T + C)}}{\sqrt{T}}\sum_{i=1}^d \sigma_i\right)$.

Proof In a single step we have that

$$\begin{aligned} L(\boldsymbol{x}_t) - L(\boldsymbol{x}_{t-1}) &\le \nabla L(\boldsymbol{x}_{t-1})^\top(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) + \frac{1}{2}\sum_{i=1}^d H_i\,(x_{t,i} - x_{t-1,i})^2 \\ &= -\eta\sum_{i=1}^d \frac{g_{t,i}\bar{g}_{t,i}}{\sqrt{v_{t,i}+\epsilon}} + \frac{1}{2}\eta^2\sum_{i=1}^d H_i\frac{g_{t,i}^2}{v_{t,i}+\epsilon}. \end{aligned}$$

If we sum over t from 1 to T and take expectation, we can get

$$\begin{aligned} \mathbb{E}\left[L(\boldsymbol{x}_T) - L(\boldsymbol{x}_0)\right] &\le -\mathbb{E}\left[\eta\sum_{i=1}^d\sum_{t=1}^T \frac{g_{t,i}\bar{g}_{t,i}}{\sqrt{v_{t,i}+\epsilon}}\right] + \frac{1}{2}\eta^2\mathbb{E}\left[\sum_{i=1}^d H_i\sum_{t=1}^T\frac{g_{t,i}^2}{v_{t,i}+\epsilon}\right] \\ &\le -\mathbb{E}\left[\eta\sum_{i=1}^d\sum_{t=1}^T \frac{g_{t,i}\bar{g}_{t,i}}{\sqrt{v_{t,i}+\epsilon}}\right] + \frac{1}{2}\eta^2\mathbb{E}\left[\sum_{i=1}^d H_i\left(T + \frac{\beta_2}{1-\beta_2}C\right)\right]. \end{aligned}$$

The second inequality comes from applying Lemma D.1 with β_1 as we did in Lemma E.5. Then we also apply Lemma E.5 to deal with the first term and get that

$$\begin{aligned} &\mathbb{E}\left[L(\boldsymbol{x}_T) - L(\boldsymbol{x}_0)\right] \\ \le& -\mathbb{E}\left[\eta\sum_{i=1}^d\sum_{t=1}^T \frac{g_{t,i}\bar{g}_{t,i}}{\sqrt{v_{t,i}}}\right] + \frac{1}{2}\eta^2\mathbb{E}\left[\sum_{i=1}^d\sum_{t=1}^T H_i\frac{g_{t,i}^2}{v_{t,i}}\right] \\ \le& -\frac{\eta}{4\sqrt{2}}\mathbb{E}\sum_{i=1}^d\sum_{t=1}^T\frac{\bar{v}_{t,i} - \beta_2^t\bar{v}_{0,i}}{\sqrt{\bar{v}_{t,i} + \sigma_i^2 + \epsilon/2}} + \eta\sum_{i=1}^d\frac{9(1+C)\eta H_i}{\sqrt{2}\sqrt{C}}T \\ &+ \eta\sum_{i=1}^d\frac{3\sqrt{C}(1+C)\beta_2}{\sqrt{2}(1-\beta_2)}\eta H_i + \eta\sum_{i=1}^d\frac{3(1-\sqrt{\beta_2})}{4\sqrt{2}}T\mathbb{E}\left|\bar{g}_{1,i}\right| + \eta\sum_{i=1}^d\frac{C}{4\sqrt{2}(1+\sqrt{\beta_2})}\mathbb{E}\left|\bar{g}_{1,i}\right| \\ &+ \eta\sum_{i=1}^d 4\sqrt{1-\beta_2}T\sigma_i + \eta\sum_{i=1}^d\frac{4\sigma_i C}{\sqrt{1-\beta_2}} + \frac{1}{2}\eta^2\sum_{i=1}^d H_i\left(T + \frac{\beta_2}{1-\beta_2}C\right) \end{aligned}$$

