Local to Global: Learning Dynamics and Effect of Initialization for Transformers

Anonymous Authors[1]

Abstract

In recent years, transformer-based models have revolutionized deep learning, particularly in sequence modeling. To better understand this phenomenon, there is a growing interest in using Markov input processes to study transformers. However, our current understanding in this regard remains limited with many fundamental questions about how transformers learn Markov chains still unanswered. In this paper, we address this by focusing on first-order Markov chains and single-layer transformers, providing a comprehensive characterization of the learning dynamics in this context. Specifically, we prove that transformer parameters trained on next-token prediction loss can either converge to global or local minima, contingent on the initialization and the Markovian data properties, and we characterize the precise conditions under which this occurs. To the best of our knowledge, this is the first result of its kind highlighting the role of initialization. We further demonstrate that our theoretical findings are corroborated by empirical evidence. Based on these insights, we provide guidelines for the initilization of transformer parameters and demonstrate their effectiveness. Finally, we outline several open problems in this arena. Code is available at: https://anonymous.4open.science/r/Local-to-Global-C70B/.

1. Introduction

Transformers have been at the forefront of recent successes across various fields including natural language processing (Vaswani et al., 2017). To obtain insights into their impressive sequential modeling capabilities, a notable emerging theme among several recent works is to model the input data as a Markov process.

Using this Markovian perspective, works such as (Nichani et al., 2024; Edelman et al., 2024; Bietti et al., 2023), among others, study the in-context learning capabilities of a two-layer transformer. (Makkuva et al., 2024) analyzes the loss-landscape for the next-token prediction task, while (Ildiz et al., 2024) shows an equivalence between the attention mechanism and Markov models. Although these works reveal interesting insights about transformers and their capabilities, many fundamental questions about their learning dynamics remain unanswered. In particular, a comprehensive characterization of their training dynamics vis-á-vis the data distributional properties and the role of initialization is still missing.

To address this gap, in this paper, we focus on the canonical setting of first-order Markov chains and single-layer transformers and make the following contributions:

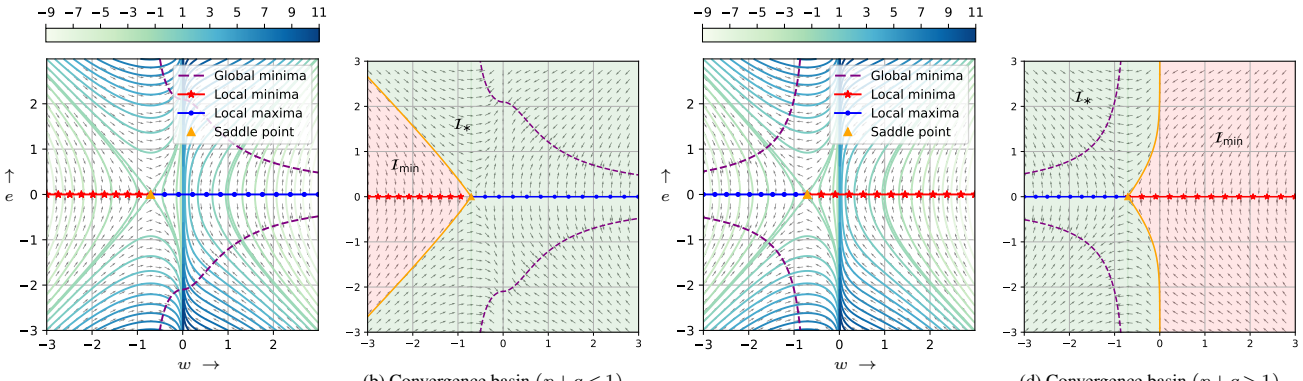
- ***Theoretical analysis:*** We precisely characterize the loss landscape and gradient flow dynamics for single-layer transformers with first-order Markov chains (Secs. 3 and 4). We demonstrate that transformer parameters trained on next-token prediction loss can converge to global or local minima, depending on the initialization and the Markovian data properties, and determine the exact conditions under which this occurs (Thms. 2, 3, and 8). To the best of our knowledge, this is the first result of its kind.

- ***Insights into initialization:*** Our theoretical analysis underscores the crucial role of initialization in transformer parameter training. Specifically, we demonstrate how the standard Gaussian initialization scheme can lead the convergence to local or global minima depending on the Markovian data properties (Thms. 2 and 8, Figs. 1 and 2).

- ***Guidelines:*** Based on these insights, we provide practical guidelines for parameter initialization, corroborated by empirical evidence demonstrating their effectiveness (§ A.3.2).

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

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



(a) Gradient flow ($p+q<1$) (b) Convergence basin ($p+q<1$) (c) Gradient flow ($p+q>1$) (d) Convergence basin ($p+q>1$)

Figure 1: Gradient flow dynamics and initialization effect for single-layer transformers. (p, q) are Markov switching probabilities, and (e, w) are the embedding and weight parameters (Sec. 2). (a), (c): The flow is aligned along energy contour lines, converging to local or global optima. (b), (d): $\mathcal{I}_\star$ is the basin of convergence for global minima, $\mathcal{I}_{\min}$ for the local minima, and yellow asymptotes for the saddle point. Notice the contrasting behavior for Gaussian initialization around origin for $p+q \lessgtr 1$.

