In-Context Learning Is Not Gradient Descent—Unless You Initialize Transformer Right

Shifeng Xie
Telecom Paris
Institut Polytechnique de Paris
France

Rui Yuan
Lexsi Labs, Paris
France

Simone Rossi
EURECOM
France

Thomas Hannagan
Stellantis
France

Abstract

We revisit the link between linear self-attention (LSA) and gradient descent (GD) beyond idealized assumptions, analyzing multi-head LSA under non-zero Gaussian prior means in linear-regression ICL. We introduce an initial-guess term in the embedding and prove an upper bound on the number of heads needed; experiments corroborate this and reveal a persistent gap to one-step GD. To close the gap, we propose y_q-LSA—a single-head LSA with a trainable initial guess—supported by theory and experiments showing alignment with one-step GD. Motivated by these insights, we further demonstrate that providing initial guesses improves ICL performance of general LLMs on a semantic similarity task.

1 Introduction

Large language models (LLMs) exhibit a remarkable phenomenon known as in-context learning (ICL)[Brown et al., 2020, Dong et al., 2024]. A prominent line of research interprets ICL as implicitly performing gradient descent (GD) within the forward pass of linear self-attention (LSA), using linear regression as a framework [Garg et al., 2022, Von Oswald et al., 2023]. However, this connection has primarily been established under simplified conditions with zero-mean Gaussian priors for the linear regression tasks and zero initialization for GD. Recent findings suggest that this equivalence breaks down under more realistic conditions. A more detailed discussion is provided in the Appendix A.

To address this challenge, we adopt the ICL linear regression framework [Von Oswald et al., 2023] and explicitly incorporate non-zero prior means into the linear regression setting. Our investigation systematically analyzes the behavior of multi-head LSA under three key variations: the number of attention heads, the deviation of the prior mean from zero, and the initialization of the query's prediction, which we term the initial guess y_q. Our observations reveal a crucial insight: when prior means are non-zero, increasing the number of heads alone cannot eliminate the performance gap between multi-head LSA and one-step GD. Motivated by these findings, we propose y_q-LSA, an architectural extension designed to neutralize the influence of both non-zero prior means and initial guesses. This approach enables even a single-head LSA to match the performance of one-step GD. Finally, inspired by the role of initial guess y_q, we show that carefully introducing a flexible initial guess consistently improves LLM ICL performance.

2 Findings

We summarize our theoretical results under the linear–regression ICL model in §B. Let $d \in \mathbb{N}$ be the input dimension and $C \in \mathbb{N}$ the number of context examples. Denote context data by $(\mathbf{X}, \mathbf{y}) \in \mathbb{R}^{C\times d} \times \mathbb{R}^C$ with rows $(\mathbf{x}_i^\top, y_i)$, a query by $\mathbf{x}_q \in \mathbb{R}^d$, and its (possibly nonzero) initial guess by $y_q \in \mathbb{R}$; the embedding matrix is $\mathbf{E}$ in (2). LSA are given by (3) and (6), respectively. The

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: What Can('t) Transformers Do?.

ICL risk is $\mathcal{R}(f) \overset{\text{def}}{=} \mathbb{E}\big[(f(\mathbf{E}) - y)^2\big]$ in (4). Tasks are generated by a prior $\widehat{\mathbf{w}} \sim \mathcal{N}(\mathbf{w}_\star, \mathbf{I}_d)$ (prior mean $\mathbf{w}_\star \in \mathbb{R}^d$), features $\mathbf{x}_i, \mathbf{x}_q \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$, and labels $y_i = \mathbf{x}_i^\top \widehat{\mathbf{w}}$, $y = \mathbf{x}_q^\top \widehat{\mathbf{w}}$. We use the flexible embedding

$$\mathbf{E}_{\mathbf{w}} \overset{\text{def}}{=} \begin{bmatrix} \mathbf{X}^\top & \mathbf{x}_q \\ \mathbf{y}^\top & \mathbf{w}^\top \mathbf{x}_q \end{bmatrix} \in \mathbb{R}^{(d+1)\times(C+1)},$$

and define the y_q-*LSA* predictor by $f_{y_q\text{-LSA}}(\mathbf{X}, \mathbf{y}, \mathbf{x}_q) \overset{\text{def}}{=} f_{\text{LSA}}(\mathbf{E}_{\mathbf{w}})$, where $\mathbf{w} \in \mathbb{R}^d$ is trainable.

Proofs and experimental validations of following findings are provided in the Appendices C and E.

Theorem 1 (Heads sufficiency). Let $H \in \mathbb{N}$ be the number of attention heads in (6). Then, for linear regression ICL, $\inf_{f \in \mathcal{F}_{(d+1)\text{-LSA}}} \mathcal{R}(f) = \inf_{f \in \mathcal{F}_{(d+2)\text{-LSA}}} \mathcal{R}(f)$. In other words, once the number of heads reaches $d+1$ (i.e., one more than the input dimension), the optimal ICL risk cannot be further reduced by adding additional heads.

Theorem 2 (No stationary point for LSA in nonzero prior). Consider any multi–head LSA $f_{\text{H-LSA}}$ with context size $C \to \infty$. If the prior mean is nonzero, $\mathbf{w}_\star \neq 0$, then $\nabla \mathcal{R}(f_{\text{H-LSA}}) \neq \mathbf{0}$. For all parameter settings, the ICL risk admits no stationary points in the infinite–context limit.

Lemma 1 (Nonconvexity of the LSA risk). The mapping from the multi–head LSA parameters to the ICL risk $\mathcal{R}(f_{\text{H-LSA}})$ is nonconvex.

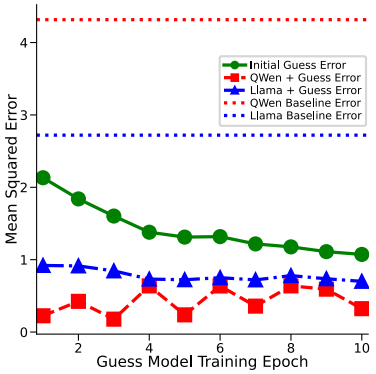
Theorem 3 (One-step GD via y_q-LSA). For any prior mean, $f_{y_q\text{-LSA}}$ realizes the one-step GD update.

Theorem 4 (Stationary point at the prior for y_q-LSA). Suppose the prior mean satisfies $\mathbf{w}_\star \neq 0$ and the context size $C \to \infty$. For y_q-LSA with parameters as in (9) and with $\mathbf{w} = \mathbf{w}_\star$, the gradient of the ICL risk vanishes: $\nabla \mathcal{R}\big(f_{y_q\text{-LSA}}\big) = \mathbf{0}$, i.e., the parameters constitutes a stationary point of $\mathcal{R}(f_{y_q\text{-LSA}})$.

Lemma 2 (Nonconvexity of the y_q-LSA risk). The ICL risk $\mathcal{R}(f_{y_q\text{-LSA}})$ is nonconvex.

Lemma 3 (y_q-LSA as a special case of the Linear Transformer Block). There exist linear maps such that the Linear Transformer Block (LSA followed by a linear MLP), f_{LTB}, satisfies $f_{y_q\text{-LSA}}(\mathbf{X}, \mathbf{y}, \mathbf{x}_q) = f_{\text{LTB}}(E)$ on the standard embedding after an appropriate reparameterization; thus $f_{y_q\text{-LSA}} \in \mathcal{F}_{\text{LTB}}$.

3 LLM experiments



Through theoretical analysis, we hypothesize that supplying an initial guess for the target in ICL serves as an optimization prior, enhancing prediction accuracy. Experiments with LLMs confirm the efficacy of this approach.

We evaluate Meta-LLaMA-3.1-8B-Instruct [Grattafiori et al., 2024], Qwen/Qwen2.5-7B-Instruct [Yang et al., 2024, Team, 2024], and the STS-Benchmark dataset (English subset) [May, 2021]. Each prompt includes 10 labeled examples and a query, with initial guesses generated by a lightweight guess model and provided as priors. The LLM predicts refined similarity scores, and performance is measured by mean squared error with and without initial guesses (see Appendix H.3). Results show that incorporating initial guesses consistently reduces MSE across conditions and improves ICL performance in both LLaMA and QWen, supporting our hypothesis that initial guesses act as priors guiding refinement.

4 Conclusion

We showed that multi-head LSA fails to match GD under nonzero priors, while our proposed y_q-LSA overcomes this limitation by incorporating a trainable initial guess. Experiments on linear regression and semantic similarity confirm that initial guesses consistently enhance ICL performance.

References

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=LziniAXEI9. (Cited on pages 5, 6, 7, and 11.)

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf. (Cited on pages 1 and 4.)

Joy Crosbie and Ekaterina Shutova. Induction heads as an essential mechanism for pattern matching in in-context learning, 2024. URL https://arxiv.org/abs/2407.07011. (Cited on page 9.)

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.64. URL https://aclanthology.org/2024.emnlp-main.64/. (Cited on page 1.)

Fabian Falck, Ziyu Wang, and Christopher C. Holmes. Is In-Context Learning in Large Language Models Bayesian? A Martingale Perspective. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=b1YQ5WKY3w. (Cited on page 4.)

Deqing Fu, Tian qi Chen, Robin Jia, and Vatsal Sharan. Transformers Learn to Achieve Second-Order Convergence Rates for In-Context Linear Regression. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=L8h6cozcbn. (Cited on page 5.)

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=flNZJ2eOet. (Cited on page 1.)

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, and al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783. (Cited on pages 2 and 23.)

Ruomin Huang and Rong Ge. Task descriptors help transformers learn linear models in-context. In *ICML 2024 Workshop on In-Context Learning*, 2024. URL https://openreview.net/forum?id=4SfCI1DJhr. (Cited on page 7.)

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. (Cited on page 23.)

Arvind V. Mahankali, Tatsunori Hashimoto, and Tengyu Ma. One Step of Gradient Descent is Provably the Optimal In-Context Learner with One Layer of Linear Self-Attention. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8p3fu56lKc. (Cited on pages 5 and 7.)

Sadegh Mahdavi, Renjie Liao, and Christos Thrampoulidis. Revisiting the equivalence of in-context learning and gradient descent: The impact of data distribution. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7410–7414. IEEE, 2024. (Cited on page 5.)

Philip May. Machine translated multilingual sts benchmark dataset., 2021. URL https://github.com/PhilipMay/stsb-multi-mt. (Cited on pages 2 and 22.)

Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-Context Learning through the Bayesian Prism. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=HX5ujdsSon. (Cited on page 4.)

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020. URL https://arxiv.org/abs/2004.09813. (Cited on page 23.)

Simone Rossi, Rui Yuan, and Thomas Hannagan. Understanding in-context learning in transformers. In *The Third Blogpost Track at ICLR 2024*, 2024. URL https://openreview.net/forum?id=lpyfAihk28. (Cited on pages 5 and 7.)

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/. (Cited on pages 2 and 23.)

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html. (Cited on pages 1, 5, 6, 7, and 15.)

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An Explanation of In-context Learning as Implicit Bayesian Inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI. (Cited on page 4.)

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. (Cited on pages 2 and 23.)

Naimeng Ye, Hanming Yang, Andrew Siah, and Hongseok Namkoong. Pre-training and In-context Learning IS Bayesian Inference a la De Finetti. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024. URL https://openreview.net/forum?id=ttupfosvgx. (Cited on page 4.)

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024a. URL http://jmlr.org/papers/v25/23-1042.html. (Cited on pages 5 and 7.)

Ruiqi Zhang, Jingfeng Wu, and Peter Bartlett. In-context learning of a linear transformer block: Benefits of the MLP component and one-step GD initialization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=Thou1rKdpZ. (Cited on pages 5, 8, and 16.)

A Related Work

Theoretical studies on ICL have analyzed its mechanisms to undestand how LLMs effectively learn from contextual examples [Brown et al., 2020]. ICL can be framed as an implicit Bayesian process where the model performs posterior inference over a latent task structure based on contextual examples, performing a form of posterior updating [Xie et al., 2022, Falck et al., 2024, Panwar et al., 2024, Ye et al., 2024]. Alternatively, a more recent perspective suggests that ICL in transformers

is akin to gradient-based optimization occurring within their forward pass. Von Oswald et al. [2023] demonstrate that self-attention layers can approximate gradient descent by constructing task-specific updates to token representations. They provide a mechanistic explanation by showing how optimized transformers can implement gradient descent dynamics with a given learning rate [Rossi et al., 2024]. While this work provides a new perspective on ICL, it limits the analysis to simple regression tasks and it simplifies the transformer architecture by considering a single-head self-attention layer without applying the $\mathsf{sfmx}(\cdot)$ function on the attention weights (also known as linear attention). Ahn et al. [2023] extend the work of Von Oswald et al. [2023] by showing how the in-context dynamics can learn to implement *preconditioned* gradient descent, where the preconditioner is implicitly optimized during pretraining. More recently, Mahankali et al. [2024] prove that a single self-attention layer converges to the *global* minimum of the squared error loss. Zhang et al. [2024b] also analyze a more complex transformer architecture with a (linear) multi-layer perceptron (MLP) after the linear self-attention layer, showing the importance of such block when pretraining for more complex regression tasks (i.e., non-zero bias). Finally, evidence also suggests that multi-layer transformers can perform gradient descent in a more complex manner, akin to the iterative Newton's method [Fu et al., 2024].

In this work, we extend the above lines of research by emphasizing more realistic priors, specifically, non-zero prior means. While Zhang et al. [2024a], Mahdavi et al. [2024] explore broader prior distributions by analyzing covariate structures or modify the distribution of input feature, our focus instead lies on the interplay between a non-zero prior mean and the capacity of LSA to emulate GD. We note that while Ahn et al. [2023], Mahankali et al. [2024], Zhang et al. [2024b] provide compelling theoretical analyses, their work does not include experimental validations. In doing so, our study builds upon and generalizes the prior-zero analyses found in Von Oswald et al. [2023], Ahn et al. [2023], illuminating new challenges and insights that arise when priors deviate from zero, both theoretically and empirically.