$$\begin{aligned}\frac{\mathbb{E}\sum_{t=1}^{T}\left\|\frac{\bar{\boldsymbol{v}}_t-\beta_2^t\bar{\boldsymbol{v}}_0}{\sqrt{\bar{\boldsymbol{v}}_t+\boldsymbol{\sigma}^2+\epsilon/2}}\right\|_1}{T} &\le \frac{4\sqrt{2}\left(L(\boldsymbol{x}_0)-L(\boldsymbol{x}_T)\right)}{\eta T}\\ &+\left(\frac{36(1+C)}{\sqrt{C}}+\sqrt{2}\right)\eta\sum_{i=1}^{d}H_i+\frac{(12\sqrt{C}(1+C)+\sqrt{C})\beta_2\eta}{(1-\beta_2)T}\sum_{i=1}^{d}H_i\\ &+6(1-\sqrt{\beta_2})\mathbb{E}\left\|\bar{\boldsymbol{g}}_1\right\|_1+\frac{C}{(1+\sqrt{\beta_2})T}\mathbb{E}\left\|\bar{\boldsymbol{g}}_1\right\|_1\\ &+16\sqrt{2}\sqrt{1-\beta_2}\sum_{i=1}^{d}\sigma_i+\frac{16\sqrt{2}C}{\sqrt{1-\beta_2}T}\sum_{i=1}^{d}\sigma_i\end{aligned}$$

When we choose $1-\beta_2=\Omega(\frac{\log T+C}{T})$ and $\eta=\Theta\left(\sqrt{\frac{L(\boldsymbol{x}_0)-L(\boldsymbol{x}_T)}{\sqrt{C}\sum_{i=1}^{d}H_iT}}\right)$, the right hand side is

$$O\left(\frac{1}{\sqrt{T}}\sqrt{\sqrt{C}\left(L(\boldsymbol{x}_0)-L(\boldsymbol{x}_T)\right)\sum_{i=1}^{d}H_i}+\frac{\sqrt{(\log T+C)}}{\sqrt{T}}\sum_{i=1}^{d}\sigma_i\right).$$

For the left hand side, we choose T_1 such that $\frac{\beta_2^{T_1}}{T(1-\beta_2)}\frac{\bar{v}_{0,i}}{\sqrt{\sigma_i^2+\epsilon}}\le\frac{\sqrt{\epsilon}}{T^{100}}$ for every $i\in[d]$. It can be relaxed into

$$\frac{\beta_2^{T_1}}{T(1-\beta_2)}\frac{\bar{v}_{0,i}}{\sqrt{\sigma_i^2+\epsilon}\sqrt{\epsilon}}\le\frac{\beta_2^{T_1}}{T(1-\beta_2)}\frac{\bar{v}_{0,i}}{\epsilon}\le\frac{\beta_2^{T_1}}{T(1-\beta_2)}e^C\le\frac{1}{T^{100}}$$

because $C=\ln\frac{G^2+\sigma^2+\epsilon}{\epsilon}\ge\ln\frac{G^2}{\epsilon}\ge\ln\frac{\bar{v}_{0,i}}{\epsilon}$. It requires that

$$T_1\ge\Omega(\frac{C+\ln(1-\beta_2)+\ln T}{1-\beta_2})$$

which can be satisfied when $1-\beta_2=\Omega(\frac{\log T+C}{T})$ and $T_1=\frac{T}{3}$. Then we have that

