Linear attention is (maybe) all you need (to understand transformer optimization)

Kwangjun Ahn⋆
MIT EECS/LIDS
kjahn@mit.edu

Xiang Cheng⋆
MIT LIDS
chengx@mit.edu

Minhak Song⋆
KAIST ISysE/Math
minhaksong@kaist.ac.kr

Chulhee Yun
KAIST AI
chulhee.yun@kaist.ac.kr

Ali Jadbabaie
MIT CEE/LIDS
jadbabai@mit.edu

Suvrit Sra
MIT EECS/LIDS
suvrit@mit.edu

Abstract

Transformer training is notoriously difficult, requiring a careful design of optimizers and use of various heuristics. We make progress towards understanding the subtleties of training transformers by carefully studying a simple yet canonical linearized *shallow* transformer model. Specifically, we train linear transformers to solve regression tasks, inspired by J. von Oswald *et al.* (ICML 2023), and K. Ahn et al. (NeurIPS 2023). Most importantly, we observe that our proposed linearized models can reproduce several prominent aspects of transformer training dynamics. Consequently, the results obtained in this paper suggest that a simple linearized transformer model could actually be a valuable, realistic abstraction for understanding transformer optimization.

1 Introduction

Transformer architectures [Vaswani et al., 2017] (henceforth, referred to as *transformers*) have shown impressive performance in various applications [Devlin et al., 2019, Bubeck et al., 2023]. However, training transformers is notoriously difficult and laborious; see, e.g., observations given by Liu et al. [2020] as well as scaling laws [Kaplan et al., 2020]. In particular, training transformers requires carefully designed optimizers as well as use of various heuristics. For instance, as illustrated in Figure 1, stochastic gradient descent (SGD)—the workhorse of most deep learning optimization problems—fails to train transformers effectively. This failure is in contrast to the success of SGD when applied to train convolutional neural networks (CNNs) on vision tasks.

Several recent papers propose a number of different explanations as to why transformer optimization is so difficult. There is a general consensus in the literature that the loss landscape of transformers has a number of distinctive features that significantly differ from standard optimization theory assumptions. Most notably, it is empirically verified through various experiments that stochastic gradient noise is heavy-tailed and non-Gaussian [Zhang et al., 2020b, Kunstner et al., 2023] and the loss landscape is significantly ill-conditioned [Zhang et al., 2020a, Jiang et al., 2022, Pan and Li, 2023]. In particular, standard assumptions are incapable of dealing with and explaining these observations, and as a result, transformer optimization has become more of an art than science.

A major obstacle in understanding transformer optimization is that full-fledged transformers are extremely complicated to model. One can probe the transformer's properties by measuring quantities, such as gradient norm or smoothness, but it is much harder to parse the inner-layer workings, and to

⋆Equal contribution, alphabetical order.

Mathematics of Modern Machine Learning Workshop at NeurIPS 2023.

satisfactorily answer questions such as: *why* does the loss landscape have such features, or *how* do algorithms like Adam perform better than SGD in transformer training?

For these reasons, having an appropriate *mathematical abstraction* is necessary for progress in understanding transformer optimization — an abstraction that is as simple as possible, while still able to capture the essence of transformer optimization. The main message of this paper is that these distinctive features of transformer training also arise in a far simpler setting: we propose that the *linear attention model* is precisely the abstraction that we are looking for. We verify that training this model on a low-dimensional linear regression task displays all the distinctive features that have been observed on the full transformer, suggesting that our surprisingly simple model can serve as a testbed for rigorous understanding of transformer optimization.

As a preliminary to our discussion, we first survey the previous works that seek to characterize and understand the transformer optimization landscape.

2 Distinctive features of transformer optimization

Numerous recent papers have identified a number of distinctive features of the transformer optimization problem, which set it apart from commonly studied optimization objectives, or even other neural networks such as CNNs. As shown in Figure 1, one of the most striking features is the following:

Adaptive method like **Adam are significantly better than SGD**!	(Adam>SGD)

This is in stark contrast with the training of other neural networks (e.g., convolutional neural networks) for which several works have shown that the values of adaptive methods are marginal [Wilson et al., 2017]. This phenomenon sparked the interest of the optimization community in investigating the main causes, and subsequently, recent works [Zhang et al., 2020b, Kunstner et al., 2023, Jiang et al., 2022, Pan and Li, 2023] have identified various "unique" features of transformer optimization.

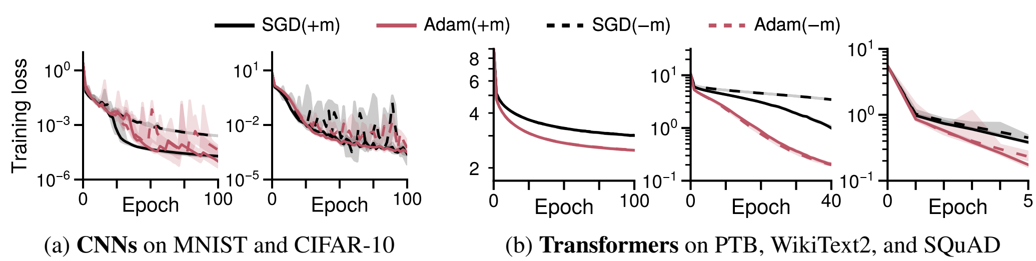


Figure 1: Adaptive optimization methods like Adam are much more effective than SGD for training transformers. This experimental result is taken from [Kunstner et al., 2023, Figure 1]. (+m) denotes "with momentum".