2. Problem setting

We succinctly define the problem setting for analysis of single-layer transformers with Markovian data (§ A).

Input data. We assume that the input word sequence $\{x_n\}_{n=1}^N \sim (\boldsymbol{\pi}, \boldsymbol{P})$ is a first-order time-homogenous Markov chain with binary state space $\{0,1\}$ and a fixed kernel $\boldsymbol{P} = [1-p, p; q, 1-q]$, where $p = \boldsymbol{P}_{01} = \mathbb{P}(x_{n+1}=1 \mid x_n=0)$ and $q = \boldsymbol{P}_{10} = \mathbb{P}(x_{n+1}=0 \mid x_n=1)$ denote the switching probabilities from the states 0 and 1 respectively. We call $p+q$ the *switching factor*. $\boldsymbol{\pi} \triangleq (\pi_0, \pi_1) = (q,p)/(p+q)$ is the stationary distribution satisfying $\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}$.

Transformer architecture. We consider a single-layer transformer with a single-head attention and ReLU nonlinearity, which for an input sequence $\{x_n\}_{n=1}^N$, performs the following mathematical operations at each $n \in [N]$:

$$\begin{aligned}
\boldsymbol{x}_n &= x_n\,\boldsymbol{e} + \boldsymbol{p}_n \in \mathbb{R}^d, && \text{(Embedding)}\\
\boldsymbol{y}_n &= \boldsymbol{x}_n + \sum_{i\in[n]} \underbrace{\mathrm{att}_{n,i}}_{\in(0,1)} \cdot \boldsymbol{W}_V\,\boldsymbol{x}_i \in \mathbb{R}^d, && \text{(Attention)}\\
\boldsymbol{z}_n &= \boldsymbol{y}_n + \boldsymbol{W}_2\,\mathrm{ReLU}(\boldsymbol{W}_1\,\boldsymbol{y}_n) \in \mathbb{R}^d, && \text{(FF)}\\
\mathrm{logit}_n &= \langle \boldsymbol{a}, \boldsymbol{z}_n\rangle + b \qquad\quad \in \mathbb{R}, && \text{(Linear)}\\
f_{\boldsymbol{\theta}}(x_1^n) &\triangleq \mathbb{P}_{\boldsymbol{\theta}}(x_{n+1}=1 \mid x_1^n) = \sigma(\mathrm{logit}_n). && \text{(Prediction)}
\end{aligned}$$

Here $\boldsymbol{\theta} \triangleq (\boldsymbol{e}, \{\boldsymbol{p}_n\}_{n=1}^N, \ldots, \boldsymbol{W}_1, \boldsymbol{W}_2, b, \boldsymbol{a}) \in \mathbb{R}^D$ denotes the full list of the transformer parameters (§ B).

Loss and training. The transformer parameters $\boldsymbol{\theta}$ are trained using gradient-based methods to minimize the cross-entropy loss on the next-token prediction, L, given by

$$L(\boldsymbol{\theta}) \triangleq -\frac{1}{N}\sum_{n\in[N]} \mathbb{E}_{x_1^{n+1}}[x_{n+1}\cdot \log f_{\boldsymbol{\theta}}(x_1^n) + (1-x_{n+1})\cdot \log(1-f_{\boldsymbol{\theta}}(x_1^n))]. \tag{1}$$

Loss landscape. A key surprising observation in (Makkuva et al., 2024) is that the loss function $L(\cdot)$ admits both global and local minima depending on the switching factor $p+q$ of the Markovian data, and the weight-tying of the embedding and linear weights ($\boldsymbol{e} = \boldsymbol{a}$) of the transformer. In view of these results, we focus on the weight-tying scenario and hence let $\boldsymbol{e} = \boldsymbol{a}$ to be a single parameter in $\mathbb{R}^d$. Thus, $\boldsymbol{\theta} = (\boldsymbol{e} = \boldsymbol{a}, \{\boldsymbol{p}_n\}_{n=1}^N, \ldots, \boldsymbol{W}_1, \boldsymbol{W}_2, b)$. We interchangeably refer to $\boldsymbol{\theta}$ as both the transformer and the set of parameters.

Our objective. While the aforementioned results detail the static landscape of the loss, they do not characterize the learning dynamics on the loss surface and the effect of initialization, which plays a central role in training machine learning models (Arora et al., 2019). In view of these shortcomings, the main objective of this paper is to address the following question:

> *(Q.1): Can we explain how the initialization and learning dynamics affect the convergence of the transformer parameters to the local or global optima?*

3. Canonical low-rank parameterization

Motivation. Given the complexity of the transformer architecture and the non-convex loss function, it is challenging to analyze the learning dynamics directly (Edelman et al., 2024). To tackle this, we capitalize on the following empiri-

cal observation (§ A.3.1) which is the motivating idea behind our approach: when trained by gradient-based methods, the weight matrices $(\boldsymbol{W}_V, \ldots, \boldsymbol{W}_1, \boldsymbol{W}_2)$ at the optima $\boldsymbol{\theta}_\star$ and $\boldsymbol{\theta}_{\min}$ exhibit *rank-one* structure, whose eigenvector is the same direction in which the both the token embedding $\boldsymbol{e}$ and the positional embeddings $\boldsymbol{p}_n$ are all aligned in. Interestingly, such low-rank solutions can also be shown to be theoretically optimal (see § C). While these observations illustrate the implicit bias towards low-rank solutions at the final convergence, a natural question arises: *if we initialize with low-rank parameters, will they remain low-rank during training?* In § A.3.1, we affirmatively address this based on a thorough empirical evaluation for single-layer transformers and inspired by these empirical phenomena, without loss of generality, we restrict our attention to these low-rank manifolds to characterize the learning dynamics.