B Preliminaries

We use $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$ to denote a feature vector and its label, respectively. We consider a fixed number of context examples, denoted by $C > 0$. We denote the context examples as $(\mathbf{X}, \mathbf{y}) \in \mathbb{R}^{C\times d} \times \mathbb{R}^C$, where each row represents a context example, denoted by $(\mathbf{x}_i^\top, y_i)$, $i \in [C]$. That is,

$$\mathbf{X} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_C^\top \end{bmatrix} \in \mathbb{R}^{C\times d} \quad \text{and} \quad \mathbf{y} \stackrel{\text{def}}{=} \begin{bmatrix} y_1 \\ \vdots \\ y_C \end{bmatrix} \in \mathbb{R}^C. \tag{1}$$

To formalize an in-context learning (ICL) problem, the input of a model is an *embedding matrix* given by

$$\mathbf{E} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{X}^\top & \mathbf{x}_q \\ \mathbf{y}^\top & y_q \end{bmatrix} \in \mathbb{R}^{(d+1)\times(C+1)}, \tag{2}$$

where $\mathbf{x}_q \in \mathbb{R}^d$ is a new query input and $y_q \in \mathbb{R}$ is an *initial guess* of the prediction for the query $\mathbf{x}_q$. The model's output corresponds to a prediction of $y \in \mathbb{R}$. Notice that the embedding matrix in (2) is a slight extension to the commonly used embedding matrix, e.g. presented in Von Oswald et al. [2023], where y_q is set to be zero by default. Its interpretation will be clearer in the next two sections.

Linear regression tasks. We formalize the linear regression tasks as follows. Assume that $(\mathbf{X}, \mathbf{y}, \mathbf{x}_q, y)$ are generated by:

- First, a task parameter is independently generated by $\widehat{\mathbf{w}} \sim \mathcal{N}(\mathbf{w}_\star, \mathbf{I}_d)$, where $\mathcal{N}(\mathbf{w}_\star, \mathbf{I}_d)$ is the *prior*, and $\mathbf{w}_\star$ is called the *prior mean*.
- The feature vectors are independently generated by $\mathbf{x}_q, \mathbf{x}_1, \ldots \mathbf{x}_C \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$.
- Then, the labels are generated by $y = \langle \widehat{\mathbf{w}}, \mathbf{x}_q \rangle$, and $y_i = \langle \widehat{\mathbf{w}}, \mathbf{x}_i \rangle$, $i \in [C]$, with no noise.

Here, $\mathbf{w}_\star \in \mathbb{R}^d$ is fixed but unknown and governs the data distribution.

A linear self-attention. We consider a *linear self-attention* (LSA) defined as

$$\begin{aligned} &f_{\mathsf{LSA}} : \mathbb{R}^{(d+1)\times(C+1)} \to \mathbb{R}, \\ &\mathbf{E} \mapsto \left[\mathbf{E} + \tfrac{1}{C}\mathbf{W}^P\mathbf{W}^V\mathbf{E}\mathbf{W}^M(\mathbf{E}^\top(\mathbf{W}^K)^\top\mathbf{W}^Q\mathbf{E})\right]_{-1,-1}, \end{aligned} \tag{3}$$

where $\mathbf{W}^K, \mathbf{W}^Q, \mathbf{W}^P, \mathbf{W}^V$ are trainable parameters, $[\,\cdot\,]_{-1,-1}$ refers to the bottom right entry of a matrix, and $\mathbf{W}^M \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{I}_C & 0 \\ 0 & 0 \end{bmatrix}$ is a mask matrix.

ICL risk. We measure the ICL risk of a model f by the mean squared error,

$$\mathcal{R}(f) \stackrel{\text{def}}{=} \mathbb{E}[(f(\mathbf{E}) - y)^2], \tag{4}$$

where the input $\mathbf{E}$ is defined in (2) and the expectation is over $\mathbf{E}$ (equivalent to over $\mathbf{X}$, $\mathbf{y}$, and $\mathbf{x}_q$) and y. The performance of different models are characterized by the ICL risk.

C Multi-Head Linear Self-Attention

In order to improve the performance of LSA, we consider the multi-head LSA. Let H be the number of heads. Similar to (3), we define the output of each transformer head as

$$\text{head}_h(\mathbf{E}) \stackrel{\text{def}}{=} \tfrac{1}{C}\mathbf{W}_h^P \mathbf{W}_h^V \mathbf{E}\mathbf{W}^M \left(\mathbf{E}^\top (\mathbf{W}_h^K)^\top \mathbf{W}_h^Q \mathbf{E}\right), \quad h \in [H], \tag{5}$$

where $\mathbf{W}_h^K, \mathbf{W}_h^Q, \mathbf{W}_h^P$ and $\mathbf{W}_h^V$ are trainable parameters for all $h \in [H]$. The multi-head LSA is defined as

$$f_{\mathsf{H-LSA}}(\mathbf{E}) \stackrel{\text{def}}{=} \big[\mathbf{E} + \sum\nolimits_{h=1}^{H} \text{head}_h(\mathbf{E})\big]_{-1,-1}. \tag{6}$$

Notice that we extend the (multi-head) LSA by introducing an initial guess y_q in the embedding (2). The predictions derived from both f_{LSA} in (3) and $f_{\mathsf{H-LSA}}$ in (6) share a common characteristic: the bottom right entry of the output matrix represents y_q augmented by the attention map's output, regardless of the number of heads. We term y_q an *initial guess* precisely because the prediction format can be conceptualized as an initial prediction y_q for the query x_q, subsequently refined by an attention map-derived update. When defaulting to $y_q = 0$ in linear regression with zero prior mean, i.e., $\mathbf{w}_\star = 0$, as considered in Von Oswald et al. [2023], a non-trivial prior knowledge is accidentally introduced for the initial estimation. In Appendix E.1.3, we empirically investigate the impact of y_q for multi-head LSA with zero prior mean through a comprehensive ablation study.

We denote the hypothesis class formed by multi-head LSA models as

$\mathcal{F}_{H-\mathsf{LSA}} = \big\{f_{\mathsf{H-LSA}} : \{\mathbf{W}_h^K, \mathbf{W}_h^Q, \mathbf{W}_h^V, \mathbf{W}_h^P\}_{h=1}^H\big\}$, where $f_{\mathsf{H-LSA}}$ is defined in (6). Our first theoretical result is about the optimal solution of the ICL risk in the setup of multi-head LSA.

Theorem 1. *Consider two multi-head LSA functions $f_{(\mathsf{d+1})-\mathsf{LSA}}$ and $f_{(\mathsf{d+2})-\mathsf{LSA}}$ with the number of heads $H = d+1$ and $d+2$, respectively. Then we have*

$$\inf_{f \in \mathcal{F}_{f_{(\mathsf{d+1})-\mathsf{LSA}}}} \mathcal{R}(f) = \inf_{f \in \mathcal{F}_{f_{(\mathsf{d+2})-\mathsf{LSA}}}} \mathcal{R}(f). \tag{7}$$

Theorem 1 establishes that, for multi-head LSA in the linear regression setting, having $d+1$ attention heads achieves the same infimum of ICL risk as having $d+2$ heads. Extending this reasoning, one can show that any additional increase in the number of heads (e.g., $d+3$, $d+4$, etc.) does not yield a strictly lower ICL risk. In other words, augmenting a multi-head LSA with more than $d+1$ heads does not improve the theoretical minimum ICL risk on linear regression tasks. In Appendix E, we verify that Theorem 1 holds in practice, for different combinations of configuration.

Next, we explore the convergence of multi-head LSA. Inspired by the analysis of Ahn et al. [2023], we analyze the stationary point of the ICL risk for multi-head LSA functions.

Theorem 2. *Consider a general multi-head LSA function $f_{\mathsf{H-LSA}}$ with context size $C \to \infty$, then there will be no stationary point for the corresponded ICL risk. That is,*

$$\nabla \mathcal{R}\left(f_{\mathsf{H-LSA}}\right) \neq 0, \text{ for all the parameters } \mathbf{W}_h^K, \mathbf{W}_h^Q, \mathbf{W}_h^P \text{ and } \mathbf{W}_h^V \text{ with } h \in [H], \tag{8}$$

except for the case when the prior mean $\mathbf{w}_\star$ is equal to zero.

Theorem 2 states that when the context size $C \to \infty$, the gradient of the multi-head LSA's ICL risk $\mathcal{R}(f_{\mathsf{H-LSA}})$ remains non-zero for the entire parameters space as long as $\mathbf{w}_\star \neq 0$. This result highlights a fundamental limitation of multi-head LSA under non-zero priors: no choice of weights $\mathbf{W}_h^K, \mathbf{W}_h^Q, \mathbf{W}_h^P$ and $\mathbf{W}_h^V$ with $h \in [H]$ can minimize the ICL risk in the infinite-context limit.

Although previous works such as Ahn et al. [2023] and Mahankali et al. [2024] provide analytical solutions corresponding to stationary points of the ICL risk, these results are derived under the assumption that the prior mean $\boldsymbol{w}_\star = 0$. In this special case, the gradient of the ICL risk can vanish, allowing the existence of a stationary point. Our analysis generalizes this observation: we prove that when $\boldsymbol{w}_\star \neq 0$, the gradient of the ICL risk remains strictly non-zero for all weights as context size $C \to \infty$, thus precluding the existence of stationary points. We adopt $C \to \infty$ as an asymptotic approach, as done by Zhang et al. [2024a], Huang and Ge [2024]. Therefore, our result does not contradict prior findings but rather extends them to the more realistic setting of non-zero prior means.

Finally, even though such a stationary point exists with finite context size, we still cannot imply that the stationary point is the global optimum, as the ICL risk of multi-head LSA $\mathcal{R}(f_{\mathsf{H-LSA}})$ is not convex, presented in the following lemma.

Lemma 1. *The ICL risk of multi-head LSA $\mathcal{R}(f_{\mathsf{H-LSA}})$ is not convex.*

Because $\mathcal{R}(f_{\mathsf{H-LSA}})$ is non-convex, any stationary point that arises—even at finite context sizes—does not guarantee a global optimum. In other words, one may encounter local minima or saddle points that satisfy the stationary condition without minimizing the overall ICL risk.

D y_q-Linear Self-Attention

To address the performance gap between one-step GD and multi-head LSA, we introduce y_q-LSA, a generalization of single-head LSA.

D.1 Formulation of y_q-LSA

Our approach builds upon the *GD-transformer* developed by Von Oswald et al. [2023], Rossi et al. [2024], which implements one-step GD in a linear regression setup when the prior mean $\mathbf{w}_\star$ is zero. The original formulation is defined by the weight matrices

$$\mathbf{W}^V = \begin{bmatrix} 0 & 0 \\ \mathbf{w}_\star^\top & -1 \end{bmatrix}, \quad \mathbf{W}^K = \mathbf{W}^Q = \begin{bmatrix} \mathbf{I}_d & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{W}^P = -\frac{\eta}{C}\mathbf{I}_{d+1}, \tag{9}$$

where η represents the GD step size. From the standard LSA formulation (3) with the given embedding (2), we derive

$$f_{\mathsf{LSA}}(\mathbf{E}) = y_q - \frac{\eta}{C}(\mathbf{w}_\star^\top \mathbf{X}^\top - \mathbf{y}^\top)\mathbf{X}\mathbf{x}_q, \tag{10}$$

where the initial guess $y_q = 0 = \mathbf{w}_\star^\top \mathbf{x}_q$ is fixed for any query $\mathbf{x}_q$, and the prior mean $\mathbf{w}_\star$ is zero. See the derivation of (10) in Appendix G for the completeness. Notably, we retain the terms for y_q and $\mathbf{w}_\star$ to facilitate future extension to non-zero scenarios. Rewriting the equation (10) with $y_q = \mathbf{w}_\star^\top \mathbf{x}_q$ yields

$$f_{\mathsf{LSA}}(\mathbf{E}) = \left(\mathbf{w}_\star - \frac{\eta}{C}\mathbf{X}^\top(\mathbf{X}\mathbf{w}_\star - \mathbf{y})\right)^\top \mathbf{x}_q. \tag{11}$$

The red term represents the gradient of the least-squares loss in linear regression. Consequently, $f_{\mathsf{LSA}}(\mathbf{E})$ becomes equivalent to a linear function $f(\mathbf{x}_q) = \mathbf{w}^\top \mathbf{x}_q$, where $\mathbf{w}$ is the one-step GD update initialized at the prior mean $\mathbf{w}_\star$.

For the more general case with a non-zero prior mean $\mathbf{w}_\star$, we relax the condition on the initial guess y_q. By allowing y_q to be a linear function of x_q, specifically $y_q = \mathbf{w}_\star^\top \mathbf{x}_q$, we obtain the prediction of the linear regression task with a given query $\mathbf{x}_q$

$$\left(\mathbf{w}_\star - \frac{\eta}{C}\mathbf{X}^\top(\mathbf{X}\mathbf{w}_\star - \mathbf{y})\right)^\top \mathbf{x}_q, \tag{12}$$

which still implements the one-step GD update. Given this, we can now define y_q-LSA.