In this section, we discuss them one by one in detail, building preliminaries for our main results. In order to discuss each feature, we first give a whirlwind tour on some background in optimization – see, e.g., monographs [Bubeck, 2015, Nesterov et al., 2018] for greater contexts.

2.1 A whirlwind tour of (convex) optimization theory

For a symmetric matrix M, we denote by $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ the largest and smallest eigenvalue of M, and by $\|M\|_2$ the spectral norm of M. For simplicity, we assume the training loss function f is twice differentiable. We introduce the following standard concepts in the optimization literature.

- **Lipschitzness.** We say f is G-Lipschitz if $\|\nabla f\|_2 \le G$.
- **Smoothness.** We say f is L-smooth if $\left\|\nabla^2 f\right\|_2 \le L$.
- **Strong convexity.** We say f is μ-strongly convex if $\lambda_{\min}(\nabla^2 f) \ge \mu$.
- **Condition number.** The (local) condition number $\kappa_f(x)$ is defined as ${\lambda_{\max}(\nabla^2 f(x))}/{\lambda_{\min}(\nabla^2 f(x))}$, provided that $\lambda_{\min}(\nabla^2 f(x)) > 0$.
- **Bounded stochastic gradient noise.** In most SGD analyses, it is assumed that the stochastic gradient $g(x)$ satisfies the *bounded variance* property: $\mathbb{E}\left\|g(x) - \nabla f(x)\right\|^2 \le \sigma^2$.

Transformers (in practice)

Shallow linear transformers
(see Appendix B and Table 1)

1. Gap between Adam vs. SGD [Zhang et al., 2020b, Kunstner et al., 2023, Jiang et al., 2022, Pan and Li, 2023]**:**

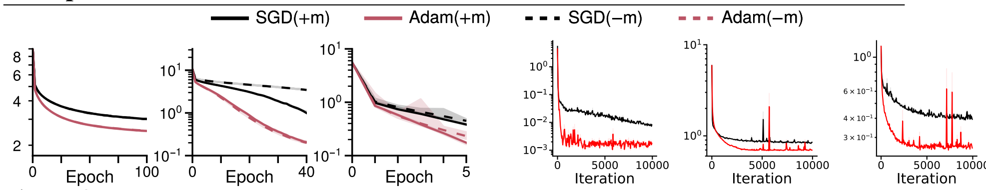


Figure 2: For transformer optimization, adaptive methods like Adam are strictly better than SGD. (+m) denotes "with momentum" and (-m) denotes without momentum. Our plots only show the momentum variants of SGD and Adam as they perform better in all cases.
Left 3 plots: Full transformers, from [Kunstner et al., 2023, Figure 1].
Right 3 plots: Shallow linear transformers (see Settings 1, 2, and 3 from Table 1).

2. Heavy-tailed stochastic gradient noise [Zhang et al., 2020b, Kunstner et al., 2023]**:**

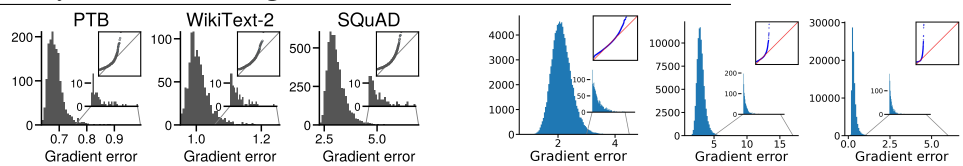


Figure 3: The stochastic gradient noise is heavy-tailed for transformer optimization.
Left 3 plots: Full transformers, from [Kunstner et al., 2023, Figure 1].
Right 3 plots: Shallow linear transformers (see Settings 1, 2, and 3 from Table 1).

3. Robust condition number of the landscape [Jiang et al., 2022]**:**

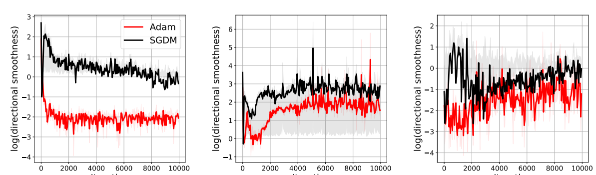
Layer#	Iteration 750 $R^{\text{SGD}}_{\text{med}}/R^{\text{Adam}}_{\text{med}}$	Iteration 1250 $R^{\text{SGD}}_{\text{med}}/R^{\text{Adam}}_{\text{med}}$
15	1.65 (0.65)	2.01 (1.00)
17	1.91 (0.53)	1.43 (0.63)
22	3.54 (1.21)	2.28 (1.18)