Parameterization. More specifically, we consider a special low-rank parameterization that is empirically observed and capitalize on it to address *(Q.1)*. Interestingly, along this low-rank manifold, it suffices to consider a reduced set of parameters $\boldsymbol{\theta} \in \mathbb{R}^2$ or $\boldsymbol{\theta} \in \mathbb{R}^3$ given by:

$$\boldsymbol{\theta} = (e, w) \in \mathbb{R}^2, \text{ or } \boldsymbol{\theta} = (e, w, a) \in \mathbb{R}^3. \tag{2}$$

Here e denotes the *embedding* scalar, w the *weight*, and a the *attention* parameter respectively. These parameters distill the essential role of the embedding vectors $(\boldsymbol{e}, \{\boldsymbol{p}_n\})$ in the Embedding layer, the weight matrices $(\boldsymbol{W}_1, \boldsymbol{W}_2)$ in the FF layer, and the attention matrices $\boldsymbol{W}_{K,Q,V}$ in the Attention layer respectively. We refer to § A.1 for a more detailed description. For the ease of exposition, here we let $a = 0$ and analyze the learning dynamics for $\boldsymbol{\theta} = (e, w) \in \mathbb{R}^2$ and defer the general case $\boldsymbol{\theta} \in \mathbb{R}^3$ to § A.2.

Using this parameterization $\boldsymbol{\theta} = (e, w)$ in the transformer architecture (Sec. 2) and the equivalence between the cross-entropy loss and the logistic loss $\ell_{\log}(z) = \log(1 + \exp(-z))$, $z \in \mathbb{R}$, the loss function in Eq. (1) can be compactly written as (Lemma 6):

$$\begin{aligned} L(\boldsymbol{\theta}) &= \frac{1}{N} \sum_{n \in [N]} \mathbb{E}\left[\ell_{\log}\left((2x_{n+1} - 1) \cdot \mathrm{logit}_n(\boldsymbol{\theta})\right)\right], \\ \mathrm{logit}_n(\boldsymbol{\theta}) &= \left(e^2(1 + 2w|w|)\, x_n + b_\star - \frac{e^2}{2}\right), \end{aligned} \tag{3}$$

where $b_\star(\boldsymbol{\theta}) = \mathrm{argmin}_{b \in \mathbb{R}} L(\boldsymbol{\theta}, b)$ is the optimal bias minimizing the loss for each $\boldsymbol{\theta}$ and has a closed form expression (Lemma 5). Empirically, this roughly translates to running the gradient-descent for the bias for more steps at each $\boldsymbol{\theta}$. In practice, one additional step is usually sufficient (§ A.3).

3.1. Loss landscape with canonical parameterization

With the new set of parameters $\boldsymbol{\theta} = (e, w)$, we are now ready to analyze the loss $L(\cdot)$ in Eq. (3). First we recall the definition of a critical point (Lee et al., 2016). A point $\boldsymbol{\theta}_\star$ is a critical point for L if $\nabla L(\boldsymbol{\theta}_\star) = 0$. Similarly, the standard notions of local & global minima and saddle points (§ A.1). Equipped with these definitions, Thm. 1 below characterizes the loss landscape in terms of these local and global optima.

Theorem 1 (All critical points)**.** *Let the input sequence be $\{x_n\}_{n=1}^N \sim (\boldsymbol{\pi}, \boldsymbol{P})$, the transformer parameters $\boldsymbol{\theta} = (e, w)$, and the next-token prediction loss L be as in Eq. (3). Then for any $(p, q) \in (0, 1)^2$ with $p + q \neq 1$ and $N \in \mathbb{N}$,*

(i) *$\boldsymbol{\Theta}_\star(p, q)$, the set of all global minima is given by*

$$\{(e, w) : e^2(1 + 2w|w|) = \log(1 - p)(1 - q)/pq\},$$

(ii) *$\boldsymbol{\Theta}_{\min}(p, q)$, the set of all local minima is given by*

$$\{(e, w) : e = 0,\ (p + q - 1)(1 + 2w|w|) > 0\},$$

(iii) *$\boldsymbol{\Theta}_{\max}(p, q)$, the set of all local maxima is given by*

$$\{(e, w) : e = 0,\ (p + q - 1)(1 + 2w|w|) < 0\},$$

(iv) *and $\boldsymbol{\Theta}_{\mathrm{sad}}(p, q)$, the set of all saddle points is*

$$\{(0, -1/\sqrt{2})\}.$$

Thus the set of all critical points is

$$\{\boldsymbol{\theta} : \nabla L(\boldsymbol{\theta}) = 0\} = \boldsymbol{\Theta}_\star \cup \boldsymbol{\Theta}_{\min} \cup \boldsymbol{\Theta}_{\max} \cup \boldsymbol{\Theta}_{\mathrm{sad}}.$$