Definition 3 (y_q-LSA)**.** *We define y_q-LSA with a flexible initial guess embedding matrix*

$$\mathbf{E}_\mathbf{w} \overset{def}{=} \begin{bmatrix} \mathbf{X}^\top & \mathbf{x}_q \\ \mathbf{y}^\top & y_q \end{bmatrix} \in \mathbb{R}^{(d+1)\times(C+1)}, \quad \textit{with } y_q = \mathbf{w}^\top \mathbf{x}_q, \tag{13}$$

where $\mathbf{w} \in \mathbb{R}^d$ *is a trainable parameter and* y_q *is the initial guess. The* y_q*-LSA function is defined as*

$$f_{y_q-\mathsf{LSA}}(\mathbf{X}, \mathbf{y}, \mathbf{x}_q) \stackrel{def}{=} f_{\mathsf{LSA}}(\mathbf{E}_{\mathbf{w}}). \tag{14}$$

The y_q-LSA extends the standard LSA by introducing an additional parameter $\mathbf{w}$ in the embedding, enabling better alignment with the query's initial guess. The trainable parameters of y_q -LSA now include $\mathbf{W}^K, \mathbf{W}^Q, \mathbf{W}^P, \mathbf{W}^V$ and $\mathbf{w}$, with inputs $\mathbf{X}, \mathbf{y}$ and $\mathbf{x}_q$.

D.2 Analysis of y_q-LSA

Similar to the analysis of multi-head LSA, we first examine the stationary point of y_q-LSA.

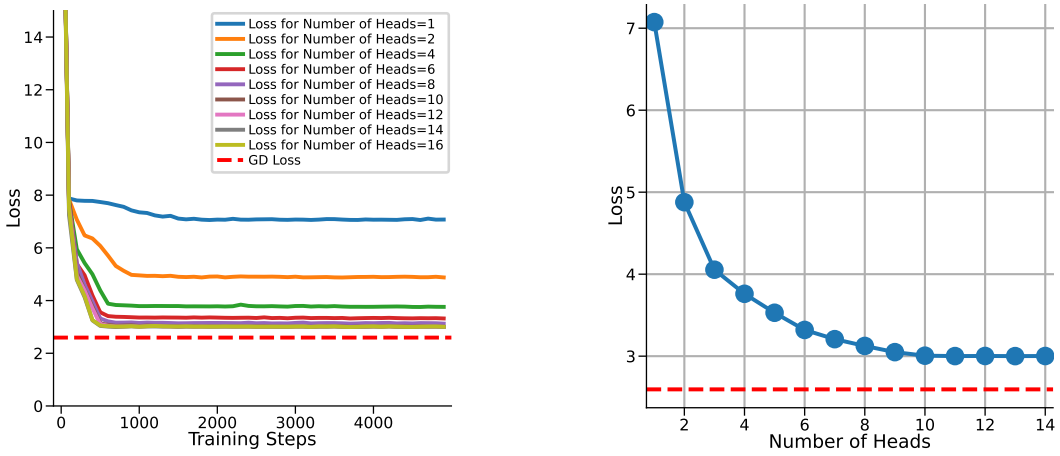
Theorem 4. *For a* y_q*-LSA function in* (14) *with a non-zero prior mean* $\mathbf{w}_\star$ *and contetxt size* $C \to \infty$*, the weights* $(\mathbf{W}^K, \mathbf{W}^Q, \mathbf{W}^P, \mathbf{W}^V, \mathbf{w}_\star)$ *in* (9) *with* $\mathbf{w} = \mathbf{w}_\star$ *constitute a stationary point of* $\mathcal{R}(f_{y_q-\mathsf{LSA}})$.

However, similar to multi-head LSA, we cannot conclusively determine that this stationary point represents the global optimum. This uncertainty comes from the non-convex nature of the y_q-LSA ICL risk, as established in the following lemma.

Lemma 2. *The ICL risk of* y_q*-LSA* $\mathcal{R}(f_{y_q-\mathsf{LSA}})$ *is not convex.*

While the non-convexity prevents a definitive proof of global optimality, our empirical investigations in Appendix E.2 suggest an intriguing hypothesis. Notably, we conjecture that the stationary point identified in Theorem 4 may indeed be the global optimum. Empirical evidence indicates that the performance of one-step gradient descent serves as a lower bound for y_q-LSA.

An additional noteworthy observation is y_q-LSA's relationship to the *linear transformer block* introduced by Zhang et al. [2024b]. Unlike y_q-LSA, LTB combines LSA with a linear multilayer perceptron (MLP) component. Critically, the global optimum of LTB implements a Newton step rather than one-step gradient descent. This approach fails to bridge the performance gap between one-step GD and single-head LSA and requires significantly more parameters through the additional MLP, in contrast to y_q-LSA's more parsimonious approach of introducing a single vector parameter $\mathbf{w}$. See Lemma 3 in Appendix G for more details.



(a) (b)

Figure 1: **Training loss of multi-head LSA with different numbers of attention heads.** In (a), we visualize the training loss curves for models with different head configurations, while (b) shows the final trained loss as a function of the number of heads.

E Experiments

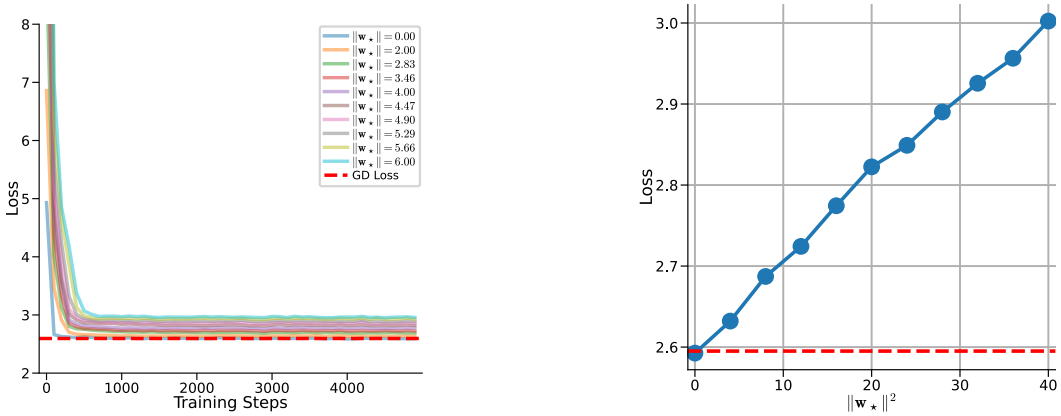
For experiments in Appendices E.1 and E.2, we focus on a simplified setting where the LSA consists of a single linear self-attention layer without LayerNorm or softmax. We generate linear functions in

a 10-dimensional input space ($d = 10$) and provide $C = 10$ context examples per task. We train for 5000 gradient steps. Further implementation details are provided in Appendix H.1.

E.1 Multi-head LSA

E.1.1 Multi-head LSA with Varying Numbers of Heads

We investigate the ICL risk (evaluation loss) of the multi-head LSA under different numbers of attention heads in the setting of a non-zero prior mean and y_q is fixed at zero (details in Table 1). Fig. 1a illustrates the loss curves over the course of training for several head configurations, while Fig. 1b summarizes the final evaluation losses as a function of the number of heads. From these results, we observe that increasing the number of heads up to $d + 1$ (here $d = 10$, see Fig. 1b) substantially enhances the in-context learning capability of multi-head LSA, as reflected by a pronounced reduction in the final evaluation loss. However, adding more than $d + 1$ heads yields negligible further improvement, indicating a saturation effect beyond this threshold. This confirms our results in Theorem 1. Notably, even at $d + 1$ heads, the multi-head LSA model does not converge to the one-step GD baseline loss, suggesting that while additional heads can capture richer in-context information[Crosbie and Shutova, 2024], they alone are insufficient for achieving full parity with the one-step GD performance in non-zero prior means setting. In other words, one-step GD loss serves as a strict lower bound of the ICL risk for multi-head LSA empirically.



(a) (b)

Figure 2: **Training loss of multi-head LSA under different prior means $w_\star$.** (a) Training loss curves for different values of $\|w_\star\|$. (b) Final trained loss as a function of $\|w_\star\|^2$.

E.1.2 Effect of Prior Mean $w_\star$n Multi-Head LSA.

We investigate how the prior mean $w_\star$, which represents the mean weight of the generated linear function, affects the performance of multi-head LSA when the number of heads is fixed at or above $d+1$ and y_q is fixed at zero. Fig. 2a shows the loss curves for different values of $\|w_\star\|$, while Fig. 2b presents the final trained loss as a function of $\|w_\star\|^2$.

Our results demonstrate that even when the number of heads is sufficiently large (i.e., $\geq d + 1$, reaching the optimal multi-head LSA configuration), multi-head LSA only matches the loss of one-step GD when the prior mean $w_\star$ is zero. For non-zero prior means, a systematic gap remains between Multi-Head LSA and one-step GD. Furthermore, this gap increases linearly with the squared ℓ_2 norm of the prior mean, $\|w_\star\|^2$, indicating that the prior mean significantly impacts the optimal loss and that larger deviations from zero result in a larger discrepancy from the GD baseline.

E.1.3 Effect of y_qn LSA

To investigate the effect of the initial guess y_q, contained in the embedding matrix (2) on the in-context learning ability of multi-head LSA, we decompose y_q into two components:

$$y_q = x_q^\top y_{\text{q\_guess}} + y_{\text{q\_bias}}.$$

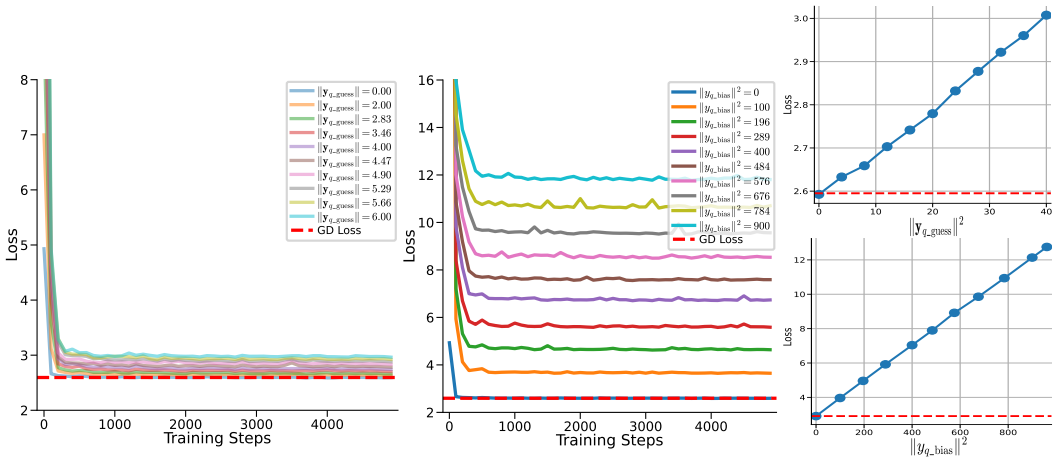
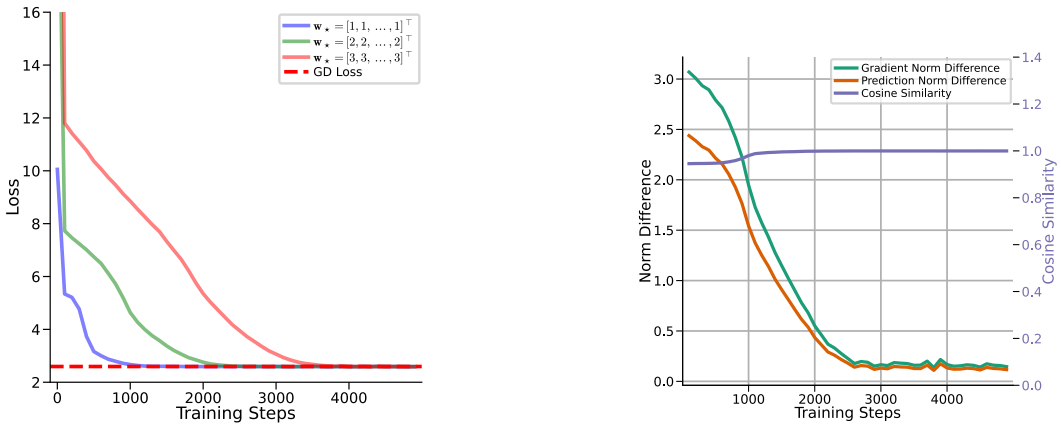


Figure 3: **Training and final loss of multi-head LSA under different initial guess configurations. Left** Training loss curves for various $\|y_{\text{q\_bias}}\|^2$, **Middle** Final trained loss as a function of $\|y_{\text{q\_bias}}\|^2$, **Right Upper** Training loss curves for various $\|\mathbf{y}_{\text{q\_guess}}\|$, and **Right Lower** Final trained loss as a function of $\|\mathbf{y}_{\text{q\_guess}}\|^2$.

We set the prior mean $\mathbf{w}_\star$ to zero and number of head is $d+1$, then conduct two separate experiments: (1) varying $\mathbf{y}_{\text{q\_guess}}$ while fixing $y_{\text{q\_bias}} = 0$, and (2) varying $y_{\text{q\_bias}}$ while fixing $\mathbf{y}_{\text{q\_guess}} = 0$. This allows us to isolate the contribution of each component to the model's behavior.

As shown in Fig. 3, multi-head LSA only converges to the same loss as one-step GD when $\mathbf{y}_{\text{q\_guess}} = 0$ (i.e., equal to the prior mean) and $y_{\text{q\_bias}} = 0$. In all other cases, a systematic gap remains between the loss of multi-head LSA and one-step GD. Moreover, this gap is directly proportional to $\|\mathbf{y}_{\text{q\_guess}}\|^2$ (the squared ℓ_2-norm of the guessed component) and $\|y_{\text{q\_bias}}\|^2$ (the squared bias term). These findings suggest that deviations in y_q from the optimal initialization introduce a persistent discrepancy in multi-head LSA's performance relative to one-step GD, regardless of the training of multi-head LSA.



(a) (b)

Figure 4: **Training loss and sensitivity analysis of y_q-LSA.** (a) Training loss curves of y_q-LSA and one-step GD. (b) Model behavior metrics including prediction norm difference, gradient norm difference, and cosine similarity.