	Iteration 750 $R^{\text{SGD}}_{\text{med}}/R^{\text{Adam}}_{\text{med}}$	Iteration 1250 $R^{\text{SGD}}_{\text{med}}/R^{\text{Adam}}_{\text{med}}$
Setting 1	1.76 (0.40)	1.58 (0.41)
Setting 2	3.14 (0.97)	5.98 (2.86)
Setting 3	9.57 (13.3)	6.53 (3.55)

Figure 4: The comparison of the robust condition number (see Subsection A.2) between SGD and Adam for transformer optimization. Numbers in parentheses show standard deviation. Left table: Full transformers, from [Jiang et al., 2022, Table 1]. Right table: Shallow linear transformers, see Table 1.

4. Directional smoothness gap between SGD v.s Adam [Zhang et al., 2020a, Pan and Li, 2023]**:**

Figure 5: Plot of log(directional smoothness) against iteration# (see Subsection A.3) for shallow linear transformers (see Settings 1, 2, 3 from Table 1).



The features defined above are typically of great importance in the theory of convex optimization, as the convergence rate of gradient-based optimizers (e.g., gradient descent) typically depends on these quantities [Bubeck, 2015, Nesterov et al., 2018].

2.2 Distinctive features of the transformer optimization

Building on the concepts from Subsection 2.1, we now discuss the previous studies on transformer optimization. We summarize them here; see Appendix A for details.

- **Heavy-tailed gradient noise [Zhang et al., 2020b, Kunstner et al., 2023].** In [Zhang et al., 2020b], it was observed that the stochastic gradient is typically more heavy-tailed for transformer optimization than other neural network optimization. In particular, they make a case that this is opposed to the standard bounded variance condition for SGD analysis – see Figure 3 and Figure 6. They posit that this phenomenon might be one of the main reasons behind (Adam>SGD).

- **Ill-conditioned landscape [Jiang et al., 2022].** In another inspiring work by Jiang et al. [2022], authors seek to understand the difficulty of transformer optimization through the lens of condition number. In particular, they consider a "robust" version of condition number defined as $R^{\mathsf{OPT}}_{\mathsf{med}} := \lambda_{\max}(\nabla^2 f)/\lambda_{\text{median}}(\nabla^2 f)$[1]. They observe that during transformer optimization, non-adaptive optimizers like SGD tend to have larger robust condition number than adaptive optimizers like Adam; they posit that this phenomenon is one of the main reasons for (Adam>SGD) – see Figure 4.
- **Directional Smoothness [Pan and Li, 2023].** In a follow up work by Pan and Li [2023], the authors again corroborate (Adam>SGD), and further observe that adaptive optimizers tend to have smaller "directional smoothness" values (formally defined in Subsection A.3). Once again, Pan and Li [2023] hypothesize that this feature is unique to transformers.

3 Linear shallow transformers have the same loss landscape as practical deep transformers

In this section, we show that a simple yet canonical transformer model exhibits all the features in Section 2. Specifically, the optimization problem to be solved is the training of **linear transformers on random instances of linear regression**, a model recently proposed for understanding of in-context learning [Garg et al., 2022, Akyürek et al., 2022, von Oswald et al., 2023, Ahn et al., 2023b, Zhang et al., 2023, Mahankali et al., 2023]. In particular, we follow the setting of Ahn et al. [2023b]; see Appendix B for precise details.

3.1 Linear transformers as a fruitful abstraction

Setting for the experiments. Having established the framework in Appendix B, we now describe details of our experiments. Our base-setup is the 3-layer linear transformer, with 5-dimensional covariates, i.e. $(L = 3, d = 5)$. This is the minimally complex setting that still recovers all of the discussed features of full transformers. Transformers with larger L or d are qualitatively similar to the $(L = 3, d = 5)$ setting, and we provide such an example in Figure 14.

$(d = 5)$	**Setting 1** [Ahn et al., 2023b]	**Setting 2** (fewer covariates)	**Setting 3** (heavy-tailed covariates)
#contexts n	20	5	20
distribution of $x^{(i)}$	$\mathcal{N}(0, I_d)$	$\mathcal{N}(0, I_d)$	$\sqrt{\Gamma_{0.1,10}} \cdot \mathrm{Unif}(\mathbb{S}^{d-1})$
distribution of $w_\star$	$\mathcal{N}(0, I_d)$	$\mathcal{N}(0, I_d)$	$\mathcal{N}(0, I_d)$

Table 1: Settings for (the right-side plots of) Figures 2, 3, 4, and 5

Our "default" setup is Setting 1 of Table 1, where the context consists of 20 context demonstrations; each context covariate is sampled from the standard Gaussian, i.e., $x^{(i)} \sim \mathcal{N}(0, I_d)$, and we draw $w_\star \sim \mathcal{N}(0, I_d)$. This is consistent with previous works [Garg et al., 2022, Akyürek et al., 2022, von Oswald et al., 2023, Ahn et al., 2023b]. In order to understand the effect of context length, we also consider the setting when context length $n = 5$ instead; this is Setting 2 of Table 1. Finally, to investigate the effect of heavy-tailed covariates on various aspects of the loss landscape, we consider Setting 3 in Table 1, where we draw each x_i instead uniformly from the unit sphere, and then scale it by the square root of a heavy-tailed Gamma random variable with shape parameter $k = 0.1$ and scale parameter $\theta = 10$. Furthermore, in Subsection C.1, we study the effect of heavy-tailedness of the covariates in more detail.