$$\begin{aligned}\sum_{t=1}^{T}\left\|\frac{\bar{\boldsymbol{v}}_t-\beta_2^t\bar{\boldsymbol{v}}_0}{\sqrt{\bar{\boldsymbol{v}}_t+\boldsymbol{\sigma}^2+\epsilon/2}}\right\|_1 &\ge\sum_{t=T_1}^{T}\left\|\frac{\bar{\boldsymbol{v}}_t-\beta_2^t\bar{\boldsymbol{v}}_0}{\sqrt{\bar{\boldsymbol{v}}_t+\boldsymbol{\sigma}^2+\epsilon/2}}\right\|_1\\ &\ge\sum_{t=T_1}^{T}\left\|\frac{\bar{\boldsymbol{v}}_t}{\sqrt{\bar{\boldsymbol{v}}_t+\boldsymbol{\sigma}^2+\epsilon/2}}\right\|_1-\sum_{t=T_1}^{T}\left\|\frac{\beta_2^t\bar{\boldsymbol{v}}_0}{\sqrt{\bar{\boldsymbol{v}}_t+\boldsymbol{\sigma}^2+\epsilon/2}}\right\|_1\\ &\ge\sum_{t=T_1}^{T}\left\|\frac{\bar{\boldsymbol{v}}_t}{\sqrt{\bar{\boldsymbol{v}}_t+\boldsymbol{\sigma}^2+\epsilon/2}}\right\|_1-\sum_{t=T_1}^{T}\left\|\frac{\beta_2^t\bar{\boldsymbol{v}}_0}{\sqrt{\boldsymbol{\sigma}^2+\epsilon/2}}\right\|_1\\ &=\sum_{t=T_1}^{T}\left\|\frac{\bar{\boldsymbol{v}}_t}{\sqrt{\bar{\boldsymbol{v}}_t+\boldsymbol{\sigma}^2+\epsilon/2}}\right\|_1-\frac{\beta_2^{T_1}}{1-\beta_2}\left\|\frac{\bar{\boldsymbol{v}}_0}{\sqrt{\boldsymbol{\sigma}^2+\epsilon/2}}\right\|_1\\ &\ge\sum_{t=T_1}^{T}\left\|\frac{\bar{\boldsymbol{v}}_t}{\sqrt{\bar{\boldsymbol{v}}_t+\boldsymbol{\sigma}^2+\epsilon/2}}\right\|_1-\frac{d\sqrt{\epsilon}}{T^{100}}.\end{aligned}$$

If we define $f(x) = \frac{x}{\sqrt{x+\sigma^2+\epsilon/2}}$, $f(x)$ is a concave function. Then we can apply Jensen's inequality on $\bar{v}_{t,i} = \beta_2^t \bar{v}_{0,i} + (1-\beta_2)\sum_{j=0}^{t-1} \beta_2^{t-i} \bar{g}_{j,i}^2$ and get that

$$f(\bar{v}_{t,i}) \geq \beta_2^t f(\bar{v}_{0,i}) + (1-\beta_2)\sum_{j=1}^{t-1} \beta_2^{t-j} f(\bar{g}_{j,i}^2).$$

Then we further have

$$\begin{aligned}
\sum_{t=T_1}^{T} \left\| \frac{\bar{\boldsymbol{v}}_t}{\sqrt{\bar{\boldsymbol{v}}_t + \boldsymbol{\sigma}^2 + \epsilon/2}} \right\|_1 &\geq \sum_{t=T_1}^{T} \left(\beta_2^t \left\| \frac{\bar{\boldsymbol{v}}_0}{\sqrt{\bar{\boldsymbol{v}}_0 + \boldsymbol{\sigma}^2 + \epsilon/2}} \right\|_1 + (1-\beta_2) \sum_{j=1}^{t-1} \beta_2^{t-j} \left\| \frac{\bar{\boldsymbol{g}}_j^2}{\sqrt{\bar{\boldsymbol{g}}_j^2 + \boldsymbol{\sigma}^2 + \epsilon/2}} \right\|_1 \right) \\
&\geq \sum_{t=T_1}^{T} (1-\beta_2^{T-t+1}) \left\| \frac{\bar{\boldsymbol{g}}_t^2}{\sqrt{\bar{\boldsymbol{g}}_t^2 + \boldsymbol{\sigma}^2 + \epsilon/2}} \right\|_1 \\
&\geq \sum_{t=T_1}^{T_2} \frac{1}{2} \left\| \frac{\bar{\boldsymbol{g}}_t^2}{\sqrt{\bar{\boldsymbol{g}}_t^2 + \boldsymbol{\sigma}^2 + \epsilon/2}} \right\|_1
\end{aligned}$$