E.2 y_q-LSA

In this section, we aim to empirically validate whether y_q-LSA, introduced in Appendix D, aligns with one-step GD across different prior settings. Fig. 4 presents the training loss of y_q-LSA. In Fig. 4a, we compare the convergence of y_q-LSA to one-step GD, demonstrating that regardless of

the prior configuration, y_q-LSA effectively matches the GD solution. Fig. 4b provides a detailed evaluation of prediction norm differences, gradient norm differences (defined in Appendix H.2), and cosine similarity between the models. The results confirm that y_q-LSA exhibits strong alignment with one-step GD in both loss convergence and gradient analysis.

F Proofs of Section C

F.1 Proof of Theorem 2

The proof of Theorem 2 is based on the analysis of Ahn et al. [2023].

Proof. **Step 1: Simplify the risk function and compute its gradient**

We first derive explicitly the expression of multi-head LSA's ICL risk and simplify it. The key idea is to decompose the ICL risk into components. That is,

$$\begin{aligned}
\mathcal{R}(f_{\mathsf{H-LSA}}) &\overset{(4)}{=} \mathbb{E}\left[(f_{\mathsf{H-LSA}}(\mathbf{E}) - y)^2\right] \quad \text{with } y = \widehat{\mathbf{w}}^\top \mathbf{x}_q \text{ and } \widehat{\mathbf{w}} \sim \mathcal{N}(\mathbf{w}_\star, \mathbf{I}_d), \\
&\overset{(6)}{=} \mathbb{E}\left[\left(\left[\mathbf{E} + \sum_{h=1}^{H} \text{head}_h(\mathbf{E})\right]_{-1,-1} - \widehat{\mathbf{w}}^\top \mathbf{x}_q\right)^2\right] \\
&\overset{(5)}{=} \mathbb{E}\left[\left(\left[\mathbf{E} + \frac{1}{C}\sum_{h=1}^{H} \mathbf{W}_h^P \mathbf{W}_h^V \mathbf{E} \mathbf{W}^M \left(\mathbf{E}^\top (\mathbf{W}_h^K)^\top \mathbf{W}_h^Q \mathbf{E}\right)\right]_{-1,-1} - \widehat{\mathbf{w}}^\top \mathbf{x}_q\right)^2\right].
\end{aligned}$$

Since the prediction of $f_{\mathsf{H-LSA}}$ is the bottom right entry of the output matrix, only the last row of the product $\mathbf{W}_h^P \mathbf{W}_h^V$ contributes to the prediction. Therefore, we write

$$\mathbf{W}_h^P \mathbf{W}_h^V \overset{\text{def}}{=} \begin{bmatrix} * \\ \mathbf{b}_h^\top \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)},$$

where $\mathbf{b}_h \in \mathbb{R}^{d+1}$ for all $h \in [H]$, and $*$ denotes entries that do not affect the ICL risk.

To simplify the computation, we also rewrite the product $(\mathbf{W}_h^K)^\top \mathbf{W}_h^Q$ and the embedding matrix $\mathbf{E}$ as

$$\begin{aligned}
(\mathbf{W}_h^K)^\top \mathbf{W}_h^Q &\overset{\text{def}}{=} \mathbf{A}_h \in \mathbb{R}^{(d+1)\times(d+1)}, \\
\mathbf{E} &\overset{\text{def}}{=} [\mathbf{z}_1 \quad \mathbf{z}_2 \quad \cdots \quad \mathbf{z}_C \quad \mathbf{z}_{C+1}] \in \mathbb{R}^{(d+1)\times(C+1)},
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{A}_h &\overset{\text{def}}{=} [\mathbf{a}_1^h \quad \mathbf{a}_2^h \quad \cdots \quad \mathbf{a}_{d+1}^h] \quad \text{with } \mathbf{a}_1^h, \cdots, \mathbf{a}_{d+1}^h \in \mathbb{R}^{d+1}, \\
\mathbf{z}_i &\overset{\text{def}}{=} \begin{bmatrix} \mathbf{x}_i \\ y_i \end{bmatrix} \in \mathbb{R}^{d+1} \quad \text{for all } i \in [C], \quad \text{and} \quad \mathbf{z}_{C+1} \overset{\text{def}}{=} \begin{bmatrix} \mathbf{x}_q \\ y_q \end{bmatrix} \in \mathbb{R}^{d+1}.
\end{aligned}$$

We define

$$\mathbf{G} \overset{\text{def}}{=} \frac{1}{C}\sum_{i=1}^{C} \mathbf{z}_i \mathbf{z}_i^\top = \frac{1}{C}\mathbf{E}\mathbf{W}^M\mathbf{E}^\top \in \mathbb{R}^{(d+1)\times(d+1)} \qquad \text{and} \qquad \widehat{\mathbf{w}} \overset{\text{def}}{=} \mathbf{w}_\star + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \in \mathbb{R}^d \sim \mathcal{N}(0, \mathbf{I}_d)$ is the noise.

Then the ICL risk can be written as

$$\begin{aligned}\mathcal{R}(f_{\mathsf{H-LSA}}) &= \mathbb{E}\left[\left(y_q + \sum_{h=1}^{H} \mathbf{b}_h^\top \mathbf{G}\mathbf{A}_h \mathbf{z}_{C+1} - \widehat{\mathbf{w}}^\top \mathbf{x}_q\right)^2\right] \\ &= \mathbb{E}\left[\left(y_q + \sum_{h=1}^{H} \mathbf{b}_h^\top \mathbf{G}\begin{bmatrix}\mathbf{a}_1^h & \mathbf{a}_2^h & \cdots & \mathbf{a}_{d+1}^h\end{bmatrix}\begin{bmatrix}\mathbf{x}_q \\ y_q\end{bmatrix} - \widehat{\mathbf{w}}^\top \mathbf{x}_q\right)^2\right] \\ &= \mathbb{E}\left[\left(y_q + \sum_{h=1}^{H}\left(\sum_{i=1}^{d} \mathbf{b}_h^\top \mathbf{G}\mathbf{a}_i^h \mathbf{x}_q[i]\right) + \mathbf{b}_h^\top \mathbf{G}\mathbf{a}_{d+1}^h y_q - \widehat{\mathbf{w}}^\top \mathbf{x}_q\right)^2\right],\end{aligned}$$

where $\mathbf{x}_q[i]$ is the i-th coordinate of the vector $\mathbf{x}_q$.

Furthermore, we know that, for all $h \in [H]$ and $i \in [d+1]$,

$$\mathbf{b}_h^\top \mathbf{G}\mathbf{a}_i^h \in \mathbb{R} = \mathrm{Tr}\left(\mathbf{b}_h^\top \mathbf{G}\mathbf{a}_i^h\right) = \mathrm{Tr}\left(\mathbf{G}\mathbf{a}_i^h \mathbf{b}_h^\top\right) = \left\langle \mathbf{G}, \mathbf{b}_h (\mathbf{a}_i^h)^\top \right\rangle,$$

where $\langle \mathbf{U}, \mathbf{V}\rangle \stackrel{\text{def}}{=} \mathrm{Tr}\left(\mathbf{U}\mathbf{V}^\top\right)$ is the Frobenius inner product for any squared matrices $\mathbf{U}$ and $\mathbf{V}$.

Hence, by using the linearity of the Frobenius inner product, we rewrite the ICL risk as

$$\begin{aligned}&\mathcal{R}(f_{\mathsf{H-LSA}}) \\ &= \mathbb{E}\left[\left(y_q + \sum_{h=1}^{H} \left\langle \mathbf{G}, \mathbf{b}_h (\mathbf{a}_{d+1}^h)^\top \right\rangle y_q + \sum_{i=1}^{d}\sum_{h=1}^{H} \left(\left\langle \mathbf{G}, \mathbf{b}_h (\mathbf{a}_i^h)^\top \right\rangle - \widehat{\mathbf{w}}[i]\right)\mathbf{x}_q[i]\right)^2\right] \\ &= \mathbb{E}\left[\left(\left(1 + \left\langle \mathbf{G}, \sum_{h=1}^{H} \mathbf{b}_h (\mathbf{a}_{d+1}^h)^\top \right\rangle\right) y_q + \sum_{i=1}^{d}\left(\left\langle \mathbf{G}, \sum_{h=1}^{H} \mathbf{b}_h (\mathbf{a}_i^h)^\top \right\rangle - \widehat{\mathbf{w}}[i]\right)\mathbf{x}_q[i]\right)^2\right],\end{aligned}$$

where $\widehat{\mathbf{w}}[i]$ is the i-th coordinate of the vector $\widehat{\mathbf{w}}$.

By reparametrizing the ICL risk, using a composite function, we have

$$\mathcal{R}(f_{\mathsf{H-LSA}}) = \mathbb{E}_{\mathbf{G},\widehat{\mathbf{w}},\mathbf{x}_q}\left[\left(\left(1 + \langle \mathbf{G}, \mathbf{M}_{d+1}\rangle\right) y_q + \sum_{i=1}^{d}\left(\langle \mathbf{G}, \mathbf{M}_i\rangle - \widehat{\mathbf{w}}[i]\right)\mathbf{x}_q[i]\right)^2\right], \tag{15}$$

where

$$\mathbf{M}_i \stackrel{\text{def}}{=} \sum_{h=1}^{H} \mathbf{b}_h (\mathbf{a}_i^h)^\top \in \mathbb{R}^{(d+1)\times(d+1)}, \qquad \text{for all } i \in [d+1].$$

Recall $\mathbf{x}_q \sim \mathcal{N}(0, \mathbf{I}_d)$. Thus, both $\mathbf{G}$ and $\widehat{\mathbf{w}}$ are independent to $\mathbf{x}_q[i]$ for all $i \in [d]$, and $\mathbf{x}_q[i] \sim \mathcal{N}(0,1)$ are i.i.d.

Expanding (15) yields

$$\begin{aligned}\mathcal{R}(f_{\mathsf{H-LSA}}) &= \mathbb{E}_{\mathbf{G}}\left[\left(1 + \langle \mathbf{G}, \mathbf{M}_{d+1}\rangle\right)^2 y_q^2\right] + \sum_{i=1}^{d} \mathbb{E}_{\mathbf{G},\widehat{\mathbf{w}}}\left[\left(\langle \mathbf{G}, \mathbf{M}_i\rangle - \widehat{\mathbf{w}}[i]\right)^2\right]\mathbb{E}_{\mathbf{x}_q}\left[\mathbf{x}_q[i]^2\right] \\ &= \mathbb{E}_{\mathbf{G}}\left[\left(1 + \langle \mathbf{G}, \mathbf{M}_{d+1}\rangle\right)^2 y_q^2\right] + \sum_{i=1}^{d} \mathbb{E}_{\mathbf{G},\widehat{\mathbf{w}}}\left[\left(\langle \mathbf{G}, \mathbf{M}_i\rangle - \widehat{\mathbf{w}}[i]\right)^2\right] \\ &= \sum_{i=1}^{d+1} \mathcal{L}_i(\mathbf{M}_i),\end{aligned} \tag{16}$$

where

$$\begin{aligned}\mathcal{L}_i(\mathbf{M}_i) &\stackrel{\text{def}}{=} \mathbb{E}_{\mathbf{G},\widehat{\mathbf{w}}}\left[\left(\langle \mathbf{G}, \mathbf{M}_i\rangle - \widehat{\mathbf{w}}[i]\right)^2\right] \qquad \text{for all } i \in [d], \\ \mathcal{L}_{d+1}(\mathbf{M}_{d+1}) &\stackrel{\text{def}}{=} \mathbb{E}_{\mathbf{G}}\left[\left(1 + \langle \mathbf{G}, \mathbf{M}_{d+1}\rangle\right)^2 y_q^2\right].\end{aligned}$$

Thus, the ICL risk (16) is decomposed into $(d+1)$ separated components $\mathcal{L}_i$ with $i \in [d+1]$. Each component is a function of $\mathbf{M}_i$. To compute the gradient of $\mathcal{R}(f_{\mathsf{H-LSA}})$, we can first compute the gradient of each component with respect to $\mathbf{M}_i$ for $i \in [d]$. That is,

$$\nabla_{\mathbf{M}_i} \mathcal{L}_i(\mathbf{M}_i) = 2\mathbb{E}_{\mathbf{G},\widehat{\mathbf{w}}} \left[\langle \mathbf{G}, \mathbf{M}_i \rangle \mathbf{G}\right] - 2\mathbb{E}_{\mathbf{G},\widehat{\mathbf{w}}} \left[\widehat{\mathbf{w}}[i]\mathbf{G}\right], \qquad \text{for } i \in [d], \tag{17}$$

$$\nabla_{\mathbf{M}_{d+1}} \mathcal{L}_{d+1}(\mathbf{M}_{d+1}) = 2y_q^2 \mathbb{E}_{\mathbf{G}} \left[(1 + \langle \mathbf{G}, \mathbf{M}_{d+1} \rangle) \mathbf{G}\right]. \tag{18}$$

Step 2: Compute $\mathbb{E}\left[\langle \mathbf{G}, \mathbf{M}_i \rangle \mathbf{G}\right]$, $\mathbb{E}\left[\widehat{\mathbf{w}}[i]\mathbf{G}\right]$ in (17)

Recall that $\widehat{\mathbf{w}} \sim \mathcal{N}(\mathbf{w}_\star, \mathbf{I}_d)$ and $\mathbf{x}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$ are independent for all $j \in [C]$, $y_j = \widehat{\mathbf{w}}^\top \mathbf{x}_j$, and $\widehat{\mathbf{w}} = \mathbf{w}_\star + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I}_d)$.