For each different setting, we pick the best learning rate from a grid search over 10 different choices. We choose the momentum parameter 0.9 for SGD, and $\beta_1 = \beta_2 = 0.9$ for Adam. We also employ the (global) gradient clipping where the thresholds are chosen to be 1 for all settings (i.e., the clipped gradient direction is the same as the non-clipped direction). All the experiments are run over 6 different random seeds. See Figures 2, 3, 4, and 5 for the results.

Discussion of results. Below we provide detailed discussion of the results.

[1]In fact, in their paper, they instead consider the maximum diagonal entry of the Hessian divided by the median diagonal entry as an approximation of this quantity.

1. **Gap between SGD and Adam.** In Figure 2 (right), we plot the training loss for the three settings in Table 1. Notice that we observe the phenomenon (Adam>SGD) over three different settings, to different extents. These loss behaviors resemble those of the practical transformer optimization (left plots of Figure 2).
2. **Heavy-tailed stochastic noise.** In Figure 3 (right), following [Zhang et al., 2020b, Kunstner et al., 2023], we plot the stochastic gradient noise at the initialization. Notice the similarity between the left plots and the right plots, showing that the shallow linear transformers also exhibit the heavy-tailed stochastic gradient noise phenomenon.
3. **Condition number of the landscape.** Following [Jiang et al., 2022], we measure the "robust" condition numbers of different optimizers along the trajectory. Figure 4 shows that the condition numbers of adaptive methods are lower than those of SGD, similar to [Jiang et al., 2022].
4. **Directional smoothness.** As observed by previous works [Zhang et al., 2020a,b, Pan and Li, 2023], in our experiments, we also observe that Adam has better directional smoothness than SGD, which correlates with the speed-up of Adam over SGD. We present this in Figure 5.

In this section, we have seen that simple linear transformers described in Appendix B suffice to recover all the main features identified in previous works (Section 2). In Appendix C, we take advantage of the concreteness and simplicity of our linear transformer to explore and understand the role of heavy-tailedness in data distribution and depth of the network. Our paper's conclusion can be found in Section 4.

4 Conclusion

The complexity of modern neural networks, especially transformers, often eludes precise mathematical understanding, and hence calls for such "physics-style" approaches (c.f. Zhang et al. [2022], Ahn et al. [2023a], Abernethy et al. [2023], Allen-Zhu and Li [2023], Li et al. [2023], Dai et al. [2023]) based on simplified models. This work presents a concrete addition to this viewpoint, and it builds a valuable, realistic proxy for understanding transformers. We hope that our work will serve as the stepping stone for building a more precise theory of transformer optimization, as well as contributing to the development of efficient training methods for transformers.

Acknowledgements

This work stems from a group project at MIT; we thank the collaborators in the group, Hadi Daneshmand, Haochuan Li, Zakaria Mhammedi, Swati Padmanabhan, Amirhossein Reisizadeh, and William Wang for their time and intriguing discussions.

Kwangjun Ahn and Ali Jadbabaie were supported by the ONR grant (N00014-23-1-2299) and MIT-IBM Watson as well as a Vannevar Bush fellowship from Office of the Secretary of Defense. Xiang Cheng and Suvrit Sra acknowledge support from NSF CCF-2112665 (TILOS AI Research Institute) and an NSF CAREER award (1846088). Minhak Song and Chulhee Yun were supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant (No. 2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)) funded by the Korea government (MSIT), two National Research Foundation of Korea (NRF) grants (No. NRF-2019R1A5A1028324, RS-2023-00211352) funded by the Korea government (MSIT), and a grant funded by Samsung Electronics Co., Ltd.

References

Jacob Abernethy, Alekh Agarwal, Teodor V Marinov, and Manfred K Warmuth. A mechanism for sample-efficient in-context learning for sparse retrieval tasks. *arXiv preprint arXiv:2305.17040*, 2023. 5

Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via the "edge of stability". *NeurIPS 2023 (arXiv:2212.07469)*, 2023a. 5

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *NeurIPS 2023 (arXiv:2306.00297)*, 2023b. 4, 9

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *International Conference on Learning Representations*, 2022. 4

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 1, context-free grammar. *arXiv preprint arXiv:2305.13673*, 2023. 5

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015. 2, 3

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 1

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signsgd. *Advances in Neural Information Processing Systems*, 35:9955–9968, 2022. 9