In addition, for any $\boldsymbol{\theta}_\star \in \boldsymbol{\Theta}_\star$, $\boldsymbol{\theta}_{\min} \in \boldsymbol{\Theta}_{\min}$, $\boldsymbol{\theta}_{\max} \in \boldsymbol{\Theta}_{\max}$, and $\boldsymbol{\theta}_{\mathrm{sad}} \in \boldsymbol{\Theta}_{\mathrm{sad}}$, the loss values satisfy

$$\begin{aligned} H(x_{n+1} \mid x_n) = L(\boldsymbol{\theta}_\star) < L(\boldsymbol{\theta}_{\min}) = L(\boldsymbol{\theta}_{\max}) \\ = L(\boldsymbol{\theta}_{\mathrm{sad}}) = H(x_{n+1}). \end{aligned}$$

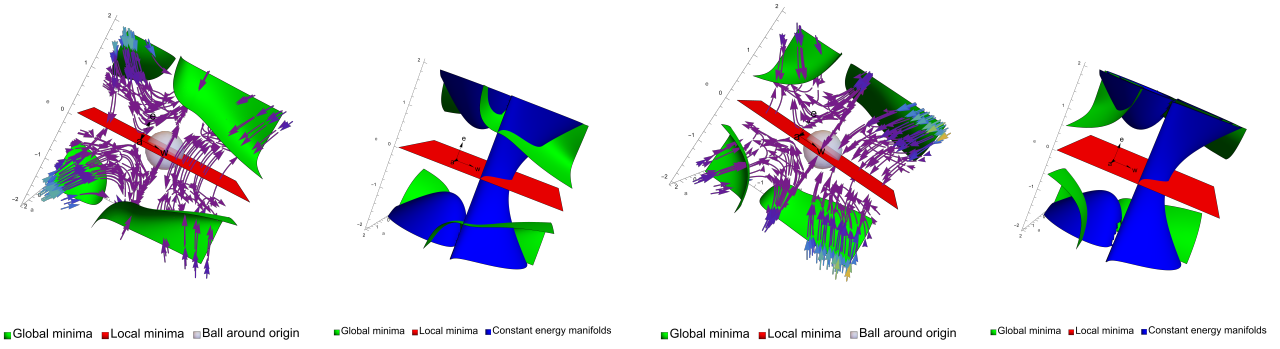
Proof. We refer to § F. □

Fig. 1 illustrates the loci of these critical points for $p + q \lessgtr 1$. Motivated by empirical observations, while (Makkuva et al., 2024) characterizes local minima for $p + q > 1$, it is interesting to note that our Thm. 1 shows that local minima also exist for $p + q < 1$ (Fig. 1a). So why did they observe the minima only for the former? The answer to this, and more broadly to question *(Q.1)* lies in the learning dynamics for $\boldsymbol{\theta}$, influenced by initialization, which we study in the next section.

4. Learning dynamics

Capitalizing on the loss landscape in Thm. 1, we now focus on the convergence of gradient-based algorithms to these critical points. Specifically, we focus on the gradient-flow of the parameters, $(\boldsymbol{\theta}_t)_{t \geq 0}$, governed by

$$\frac{\mathrm{d}\boldsymbol{\theta}_t}{\mathrm{d}t} = -\nabla L(\boldsymbol{\theta}_t), \quad \boldsymbol{\theta}_t = (e_t, w_t) \in \mathbb{R}^2,\ t \geq 0, \tag{GF}$$



(a) $p + q < 1$: Gradient flow (b) Energy manifold and minima (c) $p + q > 1$: Gradient flow (d) Energy manifold and minima

Figure 2: Gradient flow dynamics for the canonical parameters $\boldsymbol{\theta} = (e, w, a) \in \mathbb{R}^3$ with the attention scalar a (§ A.2). Notice the contrasting behavior for Gaussian initialization around origin for $p + q$ smaller and greater than one. For an enhanced view of the flow near the origin, please refer to Fig. 5.

where $\boldsymbol{\theta}_t \triangleq \boldsymbol{\theta}(t)$ is a continuously differentiable curve in $\mathbb{R}^2$ starting with a randomly initialized $\boldsymbol{\theta}_0$. To characterize these trajectories, we define an *energy function* $\mathcal{E}(\cdot, \cdot)$, which plays a crucial in the GF dynamics. It is defined as

$$\mathcal{E}(e, w) \triangleq e^2 - (w^2 + \text{sign}(w) \cdot \log |w|), \tag{4}$$

for all $(e, w) \in \mathbb{R}^2 \setminus \text{e-axis}$, where e-axis $\triangleq \{(e, w = 0)\}$. Figs. 1a and 1c illustrate these energy contour lines. The utility of the energy function is captured below.