For $\mathbb{E}_{\mathbf{G},\widehat{\mathbf{w}}} \left[\widehat{\mathbf{w}}[i]\mathbf{G}\right]$ in (17) with $i \in [d]$, we have

$$\mathbb{E}_{\mathbf{G},\widehat{\mathbf{w}}} \left[\widehat{\mathbf{w}}[i]\mathbf{G}\right] = \frac{1}{C} \sum_{j=1}^{C} \begin{bmatrix} \mathbb{E}_{\widehat{\mathbf{w}},\mathbf{x}_j}[\widehat{\mathbf{w}}[i] \cdot \mathbf{x}_j \mathbf{x}_j^\top] & \mathbb{E}_{\widehat{\mathbf{w}},\mathbf{x}_j}[\widehat{\mathbf{w}}[i] \cdot y_j \mathbf{x}_j] \\ \mathbb{E}_{\widehat{\mathbf{w}},\mathbf{x}_j}[\widehat{\mathbf{w}}[i] \cdot y_j \mathbf{x}_j^\top] & \mathbb{E}_{\widehat{\mathbf{w}},\mathbf{x}_j}[\widehat{\mathbf{w}}[i] \cdot y_j^2] \end{bmatrix}.$$

In particular, for each block of the above matrix, we have

$$\begin{aligned}
\mathbb{E}_{\widehat{\mathbf{w}},\mathbf{x}_j}[\widehat{\mathbf{w}}[i] \cdot \mathbf{x}_j \mathbf{x}_j^\top] &= \mathbb{E}_{\widehat{\mathbf{w}}} \left[\widehat{\mathbf{w}}[i]\right] \mathbb{E}_{\mathbf{x}_j} \left[\mathbf{x}_j \mathbf{x}_j^\top\right] = \mathbf{w}_\star[i] \mathbf{I}_d, \\
\mathbb{E}_{\widehat{\mathbf{w}},\mathbf{x}_j}[\widehat{\mathbf{w}}[i] \cdot y_j \mathbf{x}_j] &= \mathbb{E}_{\widehat{\mathbf{w}},\mathbf{x}_j} \left[\widehat{\mathbf{w}}[i] \widehat{\mathbf{w}}^\top \mathbf{x}_j \mathbf{x}_j\right] \\
&= \mathbb{E}_{\boldsymbol{\varepsilon},\mathbf{x}_j} \left[(\mathbf{w}_\star[i] + \boldsymbol{\varepsilon}[i])(\mathbf{w}_\star + \boldsymbol{\varepsilon})^\top \mathbf{x}_j \mathbf{x}_j\right] = \mathbf{w}_\star[i] \mathbf{w}_\star + \mathbf{e}_i, \\
\mathbb{E}_{\widehat{\mathbf{w}},\mathbf{x}_j}[\widehat{\mathbf{w}}[i] \cdot y_j^2] &= \mathbb{E}_{\widehat{\mathbf{w}},\mathbf{x}_j}[\widehat{\mathbf{w}}[i] \widehat{\mathbf{w}}^\top \mathbf{x}_j \mathbf{x}_j^\top \widehat{\mathbf{w}}] = \mathbb{E}_{\widehat{\mathbf{w}}} \left[\widehat{\mathbf{w}}[i] \widehat{\mathbf{w}}^\top \widehat{\mathbf{w}}\right] \\
&= \mathbb{E}_{\boldsymbol{\varepsilon}} \left[(\mathbf{w}_\star[i] + \boldsymbol{\varepsilon}[i])(\mathbf{w}_\star + \boldsymbol{\varepsilon})^\top (\mathbf{w}_\star + \boldsymbol{\varepsilon})\right] = \mathbf{w}_\star[i](\|\mathbf{w}_\star\|^2 + d + 2),
\end{aligned}$$

where $\mathbf{e}_i$ denotes the standard basis vector with zeros in all coordinates except the i-th position, where the value is 1.

Combining the above three components, we have

$$\mathbb{E}_{\mathbf{G},\widehat{\mathbf{w}}} \left[\widehat{\mathbf{w}}[i]\mathbf{G}\right] = \begin{bmatrix} \mathbf{w}_\star[i] \mathbf{I}_d & \mathbf{w}_\star[i] \mathbf{w}_\star + \mathbf{e}_i \\ (\mathbf{w}_\star[i] \mathbf{w}_\star + \mathbf{e}_i)^\top & \mathbf{w}_\star[i](\|\mathbf{w}_\star\|^2 + d + 2) \end{bmatrix}. \tag{19}$$

Now we compute $\mathbb{E}\left[\langle \mathbf{G}, \mathbf{M}_i \rangle \mathbf{G}\right]$ for $i \in [d]$.

We start by calculating the expected value of the product of elements in matrix $\mathbf{G}$. That is, for all $m, n, p, q \in [d+1]$,

$$\mathbb{E}\left[\mathbf{G}_{mn} \mathbf{G}_{pq}\right] = \frac{1}{C^2} \sum_{j=1}^{C} \sum_{k=1}^{C} \mathbb{E}\left[\mathbf{z}_j[m] \mathbf{z}_j[n] \mathbf{z}_k[p] \mathbf{z}_k[q]\right],$$

where $\mathbf{G}_{mn}$ is the value of matrix $\mathbf{G}$ in m-th row and n-th column position for all $m, n \in [d+1]$. By expanding the summation, we have

$$\begin{aligned}
\mathbb{E}\left[\mathbf{G}_{mn} \mathbf{G}_{pq}\right] &= \frac{1}{C^2} \sum_{\substack{1 \le j,k \le C \\ j \neq k}} \mathbb{E}\left[\mathbf{z}_j[m] \mathbf{z}_j[n] \mathbf{z}_k[p] \mathbf{z}_k[q]\right] + \frac{1}{C} \mathbb{E}\left[\mathbf{z}_1[m] \mathbf{z}_1[n] \mathbf{z}_1[p] \mathbf{z}_1[q]\right] \\
&= \frac{C(C-1)}{C^2} \mathbb{E}\left[\mathbf{z}_1[m] \mathbf{z}_1[n]\right] \mathbb{E}\left[\mathbf{z}_2[p] \mathbf{z}_2[q]\right] + \frac{1}{C} \mathbb{E}\left[\mathbf{z}_1[m] \mathbf{z}_1[n] \mathbf{z}_1[p] \mathbf{z}_1[q]\right] \\
&\approx \mathbb{E}\left[\mathbf{z}_1[m] \mathbf{z}_1[n]\right] \mathbb{E}\left[\mathbf{z}_2[p] \mathbf{z}_2[q]\right], \qquad \text{when } C \longrightarrow \infty.
\end{aligned}$$

To compute $\mathbb{E}\left[\mathbf{z}_1[m] \mathbf{z}_1[n]\right]$,

1. For $m, n \in [d]$, we have $\mathbf{z}_1[m] = \mathbf{x}_1[n]$ and $\mathbf{z}_1[n] = \mathbf{x}_1[n]$. Thus, $\mathbb{E}\left[\mathbf{z}_1[m] \mathbf{z}_1[n]\right] = \delta_{mn}$, where δ is the Kronecker delta.
2. For $m \in [d]$ and $n = d+1$, we have $\mathbf{z}_1[n] = y_1$. Thus, $\mathbb{E}\left[\mathbf{z}_1[m] \mathbf{z}_1[n]\right] = \mathbb{E}\left[\mathbf{x}_1[m] \mathbf{x}_1^\top \widehat{\mathbf{w}}\right] = \mathbf{w}_\star[m]$.
3. For $m = n = d+1$, we have $\mathbb{E}\left[\mathbf{z}_1[m] \mathbf{z}_1[n]\right] = \mathbb{E}\left[\widehat{\mathbf{w}}^\top \mathbf{x}_1 \mathbf{x}_1^\top \widehat{\mathbf{w}}\right] = \mathbb{E}\left[\widehat{\mathbf{w}}^\top \widehat{\mathbf{w}}\right] = \|\mathbf{w}_\star\|^2 + d$.

We denote

$$\mathbf{M} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{I}_d & \mathbf{w}_\star \\ \mathbf{w}_\star^\top & \|\mathbf{w}_\star\|^2 + d \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}. \tag{20}$$

By using (20), when $C \longrightarrow \infty$, we have

$$\mathbb{E}\left[\mathbf{G}_{mn}\mathbf{G}_{pq}\right] = \mathbf{M}_{mn}\mathbf{M}_{pq}. \tag{21}$$

By linearity of the Frobenius inner product, we have

$$\mathbb{E}\left[\langle \mathbf{G}, \mathbf{M}_i \rangle \mathbf{G}\right] = \langle \mathbf{M}, \mathbf{M}_i \rangle \mathbf{M}. \tag{22}$$

Combining the above equation with (19), (17) becomes

$$\begin{aligned} \nabla_{\mathbf{M}_i} \mathcal{L}_i(\mathbf{M}_i) &= 2\langle \mathbf{M}, \mathbf{M}_i \rangle \mathbf{M} - 2\begin{bmatrix} \mathbf{w}_\star[i]\mathbf{I}_d & \mathbf{w}_\star[i]\mathbf{w}_\star + \mathbf{e}_i \\ (\mathbf{w}_\star[i]\mathbf{w}_\star + \mathbf{e}_i)^\top & \mathbf{w}_\star[i](\|\mathbf{w}_\star\|^2 + d + 2) \end{bmatrix} \\ &= 2\langle \mathbf{M}, \mathbf{M}_i \rangle \mathbf{M} - 2\mathbf{w}_\star[i]\mathbf{M} - 2\mathbf{N} \\ &= (2\langle \mathbf{M}, \mathbf{M}_i \rangle - 2\mathbf{w}_\star[i])\mathbf{M} - 2\mathbf{N}, \end{aligned} \tag{23}$$

where

$$\mathbf{N} \stackrel{\text{def}}{=} \begin{bmatrix} 0 & \mathbf{e}_i \\ \mathbf{e}_i^\top & 2\mathbf{w}_\star[i] \end{bmatrix}.$$

Notice that $\mathbf{M}$ is full rank and the rank of $\mathbf{N}$ is smaller or equal to 2. Thus, for any $\mathbf{M}_i \in \mathbb{R}^{(d+1)\times(d+1)}$, we have

$$\nabla_{\mathbf{M}_i} \mathcal{L}_i(\mathbf{M}_i) \neq 0.$$

□

F.2 Proofs of Theorem 1

Observe that each $\mathbf{M}_i$ is an element of $\mathbb{R}^{(d+1)\times(d+1)}$, a $(d+1)^2$-dimensional real vector space. Because we have $\{\mathbf{M}_i\}_{i=1}^{d+1}$, we are effectively working inside the product space

$$\left(\mathbb{R}^{(d+1)\times(d+1)}\right)^{d+1},$$

whose dimension is

$$(d+1) \times (d+1)^2 \;=\; (d+1)^3.$$

Hence, any collection $\{\mathbf{M}_i\}_{i=1}^{d+1}$ occupies at least $(d+1)^3$ degrees of freedom in total.

For each index $h \in \{1, \ldots, H\}$, we introduce vectors $\mathbf{b}_h$, $\mathbf{a}_1^h, \ldots, \mathbf{a}_d^h$, $\mathbf{a}_{d+1}^h$ in $\mathbb{R}^{d+1}$. Each such block contributes

$$(d+1) \;+\; d(d+1) \;+\; (d+1) \;=\; (d+1)(d+2)$$

real parameters. Repeating for all $h = 1, \ldots, H$, the full parameter space Ω_H is:

$$\Omega_H \;=\; \left[\mathbb{R}^{d+1} \times (\mathbb{R}^{d+1})^d \times \mathbb{R}^{d+1}\right]^H,$$

and it has dimension

$$\dim\left(\Omega_H\right) \;=\; H\,(d+1)(d+2).$$

When $H \geq d+1$, we have

$$H\,(d+1)\,(d+2) \;\;\geq\;\; (d+1)^3.$$

Considering $\mathbf{M}_i$ is bilinear combined by $\mathbf{b}_h$, $\mathbf{a}_1^h, \ldots, \mathbf{a}_d^h$, $\mathbf{a}_{d+1}^h$, we need at least $d+1$ head to have a surjection from parameter space to target space.

We can construct a surjection as following:

For any $\mathbf{M}_i$,

$$\mathbf{M}_i = \sum_{h=1}^{H} \mathbf{b}_h (\mathbf{a}_i^h)^\top = \sum_{h=1}^{H} \mathbf{e}_h (\mathbf{M}_i[h])^\top,$$

in which $\mathbf{M}_i[h]$ means the h-th row in the $\mathbf{M}_i$

Therefore when $H \geq d+1$, all $\mathbf{M}$ can be constructed.

Then,

$$\inf_{f \in \mathcal{F}_{f_{(d+1)-\mathsf{LSA}}}} \mathcal{R}(f) = \inf_{f \in \mathcal{F}_{f_{(d+2)-\mathsf{LSA}}}} \mathcal{R}(f).$$

F.3 Proof of Lemma 1

Proof. From (23), we can compute the Hessian of the function $\mathcal{L}_i(\mathbf{M}_i)$, that is,

$$\nabla^2_{\mathbf{M}_i}\mathcal{L}_i(\mathbf{M}_i) = 2\mathbf{M}.$$

We verify that $\mathbf{M}$ is positive semi-definite. Indeed, let $\mathbf{u} \in \mathbb{R}^d$ and $u \in \mathbb{R}$. We have

$$\begin{aligned}
\begin{bmatrix}\mathbf{u}^\top & u\end{bmatrix}\mathbf{M}\begin{bmatrix}\mathbf{u}\\ u\end{bmatrix} &\overset{(20)}{=} \begin{bmatrix}\mathbf{u}^\top & u\end{bmatrix}\begin{bmatrix}\mathbf{I}_d & \mathbf{w}_\star\\ \mathbf{w}_\star^\top & \|\mathbf{w}_\star\|^2 + d\end{bmatrix}\begin{bmatrix}\mathbf{u}\\ u\end{bmatrix}\\
&= \begin{bmatrix}\mathbf{u}^\top & u\end{bmatrix}\begin{bmatrix}\mathbf{u} + u\mathbf{w}_\star\\ \mathbf{w}_\star^\top\mathbf{u} + u(\|\mathbf{w}_\star\|^2 + d)\end{bmatrix}\\
&= \|\mathbf{u}\|^2 + 2u\mathbf{w}_\star^\top\mathbf{u} + u^2(\|\mathbf{w}_\star\|^2 + d)\\
&= \|\mathbf{u} + u\mathbf{w}_\star\|^2 + du^2 \geq 0.
\end{aligned}$$

Since $\mathbf{M}$ is positive semi-definite, we have the function $\mathcal{L}_i$ is convex with respect to $\mathbf{M}_i$.