Yan Dai, Kwangjun Ahn, and Suvrit Sra. The crucial role of normalization in sharpness-aware minimization. *NeurIPS 2023 (arXiv:2305.15287)*, 2023. 5

J Devlin, MW Chang, K Lee, and K Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding in: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics, 4171–4186.. acl. *ACL. DOI: https://doi.org/10.18653/v1*, (19):1423, 2019. 1

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022. 4

Kaiqi Jiang, Dhruv Malik, and Yuanzhi Li. How does adaptive optimization impact local neural network geometry? *arXiv preprint arXiv:2211.02254*, 2022. 1, 2, 3, 4, 5, 8, 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. 1

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 International Conference on Learning Representations (ICLR) (arXiv:2304.13960)*, 2023. 1, 2, 3, 5, 8, 11

Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a mechanistic understanding. *International Conference on Machine Learning (ICML) (arXiv:2303.04245)*, 2023. 5

Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. In *2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 5747–5763. Association for Computational Linguistics (ACL), 2020. 1

Arvind Mahankali, Tatsunori B Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *arXiv preprint arXiv:2307.03576*, 2023. 4, 9

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018. 2, 3

Yan Pan and Yuanzhi Li. Toward understanding why adam converges faster than sgd for transformers. *arXiv preprint arXiv:2306.00204*, 2023. 1, 2, 3, 4, 5, 8, 9

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017. 1, 10

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023. 4, 9, 10

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pages 4148–4158, 2017. 2

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations (ICLR)*, 2020a. 1, 3, 5, 9

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*, 33:15383–15393, 2020b. 1, 2, 3, 5, 8, 11

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023. 4, 9

Yi Zhang, Arturs Backurs, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, and Tal Wagner. Unveiling transformers with lego: a synthetic reasoning task. *arXiv preprint arXiv:2206.04301*, 2022. 5

A Details on the features of the transformer optimization

A.1 Heavy-tailed gradient noise [Zhang et al., 2020b, Kunstner et al., 2023]

In [Zhang et al., 2020b] (entitled *Why are adaptive methods good for attention models?*), it was observed that the stochastic gradient is typically more heavy-tailed for transformer optimization than other neural network optimization. In particular, they make a case that this is opposed to the standard bounded variance condition for SGD analysis – see Figure 3 and Figure 6. They posit that this phenomenon might be one of the main reasons behind the phenomenon (Adam>SGD); they also theoretically show that adaptive step sizes in the form of gradient clipping is required for convergence.

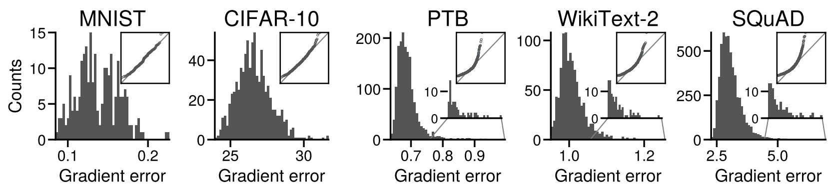


Figure 6: **The heavy-tail stochastic gradient noise for transformers.** Under the same setting as Figure 1, Kunstner et al. [2023] plot the stochastic gradient noise at the initialization. Notice that the stochastic gradient noise for the convolutional neural networks on vision tasks (MNIST, CIFAR-10) is much less heavy-tailed than the transformers on NLP tasks. We will revisit this plot in Figure 9 with shallow linear transformers.

A noteworthy follow-up work by Kunstner et al. [2023] reveal that the heavy-tailed stochastic noise may not explain the full picture. In particular, they compare the full-batch versions (hence no stochastic noise), and notice the phenomenon (Adam>SGD) still hold. Since there is no stochastic noise in this setting, the explanation based on heavy-tailed noise does not apply here.

A.2 Ill-conditioned landscape [Jiang et al., 2022]

In another inspiring work by Jiang et al. [2022], authors seek to understand the difficulty of transformer optimization through the lens of condition number. In particular, they consider a "robust" version of condition number defined as $R^{\mathsf{OPT}}_{\mathsf{med}} := {}^{\lambda_{\max}(\nabla^2 f)}/_{\lambda_{\text{median}}(\nabla^2 f)}$[2], and here the reason for $\lambda_{\mathsf{median}}$ instead of $\lambda_{\min}$ is to handle the case where the Hessian is degenerate. They observe that during transformer optimization, non-adaptive optimizers like SGD tend to have larger robust condition number than adaptive optimizers like Adam; they posit that this phenomenon is one of the main reasons for (Adam>SGD) – see Figure 4. Jiang et al. [2022] also report that this gap is not there when training convolutional neural networks on image classification tasks, and suggest that this phenomenon may be rooted in unique features of the transformer which are missing in other popular neural networks.