when choosing $T_2 = T - O(\frac{1}{1-\beta_2})$. And we can show that $\left\| \frac{\bar{\boldsymbol{g}}_t^2}{\sqrt{\bar{\boldsymbol{g}}_t^2+\boldsymbol{\sigma}^2+\epsilon/2}} \right\|_1 \geq \frac{\|\bar{\boldsymbol{g}}_t\|_1^2}{\left\|\sqrt{\bar{\boldsymbol{g}}_t^2+\boldsymbol{\sigma}^2+\epsilon/2}\right\|_1} \geq \frac{\|\bar{\boldsymbol{g}}_t\|_1^2}{\|\bar{\boldsymbol{g}}_t\|_1+\sum_{i=1}^d \sigma_i + d\sqrt{\epsilon/2}}$. Then we have that

$$\begin{aligned}
&\mathbb{E} \min_{T_1 \leq t \leq T_2} \frac{\|\bar{\boldsymbol{g}}_t\|_1^2}{\|\bar{\boldsymbol{g}}_t\|_1 + \sum_{i=1}^d \sigma_i + d\sqrt{\epsilon/2}} \\
\leq &\frac{1}{T_2 - T_1 + 1} \mathbb{E} \sum_{t=T_1}^{T_2} \frac{\|\bar{\boldsymbol{g}}_t\|_1^2}{\|\bar{\boldsymbol{g}}_t\|_1 + \sum_{i=1}^d \sigma_i + d\sqrt{\epsilon/2}} \\
\leq &\frac{2}{T_2 - T_1 + 1} \left(\mathbb{E} \sum_{t=1}^{T} \left\| \frac{\bar{\boldsymbol{v}}_t - \beta_2^t \bar{\boldsymbol{v}}_0}{\sqrt{\bar{\boldsymbol{v}}_t + \boldsymbol{\sigma}^2 + \epsilon/2}} \right\|_1 + \frac{d\sqrt{\epsilon}}{T^{100}} \right) \\
:= &E = O\left(\frac{1}{\sqrt{T}} \sqrt{\sqrt{C}\left(L(\boldsymbol{x}_0) - L(\boldsymbol{x}_T)\right) \sum_{i=1}^d H_i} \right) + O\left(\frac{\sqrt{(\log T + C)}}{\sqrt{T}} \sum_{i=1}^d \sigma_i \right).
\end{aligned}$$

Then we consider the increasing function $h(x) = \frac{x^2}{x+\sum_{i=1}^d \sigma_i + d\sqrt{\epsilon/2}}$ and solve $h(x) = E$. We can get

$$
\begin{aligned}
&\mathbb{E} \min_{T_1 \le t \le T_2} \|\bar{\boldsymbol{g}}_t\|_1 \le h^{-1}(E_1) \\
=&\frac{E_1}{2} + \sqrt{E(\sum_{i=1}^d \sigma_i + d\sqrt{\epsilon/2}) + \frac{E^2}{4}} \\
\le& E + \sqrt{E(\sum_{i=1}^d \sigma_i + d\sqrt{\epsilon/2})} \\
\le& E + \sqrt{E\sum_{i=1}^d \sigma_i} + \sqrt{Ed\sqrt{\epsilon/2}} \\
\le& O(\frac{1}{\sqrt{T}} \sqrt{\sqrt{C}\left(L(\boldsymbol{x}_0) - L(\boldsymbol{x}_T)\right) \sum_{i=1}^d H_i} + \frac{\sqrt{(\log T + C)}}{\sqrt{T}} \sum_{i=1}^d \sigma_i \\
&+ \frac{\sqrt{\sum_{i=1}^d \sigma_i} \left(\sqrt{C}\left(L(\boldsymbol{x}_0) - L(\boldsymbol{x}_T)\right) \sum_{i=1}^d H_i\right)^{1/4}}{T^{1/4}} + \frac{\left(\sum_{i=1}^d \sigma_i\right)(\log T + C)^{1/4}}{T^{1/4}} \\
&+ \frac{\sqrt{d\sqrt{\epsilon/2}} \left(\sqrt{C}\left(L(\boldsymbol{x}_0) - L(\boldsymbol{x}_T)\right) \sum_{i=1}^d H_i\right)^{1/4}}{T^{1/4}} + \frac{\sqrt{(\sum_{i=1}^d \sigma_i) d\sqrt{\epsilon/2}}(\log T + C)^{1/4}}{T^{1/4}})
\end{aligned}
$$