From (16), we know that

$$\mathcal{R}(f_{\mathsf{H-LSA}}) = \sum_{i=1}^{d+1}\mathcal{L}_i(\mathbf{M}_i).$$

Each function $\mathcal{L}_i$ is a function of $\mathbf{M}_i$. We denote

$$\mathcal{R}(f_{\mathsf{H-LSA}}) = f(\mathbf{M}_1, \cdots, \mathbf{M}_{d+1}).$$

Then the Hessian of the function f with respect to variables $\mathbf{M}_1, \cdots, \mathbf{M}_{d+1}$ is a block diagonal matrix, each block on the diagonal is $\nabla^2_{\mathbf{M}_i}\mathcal{L}_i(\mathbf{M}_i) \geq 0$. Therefore, the function f is convex with respect to $\mathbf{M}_1, \cdots, \mathbf{M}_{d+1}$.

Lastly, $\mathbf{M}_i = \sum_{h=1}^{H}\mathbf{b}_h(\mathbf{a}_i^h)^\top$ for $i \in [d+1]$. To simplify it, we can consider only one head. That is, $\mathbf{M}_i = \mathbf{b}_1(\mathbf{a}_i^1)^\top$, a bilinear function, which is known to be not convex with respect to $\mathbf{b}_1$ and $\mathbf{a}_i^1$.

To conclude, the ICL risk $\mathcal{R}(f_{\mathsf{H-LSA}})$ is a composite function with a convex function and non convex functions, which implies that $\mathcal{R}(f_{\mathsf{H-LSA}})$ is not convex. □

G Proofs of Section D

G.1 Derivation of (10)

Here we provide the derivation of (10). Recall

$$\mathbf{W}^V = \begin{bmatrix}0 & 0\\ \mathbf{w}_\star^\top & -1\end{bmatrix}, \quad \mathbf{W}^K = \mathbf{W}^Q = \begin{bmatrix}\mathbf{I}_d & 0\\ 0 & 0\end{bmatrix}, \quad \mathbf{W}^P = -\frac{\eta}{C}\mathbf{I}_{d+1}.$$

From the standard LSA formulation (3) with the given embedding in (2), we have

$$\begin{aligned}
\mathbf{K} \overset{\text{def}}{=} \mathbf{Q} \overset{\text{def}}{=} \mathbf{W}^Q\mathbf{E} &= \begin{bmatrix}\mathbf{X}^\top & \mathbf{x}_q\\ 0 & 0\end{bmatrix},\\
\mathbf{V} \overset{\text{def}}{=} \mathbf{W}^V\mathbf{E} &= \begin{bmatrix}0 & 0\\ \mathbf{w}_\star^\top\mathbf{X}^\top - \mathbf{y}^\top & \mathbf{w}_\star^\top\mathbf{x}_q - y_q\end{bmatrix}.
\end{aligned}$$

So we get the LSA simplified as

$$f_{\mathsf{LSA}}(\mathbf{E}) = \left[\mathbf{E} + \mathbf{W}^P\mathbf{V}\mathbf{W}^M\left(\mathbf{K}^\top\mathbf{Q}\right)\right]_{-1,-1}.$$

In this case, we have

$$\mathbf{V}\mathbf{W}^M\left(\mathbf{K}^\top\mathbf{Q}\right) = \begin{bmatrix}0 & 0\\ (\mathbf{w}_\star^\top\mathbf{X}^\top - \mathbf{y}^\top)\mathbf{X}\mathbf{X}^\top & (\mathbf{w}_\star^\top\mathbf{X}^\top - \mathbf{y}^\top)\mathbf{X}\mathbf{x}_q\end{bmatrix},$$

and LSA recovers the result in Von Oswald et al. [2023], which performs one-step GD on the update of the linear regression parameter initialized at $\mathbf{w}_\star = \mathbf{0}$ with $y_q = 0 = \mathbf{w}_\star^\top\mathbf{x}_q$:

$$\begin{aligned}
f_{\mathsf{LSA}}(\mathbf{E}) &= y_q - \frac{\eta}{C}(\mathbf{w}_\star^\top\mathbf{X}^\top - \mathbf{y}^\top)\mathbf{X}\mathbf{x}_q\\
&= \left(\mathbf{w}_\star - \frac{\eta}{C}\mathbf{X}^\top(\mathbf{X}\mathbf{w}_\star - \mathbf{y})\right)^\top\mathbf{x}_q,
\end{aligned}$$

that yields (10).

G.2 y_q-LSA is a Special Case of Linear Transformer Block

In this section, we show that y_q-LSA defined in (14) is a special case of *linear transformer block* (LTB) presented in Zhang et al. [2024b], which is mentioned in Appendix D.

LTB combines LSA with a linear multilayer perceptron (MLP) component. That is,

$$f_{\mathsf{LTB}} : \mathbb{R}^{(d+1)\times(C+1)} \to \mathbb{R} \tag{24}$$
$$\mathbf{E} \mapsto \left[\mathbf{W}_2^\top \mathbf{W}_1 \left(\mathbf{E} + \frac{1}{C}\mathbf{W}^P \mathbf{W}^V \mathbf{E} \mathbf{W}^M \mathbf{E}^\top (\mathbf{W}^K)^\top \mathbf{W}^Q \mathbf{E}\right)\right]_{-1,-1},$$

where $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}^P, \mathbf{W}^V, \mathbf{W}^K$ and $\mathbf{W}^Q$ are trainable parameters for f_{LTB}, and

$$\mathbf{E} = \begin{bmatrix} \mathbf{X}^\top & \mathbf{x}_q \\ \mathbf{y}^\top & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(C+1)},$$

for $\mathbf{X} \in \mathbb{R}^{C\times d}, \mathbf{y} \in \mathbb{R}^C$ and $\mathbf{x}_q \in \mathbb{R}^d$. Notice that there is no initial guess y_q involved in this embedding matrix $\mathbf{E}$.

We denote the hypothesis class formed by LTB models as

$$\mathcal{F}_{\mathsf{LTB}} \overset{\text{def}}{=} \left\{ f_{\mathsf{LTB}} : \mathbf{W}^K, \mathbf{W}^Q, \mathbf{W}^V, \mathbf{W}^P, \mathbf{W}_1, \mathbf{W}_2 \right\},$$

where f_{LTB} is defined in (24). Then we have the following lemma.

Lemma 3. *Consider $f_{y_q-\mathsf{LSA}}$ defined in* (14)*. We have*

$$f_{y_q-\mathsf{LSA}} \in \mathcal{F}_{\mathsf{LTB}}.$$

Proof. Let $\mathbf{w} \in \mathbb{R}^d$. For all $\mathbf{X} \in \mathbb{R}^{C\times d}, \mathbf{y} \in \mathbb{R}^C$ and $\mathbf{x}_q \in \mathbb{R}^d$, we have

$$f_{y_q-\mathsf{LSA}}(\mathbf{X}, \mathbf{y}, \mathbf{x}_q) = f_{\mathsf{LSA}}(\mathbf{E}_\mathbf{w}) = \left[\mathbf{E}_\mathbf{w} + \tfrac{1}{C}\mathbf{W}^P \mathbf{W}^V \mathbf{E}_\mathbf{w} \mathbf{W}^M (\mathbf{E}_\mathbf{w}^\top (\mathbf{W}^K)^\top \mathbf{W}^Q \mathbf{E}_\mathbf{w})\right]_{-1,-1},$$

with

$$\mathbf{E}_\mathbf{w} = \begin{bmatrix} \mathbf{X}^\top & \mathbf{x}_q \\ \mathbf{y}^\top & \mathbf{w}^\top \mathbf{x}_q \end{bmatrix} \in \mathbb{R}^{(d+1)\times(C+1)}.$$

We aim to find $(\mathbf{W}^K)', (\mathbf{W}^Q)', (\mathbf{W}^V)', (\mathbf{W}^P)', \mathbf{W}_1, \mathbf{W}_2$ for f_{LTB} such that $f_{y_q-\mathsf{LSA}}(\mathbf{X}, \mathbf{y}, \mathbf{x}_q) = f_{\mathsf{LTB}}(\mathbf{E})$ with

$$\mathbf{E} = \begin{bmatrix} \mathbf{X}^\top & \mathbf{x}_q \\ \mathbf{y}^\top & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(C+1)}.$$

Let choose $\mathbf{W}_2 = \mathbf{I}_{d+1}$ and

$$\mathbf{W}_1 = \begin{bmatrix} \mathbf{I}_d & \mathbf{w} \\ \mathbf{w}^\top & c \end{bmatrix} \tag{25}$$

with $c \neq \|\mathbf{w}\|^2$, then $\mathbf{W}_2^\top \mathbf{W}_1 = \mathbf{W}_1$ and $\mathbf{W}_1 \in \mathbb{R}^{(d+1)\times(d+1)}$ is invertible.

Indeed, let $\mathbf{u} \in \mathbb{R}^d$ and $u \in \mathbb{R}$ such that $\mathbf{W}_1 \begin{bmatrix} \mathbf{u} \\ u \end{bmatrix} = 0$. So we have

$$\mathbf{u} + u\mathbf{w} = 0,$$
$$\mathbf{w}^\top \mathbf{u} + cu = 0.$$

From $\mathbf{u} + u\mathbf{w} = 0$, we have $\mathbf{u} = -u\mathbf{w}$. Plugging it into $\mathbf{w}^\top \mathbf{u} + cu = 0$, we obtain

$$(c - \|\mathbf{w}\|^2)u = 0.$$

Since $c \neq \|\mathbf{w}\|^2$, we obtain $u = 0$. Thus, $\mathbf{u} = -u\mathbf{w} = 0$. This implies that $\mathbf{W}_1$ is invertible.

Next, we consider the following matrix

$$\mathbf{W}_3 = \begin{bmatrix} \mathbf{I}_d & 0 \\ \mathbf{w}^\top & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}.$$

Let

$$(\mathbf{W}^P)' = \mathbf{W}_1^{-1}\mathbf{W}^P,$$
$$(\mathbf{W}^K)' = \mathbf{W}^K\mathbf{W}_3,$$
$$(\mathbf{W}^Q)' = \mathbf{W}^Q\mathbf{W}_3,$$
$$(\mathbf{W}^V)' = \mathbf{W}^V\mathbf{W}_3.$$

We show that $f_{y_q-\mathsf{LSA}}(\mathbf{X}, \mathbf{y}, \mathbf{x}_q) = f_{\mathsf{LTB}}(\mathbf{E})$.

Indeed, by using $\mathbf{X}\mathbf{w} = \mathbf{y}$, we have

$$\mathbf{W}_3\mathbf{E} = \begin{bmatrix} \mathbf{X}^\top & \mathbf{x}_q \\ \mathbf{w}^\top\mathbf{X}^\top & \mathbf{w}^\top\mathbf{x}_q \end{bmatrix} = \mathbf{E}_\mathbf{w}.$$

So

$$\begin{aligned} f_{\mathsf{LTB}}(\mathbf{E}) &= \mathbf{W}_1\left[\left(\mathbf{E} + \frac{1}{C}\mathbf{W}_1^{-1}\mathbf{W}^P\mathbf{W}^V\mathbf{W}_3\mathbf{E}\mathbf{W}^M\mathbf{E}^\top(\mathbf{W}^K\mathbf{W}_3)^\top\mathbf{W}^Q\mathbf{W}_3\mathbf{E}\right)\right]_{-1,-1} \\ &= [\mathbf{W}_1\mathbf{E}]_{-1,-1} + \left[\left(\frac{1}{C}\mathbf{W}^P\mathbf{W}^V\mathbf{E}_\mathbf{w}\mathbf{W}^M(\mathbf{E}_\mathbf{w}^\top(\mathbf{W}^K)^\top\mathbf{W}^Q\mathbf{E}_\mathbf{w}\right)\right]_{-1,-1} \\ &= \mathbf{w}^\top\mathbf{x}_q + \left[\left(\frac{1}{C}\mathbf{W}^P\mathbf{W}^V\mathbf{E}_\mathbf{w}\mathbf{W}^M(\mathbf{E}_\mathbf{w}^\top(\mathbf{W}^K)^\top\mathbf{W}^Q\mathbf{E}_\mathbf{w}\right)\right]_{-1,-1} \\ &= f_{y_q-\mathsf{LSA}}(\mathbf{X}, \mathbf{y}, \mathbf{x}_q). \end{aligned}$$

Thus, we conclude $f_{y_q-\mathsf{LSA}} \in \mathcal{F}_{\mathsf{LTB}}$. □

G.3 Proofs of Theorem 4

The risk (loss) function with learnable vector $\mathbf{v}$ is given by:

$$\mathcal{R}(f_{y_q-\mathsf{LSA}}) = \mathbb{E}\left[\left(\left(\mathbf{E} + \frac{1}{C}\mathrm{Att}(\mathbf{E})\right)_{C+1,C+1} + \mathbf{v}^\top\mathbf{x}_q - \widehat{\mathbf{w}}^T\mathbf{x}_q\right)^2\right].$$