Lemma 1 (Constant energy along the flow)**.** *For any $(p, q) \in (0, 1)^2$ and initialization $\boldsymbol{\theta}_0 = (e_0, w_0)$, let $(\boldsymbol{\theta}_t)_{t \geq 0}$ be the corresponding GF trajectory starting from $\boldsymbol{\theta}_0$. If $\boldsymbol{\theta}_0 \in \mathbb{R}^2 \setminus \text{e-axis}$, the energy stays constant along the trajectory, i.e. for all $t \geq 0$,*

$$\mathcal{E}(\boldsymbol{\theta}_t) = e_t^2 - (w_t^2 + \text{sign}(w_t) \cdot \log |w_t|) = \mathcal{E}(\boldsymbol{\theta}_0). \tag{5}$$

On the other hand, if $\boldsymbol{\theta}_0 \in \text{e-axis}$, we have $\boldsymbol{\theta}_t \in \text{e-axis}$ for all $t \geq 0$ with $w_t = w_0 = 0$.

We are now ready to present the main results of our paper. Specifically, Thm. 2 and Thm. 8, for $p + q \gtrless 1$ respectively, highlight the role of the switching factor and the parameter initialization, $\boldsymbol{\theta}_0$, in deciding whether the GF converges to local or global optima. First we define the energy $\mathcal{E}_{\text{sad}} \triangleq \mathcal{E}(e = 0, w = -1/\sqrt{2}) = -(1 + \log 2)/2$.

Theorem 2 (GF dynamics for $p + q > 1$)**.** *Let $(p, q) \in (0, 1)^2$ with $p + q > 1$, the input sequence be $\{x_n\}_{n=1}^N \sim (\boldsymbol{\pi}, \boldsymbol{P})$, and $(\boldsymbol{\theta}_t)_{t \geq 0}$ be the corresponding GF trajectory starting from $\boldsymbol{\theta}_0$. Then for all initializations $\boldsymbol{\theta}_0 \in \mathbb{R}^2$, the gradient flow converges to a critical point of the loss L. That is, there exists a $\boldsymbol{\theta}_{\text{lim}} \in \mathbb{R}^2$ such that $\lim_{t \to \infty} \boldsymbol{\theta}_t = \boldsymbol{\theta}_{\text{lim}}$ and $\nabla L(\boldsymbol{\theta}_{\text{lim}}) = 0$. In particular, $\boldsymbol{\theta}_{\text{lim}}$ is a*

(i) *a local minimum if*

$$\boldsymbol{\theta}_0 \in \mathcal{I}_{\min} \triangleq \{(e, w) : w \in (-1/\sqrt{2}, 0), \\ e \in (-g(w), g(w)),\ g(w) = \sqrt{w^2 - \log(-w) + \mathcal{E}_{\text{sad}}}\} \\ \cup \{(e, w) : w \geq 0\},$$

(ii) *a saddle point if $\boldsymbol{\theta}_0 \in \mathcal{I}_{\text{sad}} \triangleq \{(e, w) : w \in [-1/\sqrt{2}, 0),\ e = \pm\sqrt{w^2 - \log(-w) + \mathcal{E}_{\text{sad}}}\}$,*

(iii) *a local maximum if $\boldsymbol{\theta}_0 \in \mathcal{I}_{\max} \triangleq \{(e, w) : e = 0,\ w < -1/\sqrt{2}\}$,*

(iv) *and a global minimum if $\boldsymbol{\theta}_0 \in \mathcal{I}_\star \triangleq \mathbb{R}^2 \setminus (\mathcal{I}_{\min} \cup \mathcal{I}_{\text{sad}} \cup \mathcal{I}_{\max})$.*

Consequently, when $p + q > 1$, if we use the standard initialization $\boldsymbol{\theta}_0 \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_2)$ with $\sigma^2 \ll 1/\sqrt{2}$, $\boldsymbol{\theta}_{\text{lim}}$ will be a local minimum with high probability. If $p + q < 1$, under the same initialization scheme, $\boldsymbol{\theta}_{\text{lim}}$ will be a global minimum with high probability.

Key insights and conclusion. Together, Thms. 2 and 8 address our motivating question *(Q.1)* by fully characterizing the GF dynamics in terms of initialization and input data properties. Specifically, our results explain the phenomenon in (Makkuva et al., 2024) wherein they observe local minima for $p + q > 1$ more often than for $p + q < 1$, owing to standard Gaussian initialization around origin (Figs. 1b and 1d). However, in practice, we often do not know the input switching factor, raising a natural questions: *is there a data-agnostic initialization that always converges to global minima?* Indeed, as can be seen from Figs. 1b and 1d, there is a common region of initialization, $\mathcal{I}_{\text{common}} \triangleq \{(e, w) : w < 0, |e| > \sqrt{w^2 - \log(-w) + \mathcal{E}_{\text{sad}}}\}$, that leads to the global minima convergence irrespective of the switching (§ A.3.2). We believe our findings open interesting avenues of future research for GF analysis with deeper architectures and higher order Markov chains.

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.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks, 2019.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Workshop on Efficient Systems for Foundation Models @ ICML2023*, 2023.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Pritam Chandra, Tanmay Kumar Sinha, Kabir Ahuja, Ankit Garg, and Navin Goyal. Towards analyzing self-attention via linear neural network, 2024. URL https://openreview.net/forum?id=4fVuBf5HE9.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality, 2024.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2nd edition, 2006.

John M Danskin. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664, 1966.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A survey on in-context learning, 2023. URL https://arxiv.org/abs/2301.00234.