■

Appendix F. Experiment Details

We conduct our experiments on a GPT-2 [18] of size 125M parameters. We train the model on the OpenWebText corups containing more than 9B tokens for 100k iterations.

F.1. Training Adam On a Rotated Loss

A key difficulty in implementing `rotated Adam` arises from applying an orthogonal rotation on the parameters before calculating the loss. It is computationally infeasible to apply a 125M $\times$ 125M orthogonal matrix on the 125M-sized parameter vector. To avoid such computation, we design a new orthogonal transformer to rotate the parameters of the network. In what follows, we elaborate on this rotation.

`RandPerm`. Given a vector v of size d, we can orthogonally rotate it by repeatedly applying these consecutive operations: 1. Permute the entries of the vector according to a randomly chosen permutation $\pi \in \mathbb{S}_d$. 2. Reshape the permuted vector into a 3D tensor of size $[s_1, s_2, s_3]$, apply a fixed orthogonal rotation of size $s \times s$ on each side of the tensor and then reshape it back to a vector of size d.

This operation performs an orthogonal transformation $\mathcal{R}$ on the input vector v. We can chain multiple operations of this kind and construct $\texttt{RandPerm}^k$, where k is a positive number indicating

the number of consecutive `RandPerm` s applied. Building upon this rotation, we train GPT-2 125M with `Adam` on $L \circ \texttt{RandPerm}^2$ to analyze our hypothesis regarding the ℓ_∞ geometry of the loss landscape and to verify that `Adam` will indeed suffer from the induced orthogonal equivariance. Figure 1 confirms our findings, as the performance of `rotated Adam` with $\texttt{RandPerm}^2$ is significantly worse than `Adam`. This suggests that `Adam` is highly sensitive to the rotation and adaptivity alone can't explain its advantage.

F.2. Computation of $(1,1)$-norm

It is impossible to get the full Hessian matrix and sum over all the entries. Therefore, we propose Algorithm 3 that leverages Hessian vector product function in pytorch to estimate the $(1,1)$-norm of Hessian matrix.

Algorithm 3 $(1,1)$-Norm Estimation

Require: C: number of Cauchy vectors
Require: θ: parameters of the network, vector of size $\mathbb{R}^d$.
Require: $B = \{b_1, b_2, \cdots b_g\}$: set of batches of data of size g
Require: $\mathcal{T} : \mathbb{R}^d \to \mathbb{R}^d$: a transformation
Require: $L : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}^+$ a loss function mapping data and parameters to a positive scalar.
H $\leftarrow \{\}$.
for $i = 1 \to C$:
  Sample a Cauchy vector $v \in \mathbb{R}^d$ from $\Gamma(0,1)/d$.
  Set $h = \mathbf{0}$
  for $b \in B$:
    Set $L_b \leftarrow L(b, \cdot)$
    $h \leftarrow h \ + \ \nabla^2(L_b \circ \mathcal{T})(\mathcal{T}^{-1}(\theta))[v]$
  Append h/g to H.
return $\sum(\text{median}(\text{abs}(H), \text{axis} = 1))$