Similar as Appendix F, we rewrite the risk:

$$\begin{aligned} \mathcal{R}(f_{y_q-\mathsf{LSA}}) &= \mathbb{E}\Big[\big((1 + \mathbf{b}^T\mathbf{G}\mathbf{a}_{d+1})y_q + (\mathbf{b}^T\mathbf{G}\mathbf{A}_{:d} - \widehat{\mathbf{w}}^\top)\mathbf{x}_q\big)^2\Big] \\ &= \mathbb{E}\left[\big((1 + \mathbf{b}^T\mathbf{G}\mathbf{a}_{d+1})\mathbf{v}^\top + (\mathbf{b}^T\mathbf{G}\mathbf{A}_{:d} - \widehat{\mathbf{w}}^\top)\big)\,\mathbf{x}_q\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{d}\Big(\langle\mathbf{G}, \mathbf{b}\mathbf{a}_j^\top\rangle + \langle\mathbf{G}, \mathbf{b}\mathbf{a}_{d+1}^\top\rangle\mathbf{v}[j] + \mathbf{v}[j] - \widehat{\mathbf{w}}[j]\Big)^2\right] \end{aligned}$$

We define, for each j:

$$t_j \;=\; \langle\mathbf{G},\ \mathbf{b}\,\mathbf{a}_j^\top\rangle \;+\; \langle G,\ \mathbf{b}\,\mathbf{a}_{d+1}^\top\rangle\,\mathbf{v}[j] \;+\; \mathbf{v}[j] \;-\; \widehat{\mathbf{w}}[j].$$

Then

$$f_{y_q-\mathsf{LSA}} \;=\; \sum_{j=1}^{d}\mathbb{E}\big[t_j^2\big].$$

Step 1: Gradient for parameters

We list the first-order partial derivatives with respect to $\mathbf{b}, \mathbf{a}_j, \mathbf{a}_{d+1}$, and $\mathbf{v}[j]$. j is from 1 to d

- **Gradient w.r.t.** $\mathbf{b}$

$$\frac{\partial t_j}{\partial \mathbf{b}} \;=\; \mathbf{G}\,\mathbf{a}_j \;+\; \mathbf{v}[j]\,\mathbf{G}\,\mathbf{a}_{d+1}.$$

$$\frac{\partial}{\partial \mathbf{b}}\big(t_j^2\big) \;=\; 2\,t_j\,\frac{\partial t_j}{\partial \mathbf{b}} \;=\; 2\,t_j\,\big(\mathbf{G}\,\mathbf{a}_j + \mathbf{v}[j]\,\mathbf{G}\,\mathbf{a}_{d+1}\big).$$

$$\frac{\partial f}{\partial \mathbf{b}} \;=\; \sum_{j=1}^{d} \mathbb{E}\Big[\,2\,t_j\,\big(\mathbf{G}\,\mathbf{a}_j + \mathbf{v}[j]\;\mathbf{G}\,\mathbf{a}_{d+1}\big)\Big].$$

- **Gradient w.r.t.** $\mathbf{a}_j$

$$\frac{\partial t_j}{\partial \mathbf{a}_j} \;=\; \mathbf{G}^\top \mathbf{b}.$$

$$\frac{\partial}{\partial \mathbf{a}_j}\big(t_j^2\big) \;=\; 2\,t_j\,\big(\mathbf{G}^\top \mathbf{b}\big).$$

Only the j-th term depends on $\mathbf{a}_j$, so

$$\frac{\partial f_{y_q-\mathsf{LSA}}}{\partial \mathbf{a}_j} \;=\; 2\,\mathbb{E}\big[t_j\,(\mathbf{G}^\top \mathbf{b})\big].$$

- **Gradient w.r.t.** $\mathbf{a}_{d+1}$

$$\frac{\partial t_j}{\partial \mathbf{a}_{d+1}} \;=\; \mathbf{v}[j]\;\big(\mathbf{G}^\top \mathbf{b}\big).$$

$$\frac{\partial}{\partial \mathbf{a}_{d+1}}\big(t_j^2\big) \;=\; 2\,t_j\,\big(\mathbf{v}[j]\;\mathbf{G}^\top \mathbf{b}\big).$$

$$\frac{\partial f_{y_q-\mathsf{LSA}}}{\partial \mathbf{a}_{d+1}} \;=\; 2\sum_{j=1}^{d} \mathbb{E}\Big[t_j\;\mathbf{v}[j]\;(\mathbf{G}^\top \mathbf{b})\Big].$$

- **Gradient w.r.t.** $v[j]$

We have

$$t_j \;=\; \mathbf{b}^\top \mathbf{G}\,\mathbf{a}_j \;+\; v[j]\,\big(\mathbf{b}^\top \mathbf{G}\,\mathbf{a}_{d+1} + 1\big) \;-\; \big(\mathbf{w}[j] + \mathbf{w}_\star[j]\big).$$

$$\frac{\partial t_j}{\partial v[j]} \;=\; \big(\mathbf{b}^\top \mathbf{G}\,\mathbf{a}_{d+1} + 1\big).$$

$$\frac{\partial f_{y_q-\mathsf{LSA}}}{\partial v[j]} \;=\; 2\,\mathbb{E}\Big[t_j\,\big(\mathbf{b}^\top \mathbf{G}\,\mathbf{a}_{d+1} + 1\big)\Big].$$

Step 2: Plug in One Step GD

we verify when $\mathbf{b} = \begin{bmatrix} -\mathbf{w}_\star \\ 1 \end{bmatrix}$, $\mathbf{a}_j = \begin{bmatrix} \mathbf{e}_j \\ 0 \end{bmatrix}$, $\mathbf{a}_{d+1} = 0$, $\mathbf{v} = \mathbf{w}_\star$, the gradients equal to zero

we define $\mathbf{w} = \widehat{\mathbf{w}} - \mathbf{w}_\star$, We have the following intermediate formula:

$$\mathbf{b}^T\mathbf{G}\mathbf{a}_j = [-\mathbf{w}_\star^T, 1]\sum_{i=1}^{C}\begin{bmatrix}\mathbf{x}_i\mathbf{x}_i^T & \mathbf{x}_i y_i^T \\ y_i\mathbf{x}_i^T & y_i^2\end{bmatrix}\begin{bmatrix}\mathbf{e}_j \\ 0\end{bmatrix} = \frac{\sum_{i=1}^{C}}{C}\left[\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i^T, \mathbf{w}^T\mathbf{x}_i y_i\right]\left[\begin{bmatrix}\mathbf{e}_j \\ 0\end{bmatrix}\right] = \frac{\sum_{i=1}^{C}}{C}\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i[j]$$

$$v[i](\mathbf{b}^T\mathbf{G}\mathbf{a}_{d+1}) = 0$$

$$t_j = \frac{\sum_{i=1}^{C} \mathbf{w}^T \mathbf{x}_i \mathbf{x}_i[j]}{C} - \mathbf{w}[j]$$

$$\mathbf{G} a_j = \frac{1}{C} \sum_{i=1}^{C} \begin{bmatrix} \mathbf{x}_i \mathbf{x}_i[j] \\ y_i \mathbf{x}_i[j] \end{bmatrix}$$

- **Gradient w.r.t. $\mathbf{b}$**

$$\frac{\partial f_{y_q-\mathsf{LSA}}}{\partial b} = 2 \sum_{j=1}^{d} \mathbb{E} \left[\left(\frac{\sum_{i=1}^{C} \mathbf{w}^T \mathbf{x}_i \mathbf{x}_i[j]}{C} - \mathbf{w}[j] \right) \frac{1}{C} \sum_{i=1}^{C} \begin{bmatrix} \mathbf{x}_i \mathbf{x}_i[j] \\ y_i \mathbf{x}_i[j] \end{bmatrix} \right]$$

Calculate each part:

$$-\mathbf{w}[j] \frac{1}{C} \sum_{i=1}^{C} \mathbf{x}_i \mathbf{x}_i[j] = 0,$$

$$-\mathbf{w}[j] \frac{1}{C} \sum_{i=1}^{C} y_i \mathbf{x}_i[j] = -\mathbf{w}[j] \frac{1}{C} \sum_{i=1}^{C} (\mathbf{w}_\star^T + \mathbf{w}^T) \mathbf{x}_i \mathbf{x}_i[j] = -\mathbf{w}[j] \frac{1}{C} \sum_{i=1}^{C} \mathbf{w}^T \mathbf{x}_i \mathbf{x}_i[j] = -1,$$

$$\frac{\sum_{i=1}^{C} \mathbf{w}^T \mathbf{x}_i \mathbf{x}_i[j] \mathbf{x}_i \mathbf{x}_i[j]}{C} = 0,$$

$$\frac{\sum_{i=1}^{C} \mathbf{w}^T \mathbf{x}_i \mathbf{x}_i[j]}{C} \frac{1}{C} \sum_{i=1}^{C} (\mathbf{w}_\star^T + \mathbf{w}^T) \mathbf{x}_i \mathbf{x}_i[j] = \frac{\sum_{i=1}^{C} \mathbf{w}^T \mathbf{x}_i \mathbf{x}_i[j]}{C} \frac{\sum_{i=1}^{C} \mathbf{w}^T \mathbf{x}_i \mathbf{x}_i[j]}{C}$$

$$= \frac{1}{C^2} \mathbb{E} \left[\Big(\sum_{i=1}^{C} \mathbf{x}_i[j] \, \mathbf{x}_i^T \Big) \, \Big(\sum_{k=1}^{C} \mathbf{x}_k[j] \, \mathbf{x}_k \Big) \right],$$

compute $\mathbb{E} \left[\Big(\sum_{i=1}^{C} \mathbf{x}_i[j] \, \mathbf{x}_i^T \Big) \, \Big(\sum_{k=1}^{C} \mathbf{x}_k[j] \, \mathbf{x}_k \Big) \right]$

when $i \neq k$,

$$\mathbb{E}[\mathbf{x}_i[j] \mathbf{x}_k[j] (\mathbf{x}_i^T \mathbf{x}_k)] = 1.$$

$$\sum_{i \neq k} \mathbb{E}[\mathbf{x}_i[j] \mathbf{x}_k[j] (\mathbf{x}_i^T \mathbf{x}_k)] = C(C-1) \cdot 1 = C(C-1).$$

when $i = k$,

$$\mathbf{x}_i[j] \mathbf{x}_i[j] (\mathbf{x}_i^T \mathbf{x}_i) = \mathbf{x}_i[j]^2 \sum_{m=1}^{d} \mathbf{x}_i[m]^2 = \mathbf{x}_i[j]^2 (\mathbf{x}_i^T \mathbf{x}_i) = d + 2.$$

Because $\mathbb{E}[\mathbf{x}[j]^2 (\mathbf{x}^T \mathbf{x})] = \mathbb{E}[\mathbf{x}[j]^4] + \sum_{m \neq j} \mathbb{E}[\mathbf{x}[j]^2 \mathbf{x}[m]^2]$, $\mathbb{E}[\mathbf{x}[j]^4] = 3$.

$$\Big(\sum_{i=1}^{C} \mathbf{x}_i[j] \, \mathbf{x}_i^T \Big) \, \Big(\sum_{k=1}^{C} \mathbf{x}_k[j] \, \mathbf{x}_k \Big) = C(C-1) + C(d+2)$$

if we have very large C, we have:

$$\frac{\sum_{i=1}^{C} \mathbf{w}^T \mathbf{x}_i \mathbf{x}_i[j]}{C} \frac{1}{C} \sum_{i=1}^{C} y_i \mathbf{x}_i[j] = 1.$$

So that $\frac{\partial f_{y_q-\mathsf{LSA}}}{\partial \mathbf{b}} = 0$

- **Gradient w.r.t. $\mathbf{a}_j$**

$$\frac{\partial f_{y_q-\mathsf{LSA}}}{\partial \mathbf{a}_j} \;=\; 2\,\mathbb{E}\big[t_j\,(\mathbf{G}^\top \mathbf{b})\big] = \mathbb{E}\big[\frac{\sum_{i=1}^{C}}{C}\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i[j] - \mathbf{w}[j])\,\frac{\sum_{i=1}^{C}}{C}\begin{bmatrix}\mathbf{x}_i\mathbf{x}_i^T\mathbf{w}\\ y_i\mathbf{x}_i^T\mathbf{w}\end{bmatrix}\big]$$

$$\mathbb{E}\big[-\mathbf{w}[j]\frac{\sum_{i=1}^{C}}{C}\begin{bmatrix}\mathbf{x}_i\mathbf{x}_i^T\mathbf{w}\\ (\mathbf{w}^T+\mathbf{w}_\star^T)\mathbf{x}_i\mathbf{x}_i^T\mathbf{w}\end{bmatrix}\big] = \mathbb{E}\big[-\begin{bmatrix}\mathbf{e}_j\\ \mathbf{w}[j](\mathbf{w}^T+\mathbf{w}_\star^T)\mathbf{w}\end{bmatrix}\big] = -\begin{bmatrix}\mathbf{e}_j\\ \mathbf{w}_\star[j]\end{bmatrix}$$

compute $\frac{\sum_{i=1}^{C}}{C}\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i[j]\frac{\sum_{i=1}^{C}}{C}\mathbf{x}_i\mathbf{x}_i^T\mathbf{w}$

We aim to compute the expectation:

$$\mathbb{E}\left[\mathbf{w}^T\left(\sum_{i=1}^{C}\mathbf{x}_i\mathbf{x}_i[j]\right)\left(\sum_{k=1}^{C}\mathbf{x}_k\mathbf{x}_k^T\right)\mathbf{w}\right],$$

First, expand the product inside the expectation:

$$\mathbf{w}^T\left(\sum_{i=1}^{C}\mathbf{x}_i\mathbf{x}_i[j]\right)\left(\sum_{k=1}^{C}\mathbf{x}_k\mathbf{x}_k^T\right)\mathbf{w} = \sum_{i=1}^{C}\sum_{k=1}^{C}\mathbf{w}^T\mathbf{x}_i\,\mathbf{x}_i[j]\,\mathbf{x}_k^T\mathbf{w}\cdot\mathbf{x}_k.$$