Benjamin L. Edelman, Ezra Edelman, Surbhi Goel, Eran Malach, and Nikolaos Tsilivis. The evolution of statistical induction heads: In-context learning markov chains, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Hengyu Fu, Tianyu Guo, Yu Bai, and Song Mei. What can a single attention layer learn? A study through the random features lens. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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.

M Emrullah Ildiz, Yixiao Huang, Yingcong Li, Ankit Singh Rawat, and Samet Oymak. From self-attention to markov models: Unveiling the dynamics of generative transformers. *arXiv preprint arXiv:2402.13512*, 2024.

Samy Jelassi, Michael Eli Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257. PMLR, 2016.

Yingcong Li, M. Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: generalization and stability in in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, 2023a.

Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a mechanistic understanding, 2023b.

Ashok Vardhan Makkuva, Marco Bondaschi, Adway Girish, Alliot Nagle, Martin Jaggi, Hyeji Kim, and Michael Gastpar. Attention with Markov: A framework for principled analysis of transformers via Markov chains. *arXiv preprint arXiv:2402.04161*, 2024.

Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. *arXiv preprint arXiv:2402.14735*, 2024.

Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Matteo Pagliardini. GPT-2 modular codebase implementation. *https://github.com/epfml/llm-baselines*, 2023.

Jorge Pérez, Pablo Barceló, and Javier Marinkovic. Attention is Turing-complete. *Journal of Machine Learning Research*, 22(75):1–35, 2021.

Charlie Snell, Ruiqi Zhong, Dan Klein, and Jacob Steinhardt. Approximating how single head attention learns. *CoRR*, abs/2103.07601, 2021. URL https://arxiv.org/abs/2103.07601.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Shaolei Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. In *Conference on Parsimony and Learning (Recent Spotlight Track)*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

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, 2023.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent, 2023.

Colin Wei, Yining Chen, and Tengyu Ma. Statistically meaningful approximation: a case study on approximating Turing machines with transformers. In *Advances in Neural Information Processing Systems*, volume 35, pages 12071–12083, 2022.

Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like transformers. In *International Conference on Machine Learning*, pages 11080–11090, 2021.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021. URL https://arxiv.org/abs/2111.02080.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020.

Contents

A. Supporting results to the manuscript

We provide additional theoretical and empirical details for our main results in the manuscript. First we define the notation.

Notation. We denote scalars by italic lower case letters like x, y and Euclidean vectors and matrices in bold: $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{M}$, etc. $\|\cdot\|$ denotes the ℓ_2-norm for Euclidean vectors and Frobenius norm for matrices. $[k] \triangleq \{1, \ldots, k\}$, and for a sequence $(x_n)_{n\geq 1}$, define $x_k^m \triangleq (x_k, \ldots, x_m)$ if $k \geq 1$ and $(x_1, \ldots, x_m)$ otherwise. For $z \in \mathbb{R}$, the sigmoid $\sigma(z) \triangleq 1/(1 + e^{-z})$,

$\mathrm{ReLU}(z) \triangleq \max(0, z)$ and the convex logistic loss $\ell_{\log}(z) \triangleq \log(1 + \exp(-z)) \in (0, \infty)$. For events A and B, $\mathbb{P}(A)$ denotes the probability of A whereas $\mathbb{P}(A \mid B)$ the conditional probability. Let (x, y) be a pair of discrete random variables on $[k] \times [k]$ with the probability mass function (pmf) of x being $\boldsymbol{p}_x = (p_1, \ldots, p_k) \in [0, 1]^k$. Then its Shannon entropy is defined as $H(x) = H(\boldsymbol{p}_x) \triangleq -\sum_{i \in [k]} p_i \log p_i$. The conditional entropy is defined to be $H(y|x) \triangleq H(x, y) - H(x)$. The entropy rate of a stochastic process $(x_n)_{n \geq 1}$ is defined as $\lim_{n \to \infty} H(x_1^n)/n$. We simply write $x = y$ to mean $\mathbb{P}(x = y) = 1$. We also use the shorthand $\mathbb{P}(y = j \mid x)$ for $\mathbb{P}(y = j \mid x = x)$ as a function of the random variable x. For $p \in (0, 1)$, the binary entropy function $h(\cdot)$ is defined as $h(p) \triangleq -p \log p - (1 - p) \log(1 - p)$.

A.1. Canonical low-rank parameterization

Recall from Sec. 3 that the reduced set of parameters $\boldsymbol{\theta} \in \mathbb{R}^2$ or $\boldsymbol{\theta} \in \mathbb{R}^3$ is given by:

$$\boldsymbol{\theta} = (e, w) \in \mathbb{R}^2, \text{ or } \boldsymbol{\theta} = (e, w, a) \in \mathbb{R}^3. \tag{Reparameterization}$$