A.3 Directional Smoothness [Pan and Li, 2023]

In a follow up work by Pan and Li [2023] (entitled *Toward understanding why Adam converges faster than SGD for transformers*), the authors again corroborate (Adam>SGD). In addition, they further observe in [Pan and Li, 2023, Figure 6] that proper gradient clipping techniques further accelerate optimization. In order to understand this phenomenon, they propose an explanation based on "directional smoothnesss" along the iterates x_t. More formally, they consider the following Taylor expansion along the iterates:

$$f(x_{t+1}) - f(x_t) = \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{1}{2}(x_{t+1} - x_t)^\top \nabla^2 f(x_t)(x_{t+1} - x_t) + O(\eta^3)\,,$$

and define the directional smoothness as ${}^{(x_{t+1}-x_t)^\top \nabla^2 f(x_t)(x_{t+1}-x_t)}/_{\|x_{t+1}-x_t\|^2}$. In particular, based on the above calculations, one can infer that smaller directional smoothness implies better optimization as $f(x_{t+1}) - f(x_t)$ becomes smaller. They claim that the directional smoothness holds the key to understanding (Adam>SGD) (as well as transformer optimization in general). They also verify that adaptive optimizers tend to have smaller directional smoothness values, and employing gradient

[2]In fact, in their paper, they instead consider the maximum diagonal entry of the Hessian divided by the median diagonal entry as an approximation of this quantity.

clipping further reduces the directional smoothness. Once again, Pan and Li [2023] hypothesize that this feature is unique to transformers, as they observe that adaptive algorithms can demonstrate *worse directional smoothness* than SGD for, e.g., ResNet training.

A.4 Generalized smoothness [Zhang et al., 2020a]

We discuss one more noteworthy work [Zhang et al., 2020a] that identifies another unconventional feature. Here we highlight that the main motivation of [Zhang et al., 2020a] was not about understanding (Adam>SGD), and they also observe their proposed feature in some other non-transformer neural networks such as ResNet. The main observation made by [Zhang et al., 2020a] is that the standard smoothness assumption is not suitable for neural network training. Instead, they observed that the spectral norm of Hessian typically grows with the norm of gradient at the current iterate. Based on this observation, the authors define the following notion of generalized smoothness:

Definition 1. *We say f is (L_0, L_1)-smooth if $\left\|\nabla^2 f(x)\right\| \leq L_0 + L_1 \left\|\nabla f(x)\right\|$. When $L_1 = 0$, this condition recovers the standard smoothness condition.*

A coordinate-wise version of Definition 1 was considered in [Crawshaw et al., 2022]. Under Definition 1, they demonstrate that non-adaptive SGD needs more iterations to converge than an adaptive method based on the global clipping of gradients.

B Details of linear transformer on linear regression

Data distribution. The data distribution can be thought of as the random instances of linear regression. Concretely, let $X \in \mathbb{R}^{(n+1)\times d}$ be the matrix of covariates of the regression whose row i contains tokens $x^{(i)} \in \mathbb{R}^d$ drawn i.i.d. from a distribution $D_{\mathcal{X}}$. We then draw $w_\star \sim D_{\mathcal{W}}$ and then generate the scalar responses $y = [\langle x^{(1)}, w_\star\rangle, \ldots, \langle x^{(n)}, w_\star\rangle] \in \mathbb{R}^n$. Now the input of the data set consists of these linear regression examples:

$$\text{Input matrix: } Z_0 = \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(n)} & x^{(n+1)} \\ y^{(1)} & y^{(2)} & \cdots & y^{(n)} & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(n+1)}.$$

The goal is to predict the missing $y^{(n+1)}$, as we detail below.

Optimization objective. Let $\mathsf{TF}_L(\cdot; W) : \mathbb{R}^{(n+1)\times(d+1)} \to \mathbb{R}$ denote the prediction of the linear transformer with parameters W. Our optimization objective is given by

$$f(W) := \mathbb{E}_{(Z_0, w_\star)}\left[\left(\mathsf{TF}_L(Z_0; W) - w_\star^\top x^{(n+1)}\right)^2\right].$$

In words, we train the linear transformer to predict $y^{(n+1)}$ using $\mathsf{TF}_L(Z_0; W)$; we will formally define the linear transformer architecture below. This objective was the center of study in a number of recent empirical and theoretical works on understanding transformers [von Oswald et al., 2023, Ahn et al., 2023b, Zhang et al., 2023, Mahankali et al., 2023].

Linear transformer (self-attention) architecture. We will now present the neural network architecture that will be used throughout this paper. Given matrices $P, Q \in \mathbb{R}^{(d+1)\times(d+1)}$, we define the **linear self-attention** architecture as

$$\mathsf{Attn}_{P,Q}(Z) = PZM(Z^\top QZ) \quad \text{where } M := \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}. \tag{1}$$

Finally, for a positive integer L, we define an **L-layer linear transformer** TF_L as a stack of L linear attention units. Specifically, let the output of the L^{th} layer attention, Z_L, be recursively defined as

$$Z_{\ell+1} = Z_\ell + \frac{1}{n}\mathsf{Attn}_{P_\ell, Q_\ell}(Z_\ell) \quad \text{for } \ell = 0, 1, \ldots, L-1.$$

Then we define $\mathsf{TF}_L(Z_0; \{P_\ell, Q_\ell\}_{\ell=0}^{L-1}) = -[Z_L]_{(d+1),(n+1)}$, i.e., the $(d+1, n+1)$-th entry of Z_ℓ. The reason for the minus sign is to be consistent with [von Oswald et al., 2023, Ahn et al., 2023b], where such a choice was motivated by theoretical considerations.