Taking expectation:

$$\mathbb{E}\left[\sum_{i=1}^{C}\sum_{k=1}^{C}\mathbf{w}^T\mathbf{x}_i\,\mathbf{x}_i[j]\,\mathbf{x}_k^T\mathbf{w}\cdot\mathbf{x}_k\right] = \sum_{i=1}^{C}\sum_{k=1}^{C}\mathbb{E}\left[\mathbf{w}^T\mathbf{x}_i\,\mathbf{x}_i[j]\,\mathbf{x}_k^T\mathbf{w}\cdot\mathbf{x}_k\right].$$

Case 1: $i \neq k$

Since $\mathbf{x}_i$ and $\mathbf{x}_k$ are independent:

$$\mathbb{E}\left[\mathbf{w}^T\mathbf{x}_i\,\mathbf{x}_i[j]\,\mathbf{x}_k^T\mathbf{w}\cdot\mathbf{x}_k\right] = \mathbb{E}\left[\mathbf{w}^T\mathbf{x}_i\,\mathbf{x}_i[j]\right]\mathbb{E}\left[\mathbf{x}_k^T\mathbf{w}\cdot\mathbf{x}_k\right].$$

Given $\mathbf{x}_i \sim \mathcal{N}(0, I_d)$:

$$\mathbb{E}\left[\mathbf{w}^T\mathbf{x}_i\,\mathbf{x}_i[j]\right] = \mathbf{w}[j], \quad \mathbb{E}\left[\mathbf{x}_k^T w\cdot\mathbf{x}_k\right] = \mathbf{w}.$$

Thus, for $i \neq k$:

$$\mathbb{E}\left[\mathbf{w}^T\mathbf{x}_i\,\mathbf{x}_i[j]\,\mathbf{x}_k^T\mathbf{w}\cdot\mathbf{x}_k\right] = \mathbf{w}[j]\mathbf{w}.$$

There are $C(C-1)$ such terms, contributing:

$$C(C-1)\mathbf{w}[j]w = C(C-1)e_j.$$

Case 2: $i = k$

For $i = k$:

$$\mathbb{E}\left[\mathbf{w}^T\mathbf{x}_i\,\mathbf{x}_i[j]\,\mathbf{x}_i^T w\cdot\mathbf{x}_i\right] = \mathbb{E}\left[(\mathbf{w}^T\mathbf{x}_i)^2\mathbf{x}_i[j]\mathbf{x}_i\right].$$

Using properties of Gaussian vectors:

$$\mathbb{E}\left[(\mathbf{w}^T\mathbf{x}_i)^2\mathbf{x}_i[j]\mathbf{x}_i\right] = 2\mathbf{w}_j\mathbf{w} + \|w\|^2\mathbf{e}_j,$$

where e_j is the j-th standard basis vector. There are C such terms, contributing:

$$C(2\mathbf{w}_j w + \|\mathbf{w}\|^2\mathbf{e}_j).$$

Adding contributions from both cases:

$$\mathbb{E}\left[\mathbf{w}^T\left(\sum_{i=1}^{C}\mathbf{x}_i\mathbf{x}_i[j]\right)\left(\sum_{k=1}^{C}\mathbf{x}_k\mathbf{x}_k^T\right)w\right] = C(C-1)\mathbf{w}_j\|\mathbf{w}\|^2 + C(2\mathbf{w}_j\mathbf{w} + \|\mathbf{w}\|^2\mathbf{e}_j).$$

Simplifying:

$$= C(C+1)\mathbf{w}_j\mathbf{w} + C\|\mathbf{w}\|^2\mathbf{e}_j.$$

Thus, the expectation is:

$$\mathbb{E}\left[\mathbf{w}^T\left(\sum_{i=1}^{C}\mathbf{x}_i\mathbf{x}_i[j]\right)\left(\sum_{k=1}^{C}\mathbf{x}_k\mathbf{x}_k^T\right)\mathbf{w}\right] = C(C+1)\mathbf{w}_j\mathbf{w} + C\|\mathbf{w}\|^2\mathbf{e}_j.$$

when C is large $\mathbb{E}\left[\frac{\sum_{i=1}^{C}}{C}\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i[j]\frac{\sum_{i=1}^{C}}{C}\mathbf{x}_i\mathbf{x}_i^T\mathbf{w}\right] = \mathbf{e}_j$

compute $\frac{\sum_{i=1}^{C}}{C}\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i[j]\frac{\sum_{i=1}^{C}}{C}y_i\mathbf{x}_i^T w$

$$\frac{\sum_{i=1}^{C}}{C}\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i[j]\frac{\sum_{k=1}^{C}}{C}(\mathbf{w}_\star^T + \mathbf{w}^T)\mathbf{x}_k\mathbf{x}_k^T w$$

From our previous experience, we only need calculate case when $k \neq i$

$$\mathbb{E}\left[\frac{\sum_{i=1}^{C}}{C}\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i[j]\frac{\sum_{k=1}^{C}}{C}(\mathbf{w}_\star^T + \mathbf{w}^T)\mathbf{x}_k\mathbf{x}_k^T w\right] = \mathbb{E}\left[\mathbf{w}[j](\mathbf{w}_\star^T + \mathbf{w}^T)w\right] = \mathbf{w}_\star[j]$$

So that we have $\frac{\partial f_{y_q-\mathsf{LSA}}}{\partial \mathbf{a}_j} = 0$

- **Gradient w.r.t.** $\mathbf{a}_{d+1}$

$$\frac{\partial f_{y_q-\mathsf{LSA}}}{\partial \mathbf{a}_{d+1}} \;=\; 2\sum_{j=1}^{d}\mathbb{E}\Big[t_j\ \mathbf{v}[j]\ (\mathbf{G}^\top\mathbf{b})\Big] \;=\; 2\sum_{j=1}^{d}\mathbb{E}\Big[t_j\ \mathbf{w}_\star[j]\ (\mathbf{G}^\top\mathbf{b})\Big].$$

we already have $\frac{\partial f_{y_q-\mathsf{LSA}}}{\partial \mathbf{a}_j} \;=\; 2\,\mathbb{E}\big[t_j\,(\mathbf{G}^\top\mathbf{b})\big]$

So that we have $\frac{\partial f_{y_q-\mathsf{LSA}}}{\partial \mathbf{a}_{d+1}} = 0$

- **Gradient w.r.t.** $v[j]$

$$\frac{\partial f_{y_q-\mathsf{LSA}}}{\partial v[j]} \;=\; 2\,\mathbb{E}\Big[t_j\,\big(\mathbf{b}^\top\mathbf{G}\,\mathbf{a}_{d+1}+1\big)\Big] \;=\; 2\,\mathbb{E}\Big[\frac{(\sum_{i=1}^{C}\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i[j]-\mathbf{w}[j])}{C}\,\big(\frac{\sum_{i=1}^{C}}{C}\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i[j]+1\big)\Big]$$

$$2\,\mathbb{E}\Big[\frac{(\sum_{i=1}^{C}}{C}\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i[j] - \mathbf{w}[j])\,1\Big] = 0$$

$$2\,\mathbb{E}\Big[\frac{(\sum_{i=1}^{C}}{C}\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i[j] - \mathbf{w}[j])\,\frac{\sum_{k=1}^{C}}{C}\mathbf{w}^T\mathbf{x}_k\mathbf{x}_k[j]\Big]$$

we still only consider the case $i \neq k$

$$2\,\mathbb{E}\Big[\frac{(\sum_{i=1}^{C}}{C}\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i[j] - \mathbf{w}[j])\,\frac{\sum_{k=1}^{C}}{C}\mathbf{w}^T\mathbf{x}_k\mathbf{x}_k[j]\Big] = 2\,\mathbb{E}\Big[\mathbf{w}[j] - \mathbf{w}[j])\,\mathbf{w}^T\Big] = 0$$

we verify that $\mathbf{b} = \begin{bmatrix}-\mathbf{w}_\star\\ 1\end{bmatrix}$**,** $\mathbf{a}_j = \begin{bmatrix}\mathbf{e}_j\\ 0\end{bmatrix}$ $\mathbf{a}_{d+1} = 0$ $v = \mathbf{w}_\star$**, is a stationary point for loss** $f_{y_q-\mathsf{LSA}}$

G.4 Proof of Lemma 2

Proof. Based on the proof of Lemma 3, we consider the following matrix

$$\mathbf{W} = \begin{bmatrix}\mathbf{I}_d & 0\\ \mathbf{w}^\top & 0\end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}.$$

Now for any $f_{y_q-\mathsf{LSA}}$'s inputs $(\mathbf{X}, \mathbf{y}, \mathbf{x}_q)$, by using $\mathbf{X}\mathbf{w} = \mathbf{y}$, we have

$$\mathbf{W}\mathbf{E} = \begin{bmatrix} \mathbf{X}^\top & \mathbf{x}_q \\ \mathbf{w}^\top \mathbf{X}^\top & \mathbf{w}^\top \mathbf{x}_q \end{bmatrix} = \mathbf{E}_\mathbf{w}.$$

Thus,

$$f_{y_q-\mathsf{LSA}}(\mathbf{X}, \mathbf{y}, \mathbf{x}_q) = f_{\mathsf{LSA}}(\mathbf{E}_\mathbf{w}) = f_{\mathsf{LSA}}(\mathbf{W}\mathbf{E}).$$

By using Lemma 1 with one-single head, we know that $\mathcal{R}(f_{\mathsf{LSA}})$ is non-convex. Thus, we conclude that $\mathcal{R}(f_{y_q-\mathsf{LSA}})$ is non-convex, as it is a composite function with a non-convex function $\mathcal{R}(f_{\mathsf{LSA}})$ and a linear function. □

H Details of Experiment

H.1 Implementation Settings.

The experiments use JAX to implement and train the LSA models. We set the learning rate to $lr = 5 \times 10^{-4}$ and a batch size of 2,048. A single linear attention layer is used, without any LayerNorm or softmax operations. We will release our code repository upon publication to facilitate reproducibility.

Table 1: Overview of the experimental setups. Each experiment modifies one factor (number of attention heads, prior mean, or y_q) while holding the others fixed.

Experiment	**Number of Heads**	**Prior Mean**	$\boldsymbol{y_q}$
Head Appendix E.1.1	Varies	$[2, 2, \ldots, 2]$	0
Prior Mean Appendix E.1.2	11	Varies	0
y_q Appendix E.1.3	11	$[0, 0, \ldots, 0]$	Varies

H.2 Detailed Metric Definitions

Prediction Norm Difference The *prediction norm difference* measures the discrepancy between the outputs of y_q-LSA and one-step GD (f_{GD}). Given a test input $\mathbf{x}_q$, we define the difference as:

$$\|f_{y_q-\mathsf{LSA}}(\mathbf{x}_q) - f_{GD}(\mathbf{x}_q)\|.$$

This metric quantifies how closely y_q-LSA approximates the predictions of the explicit one-step GD solution.

Gradient Norm Difference The *gradient norm difference* assesses the deviation between the sensitivity of the model predictions to the input. Given the gradient of the output with respect to the input $\mathbf{x}_q$, we compute:

$$\left\| \frac{\partial f_{GD}(\mathbf{x}_q)}{\partial \mathbf{x}_q} - \frac{\partial f_{y_q-\mathsf{LSA}}(\mathbf{x}_q)}{\partial \mathbf{x}_q} \right\|.$$

This metric evaluates whether y_q-LSA captures the same local sensitivity as one-step GD.

Cosine Similarity The *cosine similarity* measures the angular alignment between the gradients of the two models. It is defined as:

$$\frac{\left\langle \frac{\partial f_{GD}(\mathbf{x}_q)}{\partial \mathbf{x}_q}, \frac{\partial f_{y_q-\mathsf{LSA}}(\mathbf{x}_q)}{\partial \mathbf{x}_q} \right\rangle}{\left\| \frac{\partial f_{GD}(\mathbf{x}_q)}{\partial \mathbf{x}_q} \right\| \left\| \frac{\partial f_{y_q-\mathsf{LSA}}(\mathbf{x}_q)}{\partial \mathbf{x}_q} \right\|}.$$

A cosine similarity of 1 indicates perfect alignment between the two models, while lower values suggest deviations in the learned representations.

H.3 LLM Experimental Settings

We conducted our experiments using the STS-Benchmark dataset (English subset)[May, 2021], which consists of sentence pairs labelled with semantic similarity scores ranging from 0 to 5. The LLM

used in our study was Meta-LLaMA-3.1-8B-Instruct[Grattafiori et al., 2024] and Qwen/Qwen2.5-7B-Instruct[Yang et al., 2024, Team, 2024]. The model's generation parameters included a maximum of 150 new tokens and deterministic decoding.

The guess model was trained to generate initial similarity score guesses. It consisted of a two-layer feedforward architecture, taking as input the concatenated embeddings of two sentences computed by the SentenceTransformer model all-MiniLM-L6-v2[Reimers and Gurevych, 2020]. The first layer mapped the concatenated embeddings to a 16-dimensional space with ReLU activation, followed by a second layer that outputs a single scalar value as the predicted similarity score. The model was trained using Adam Optimizer[Kingma, 2014] with a learning rate of 1e-3 and a mean squared error loss function. Training was performed over 10 epochs, with a batch size of 8. Sentence embeddings were dynamically computed during training. The loss for training the guess model was computed as the MSE between the predicted and ground truth scores.

For each prompt, a context was constructed by randomly sampling 10 labelled examples from the dataset. Each labelled example included two sentences, a ground truth similarity score, and an initial guess for the similarity score generated by a lightweight guess model. The query example included two sentences and its guessed similarity score and an explicit instruction for the LLM to refine the guess and provide a similarity score between 0 and 5.

To evaluate the effectiveness of the initial guess, we calculated the MSE between the LLM's predicted similarity scores and the ground truth scores across 100 experimental runs. The baseline performance, derived from the initial guesses provided was compared to the refined predictions generated by the LLM.