Here e denotes the *embedding* scalar, w the *weight*, and a the *attention* parameter respectively. Now we describe the parameterization of the transformer vis-á-vis these scalars and refer to App. D for a more detailed descripton. Let the input $\{x_n\}_{n=1}^N$ be a first-order Markov chain as in Sec. 2 and let $n \in [N]$ be fixed. Then we have

$$\text{Embedding}: \boldsymbol{e} = e \cdot \boldsymbol{\alpha},\ \boldsymbol{p}_n = \left(-\frac{e}{2}\right) \cdot \boldsymbol{\alpha} \to \boldsymbol{x}_n = e\left(x_n - \frac{1}{2}\right)\boldsymbol{\alpha}, \quad e \in \mathbb{R}, \boldsymbol{\alpha} \in \{\pm 1\}^d/\sqrt{d},$$

$$\text{Attention}: \boldsymbol{W}_V = \boldsymbol{\alpha}\,\boldsymbol{v}^\top \to \boldsymbol{y}_n = e\left(x_n - \frac{1}{2}\right)\boldsymbol{\alpha} + \underbrace{\langle \boldsymbol{v}, \boldsymbol{\alpha} \rangle}_{\triangleq a \approx 0}\left(\sum_{i \in [n]} \mathrm{att}_{n,i} \cdot e\left(x_i - \frac{1}{2}\right)\right)\boldsymbol{\alpha}, \boldsymbol{v} \in \mathbb{R}^d.$$

The scalar a is the product of $\langle \boldsymbol{v}, \boldsymbol{\alpha} \rangle$ and the scaling in the attention weights $\mathrm{att}_{n,i}$, which is empirically close to zero for first-order Markov chains. Hence for the ease of exposition, we first omit it by letting $a = 0$ and analyze the general case when $a \in \mathbb{R}$ in App. A.2. We continue:

$$\text{FF}: \boldsymbol{W}_1 = \frac{|w|}{\sqrt{d}}\,\mathbf{1}\,\boldsymbol{\alpha}^\top, \boldsymbol{W}_2 = \frac{w}{\sqrt{d}}\,\boldsymbol{\alpha}\,\mathbf{1}^\top \to \boldsymbol{z}_n = e\left(x_n - \frac{1}{2}\right)(1 + 4w|w|x_n)\,\boldsymbol{\alpha}, w \in \mathbb{R}.$$

$\mathbf{1}$ is the all-one vector in $\mathbb{R}^r$ with $r = 4d$ typically in practice. Substituting this $\boldsymbol{z}_n$ in the linear layer with $\boldsymbol{e} = \boldsymbol{a}$ and bias $b \in \mathbb{R}$, the logits and the probabilities simplify to:

$$\text{Linear}: \mathrm{logit}_n(e, w, b) = e^2(1 + 2w|w|)\,x_n + b - \frac{e^2}{2} \in \mathbb{R}, \tag{6}$$

$$\text{Prediction}: f_{(\boldsymbol{\theta}, b)}(x_1^n) = \sigma(\mathrm{logit}_n) \in (0, 1), \quad \boldsymbol{\theta} \triangleq (e, w). \tag{7}$$

Finally, using the equivalence between the cross-entropy loss and the logistic loss $\ell_{\log}(\cdot)$, the loss function in Eq. (1) can be compactly written as (Lemma 6):

$$L(\boldsymbol{\theta}, b) = \frac{1}{N}\sum_{n \in [N]} \mathbb{E}[\ell_{\log}((2x_{n+1} - 1) \cdot \mathrm{logit}_n(\boldsymbol{\theta}))], \quad \boldsymbol{\theta} \in \mathbb{R}^2, b \in \mathbb{R}. \tag{8}$$

Due to convexity of $\ell_{\log}(\cdot)$, it follows that $L(\boldsymbol{\theta}, b)$ is convex in the bias b for any fixed $\boldsymbol{\theta}$, whose minimizer, $b_\star(\boldsymbol{\theta}) = \mathrm{argmin}_{b \in \mathbb{R}} L(\boldsymbol{\theta}, b)$, has a closed form expression (Lemma 5). Hence, without loss of generality, we consider the loss with this optimal bias $b_\star$:

$$L(\boldsymbol{\theta}) \triangleq L(\boldsymbol{\theta}, b_\star) = \frac{1}{N}\sum_{n \in [N]} \mathbb{E}\left[\ell_{\log}\left((2x_{n+1} - 1)\left(e^2(1 + 2w|w|)\,x_n + b_\star - \frac{e^2}{2}\right)\right)\right]. \tag{9}$$

Empirically, this roughly translates to running the gradient-based algorithm for the bias for more steps at each $\boldsymbol{\theta}$. In practice, one additional step is usually sufficient (see App. A.3). Eq. (9) resembles the standard logistic regression loss (Soudry et al., 2018) whose binary labels are $2x_{n+1} - 1 \in \{\pm 1\}$ and the logits given by $e^2(1 + 2w|w|)\,x_n + b_\star - e^2/2$, for each $n \in [N]$. The key difference here is that the logits are a non-linear function of the parameters (e, w) unlike in the standard setting.

Definitions of local and global optima. For the sake of completeness, we first recall the definition of a critical point (Lee et al., 2016). A point $\boldsymbol{\theta}_\star \in \mathbb{R}^2$ is a critical or a stationary point for L if $\nabla L(\boldsymbol{\theta}_\star) = 0$. A critical point $\boldsymbol{\theta}_\star$ is a *local minimum* if there exists a neighborhood U around $\boldsymbol{\theta}_\star$ such that $L(\boldsymbol{\theta}_\star) \leq L(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in U$, and a *local maximum* if $L(\boldsymbol{\theta}_\star) \geq L(\boldsymbol{\theta})$. If the neighborhood U is whole of $\mathbb{R}^2$, it is a *global minimum/maximum*. On the other hand, a critical point is a saddle point if for all neighborhoods U around $\boldsymbol{\theta}_\star$, there are $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in U$ such that $L(\boldsymbol{\theta}_1) \leq L(\boldsymbol{\theta}_\star) \leq L(\boldsymbol{\theta}_2)$.