We emphasize here that the linear attention unit, defined in (1), differs from the standard attention unit in [Vaswani et al., 2017] in two ways: we use a single matrix Q to represent the product of key, query matrices, and more importantly, *we remove the softmax activation outside* $Z^\top QZ$. There are two key reasons for our choice:

1. The linear attention unit is *much better suited to the task of linear regression.* For instance, [von Oswald et al., 2023, Appendix A.9] demonstrates that the performance of softmax transformer with twice many heads matches that of linear transformers; in other words, we need two softmax attention heads to recover the performance of a single linear head. In Figure 7, we show that linear attention performs significantly better than standard attention with softmax.
2. Our goal in this paper is to *find the simplest abstraction* which is representative of the transformer's optimization landscape. As we will see in Subsection 3.1, the loss landscape of the linear transformer well approximates that of the actual transformer, even without the softmax activation.

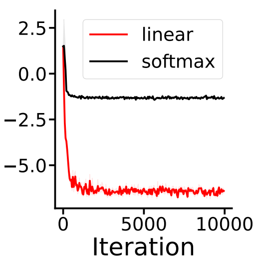


Figure 7: $\log(\text{loss})$ against iteration. Comparison between linear attention and softmax attention for the 3-layer transformers. Note that the loss of linear transformer decreases much faster.

C Understanding features based on linear transformers

The main advantage of our toy linear transformer comes from its simplicity and concreteness. In particular, thanks to the concreteness of the setting, one can conduct various "controlled" experiments to understand the features observed in Subsection 3.1. Recall that the data set used in our experiments consists of nothing but random linear regression instances. This data set is far simpler and more concrete than the language modeling data sets (e.g., Wikipedia texts, question&answering) of the previous works discussed in Section 2.

We first take advantage of the concreteness of our data distribution, and look deeper into how the main distinctive features of transformer optimization arise. We first investigate how the "heavy-tailedness" of the data distribution affects the extent of the features from Section 2.

Spherical $x^{(i)}$'s **Heavy-tailed $x^{(i)}$'s**

1. Comparing SGD v.s Adam:

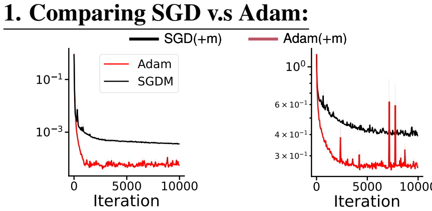


Figure 8: Plot of log(loss) against iteration for SGD and Adam.

2. Stochastic gradient noise:

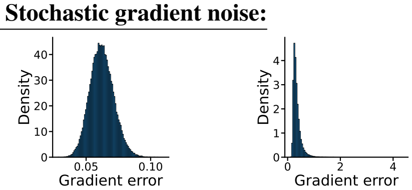


Figure 9: Comparing distribution of stochastic gradient noise at Epoch 0.

3. Robust condition number:

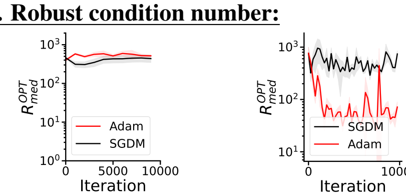


Figure 10: Comparing the robust condition number from Jiang et al. [2022]

C.1 Effect of data distribution

Given that we observe the "heavy-tailedness" of stochastic gradient noise, perhaps a natural question to ask is the following:

Q. *Does the "heavy-tailedness" of data distribution exacerbate the features in Section 2?*

Settings. In order to investigate the above question, we consider the following distributions for the covariates $x^{(i)}$'s of linear regression:

- **Spherical covariates.** We sample $x^{(i)}$'s uniformly at random from the unit sphere $\mathbb{S}^{d-1}$.
- **Heavy-tailed covariates.** We first sample $x^{(i)}$'s uniformly at random from the unit sphere $\mathbb{S}^{d-1}$, and then multiply each covariate by a random scale drawn *i.i.d* from a heavy-tailed distribution, specifically the square root of a Gamma random variable from $\Gamma_{k,\theta}$. Note that $k = 2.5$ and $\theta = 2$

precisely corresponds to the case where $x^{(i)} \sim \mathcal{N}(0, I_5)$. In our experiments, we use $k = 0.1$ and $\theta = 10$ to make the distribution more heavy-tailed, while keeping the variance the same.

Discussion. We now discuss the experimental results presented in Figures 8, 9, and 10:

- In Figure 9, we see that "heavy-tailed"-ness of covariates is reflected in the "heavy-tailed"-ness of the stochastic gradient. Notably, the contrast between the two plots in Figure 9 reminds us of the contrast we see between CNNs and transformers in Figure 6.
- In Figure 10, it appears that there is some correlation between the gap in robust condition number, and the "heavy-tailed"-ness of the data distribution, with heavier tails leading to larger gaps.
- Finally, Figure 8 shows how the optimization speed of SGD and Adam vary with the heavy-tailedness of covariates. First, given spherical (light-tailed) covariates, both SGD and Adam converge much faster than Gamma-scaled covariates (heavy-tailed). On the other hand, the *relative gap* between the speed of Adam and SGD does not seem to improve noticeably under light-tailed noise.
- Together, Figure 8 and Figure 9 show that the relationship between heavy-tailed gradient noise and optimization speed may be a little more complicated than suggested in [Zhang et al., 2020b]. Specifically, adaptivity seems to be equally beneficial regardless of the heavy-tailedness of the gradient noise. Instead, these two plots seem to align more with the message in [Kunstner et al., 2023] – that noise may not be the sole contributor of (Adam>SGD).

We next take advantage of the concreteness of our model, and investigate the effect of the number of layers on the optimization.

1. Loss against time

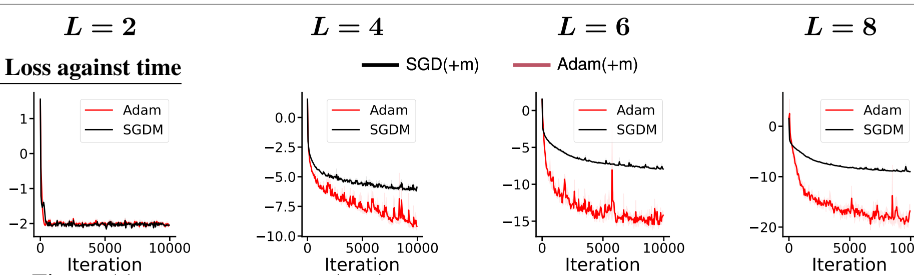


Figure 11: Comparison of $\log(\text{loss})$ between SGD and Adam for different number of layers.

2. Stochastic gradient noise:

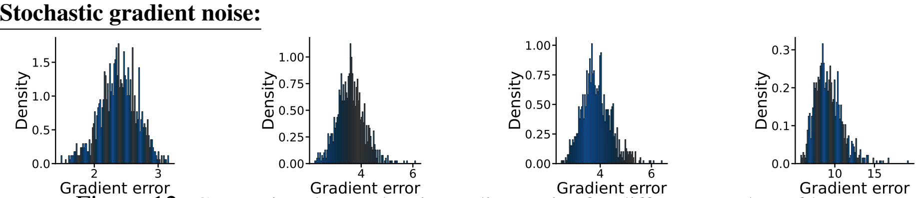


Figure 12: Comparing the stochastic gradient noise for different number of layers.

3. Robust condition number:

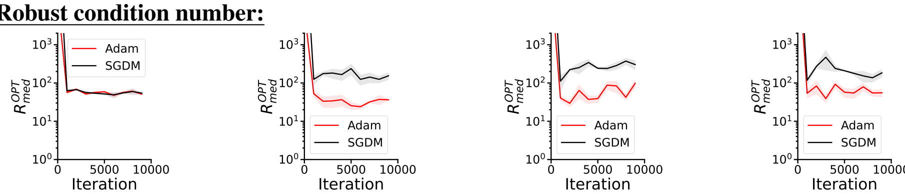


Figure 13: Comparing the robust condition number for different number of layers.

C.2 Effect of more layers

We investigate the effect of the number of layers L on the optimization. Specifically,

Q. Will a deeper linear transformer exacerbate the features in Section 2?

Settings. In order to investigate the above question, we consider repeating the experiments in Subsection 3.1 for the number of layers $L \in \{2, 4, 6, 8\}$.

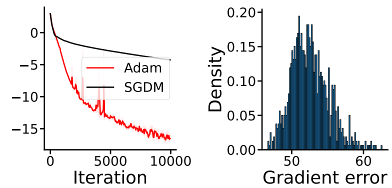


Figure 14: Plots for 8-layer linear transformer with covariate dimension $d = 20$ and context length $n = 60$. Left: log(loss) against iterations. Right: histogram of stochastic gradient noise at Epoch 0.

Discussion. We present the experimental results presented in Figures 11, 12, and 13.

- As one can see from Figure 11, the gap in loss between adaptive methods and SGD become more and more pronounced as we increase the number of layers.
- On the other hand, the absolute value of the loss decreases with increasing depth, for both SGD and Adam, which makes sense considering the larger capacity of deeper models.
- We plot the stochastic gradient noise for different settings in Figure 12. We do see that the noise for the case of $L = 6, 8$ are more heavy-tailed than the case of $L = 2, 4$. In particular, the noise distribution for $L = 2$ is much less heavy-tailed than that for $L = 8$.
- Lastly, we observe in Figure 13 that the gap in the robust condition number of SGD and Adam is more pronounced in deeper models ($L = 4, 6, 8$) than the shallow model ($L = 2$).