A.2. Gradient flow with attention

In this section, the consider the attention scalar $a \in \mathbb{R}$ (Sec. 3) and study the gradient flow dynamics with the parameters $\boldsymbol{\theta} = (e, w, a) \in \mathbb{R}^3$. The parameter a captures the overall scaling from the value, key, and query components in the attention layer. Recall that the soft-max attention weights are given by $\mathrm{att}_{n,i} \propto \exp(\langle \boldsymbol{q}_n, \boldsymbol{k}_i \rangle / \sqrt{d})$, where $\boldsymbol{q}_n = \boldsymbol{W}_Q \boldsymbol{x}_n$ and $\boldsymbol{k}_i = \boldsymbol{W}_K \boldsymbol{x}_i$ are the query and key embeddings for any position $i \in [n]$. Using the low-rank structure of the query and key matrices, satisfying $\boldsymbol{W}_Q^\top \boldsymbol{W}_K = (q^2 d)\, \boldsymbol{\alpha}\boldsymbol{\alpha}^\top$ and the value matrix $\boldsymbol{W}_V = \boldsymbol{\alpha}\boldsymbol{v}^\top$ for some $q \in \mathbb{R}$ and $\boldsymbol{v} \in \mathbb{R}^d$ (App. H), and assuming linear attention $\mathrm{att}_{n,i} \propto \langle \boldsymbol{q}_n, \boldsymbol{k}_i \rangle / \sqrt{d}$, we define a single scalar $a \triangleq \langle \boldsymbol{v}, \boldsymbol{\alpha} \rangle q^2 d^{5/2}/4$ that captures the essence of the attention layer. We note that linear attention weights are a standard assumption in the transformer analysis literature (Ahn et al., 2023; von Oswald et al., 2023). Using this parameterization, similar to the steps in Sec. 3, we obtain the final loss function to be

$$L(\boldsymbol{\theta}) = \mathbb{E}\left[\ell_{\log}\left((2Y-1)\left(e^2\left[\left(X - \frac{1}{2}\right)(1 + ae^2)(1 + 2w|w|) + w|w(1+ae^2)|\right] + b_\star\right)\right)\right],$$

where $\boldsymbol{\theta} = (e, w, a)$ and $b_\star$ is the corresponding optimal bias. L recovers the loss in Eq. (9) when $a = 0$. In Thm. 10, we determine the set of all critical points of L in terms of global minima and local optima in closed-form expressions, analogous to Thm. 1. Capitalizing on this characterization, we now shift our focus to the analysis of the gradient flow in $\mathbb{R}^3$. To this end, let $(\boldsymbol{\theta}_t)_{t \geq 0}$ be a C^1 (continuously differentiable) curve in $\mathbb{R}^3$ governed by

$$\frac{\mathrm{d}\boldsymbol{\theta}_t}{\mathrm{d}t} = -\nabla L(\boldsymbol{\theta}_t), \quad \boldsymbol{\theta}_t = (e_t, w_t, a_t) \in \mathbb{R}^3,\ t \geq 0, \tag{GF-attn}$$

starting with a randomly initalized $\boldsymbol{\theta}_0$. We define the *energy function* $\mathcal{E}(\cdot, \cdot, \cdot)$ as

$$\begin{gathered}\mathcal{E}(e, w, a) \triangleq e^2 - (w^2 + \mathrm{sign}(w) \cdot \log|w|) - 2a^2, \\ \forall (e, w, a) \in \mathbb{R}^3 \setminus \text{ea-plane},\end{gathered} \tag{10}$$

where ea-plane $\triangleq \{(e, w = 0, a)\}$. It is similar to its counterpart in Eq. (4), except for the $2a^2$ term. Fig. 2 visualizes this energy surface and the set of critical points, which reveal close resemblance to that of Fig. 1 in $\mathbb{R}^2$. Capitalizing on the energy function, we now present our main result with the attention.

Theorem 3 (GF dynamics with attention)**.** *For any $(p, q) \in (0, 1)^2$ and initialization $\boldsymbol{\theta}_0 \in \mathbb{R}^3$, let $(\boldsymbol{\theta}_t)_{t \geq 0}$ be the corresponding GF-attn trajectory starting from it. Then for all $\boldsymbol{\theta}_0 \in \mathbb{R}^3$, the gradient flow converges to a critical point of the loss L. That is, there exists a $\boldsymbol{\theta}_{\lim} \in \mathbb{R}^3$ such that $\lim_{t \to \infty} \boldsymbol{\theta}_t = \boldsymbol{\theta}_{\lim}$ and $\nabla L(\boldsymbol{\theta}_{\lim}) = 0$. Further,*

(i) *if $\boldsymbol{\theta}_0 \in \mathbb{R}^3 \setminus$ ea-plane, we have $\mathcal{E}(\boldsymbol{\theta}_{\lim}) = \mathcal{E}(\boldsymbol{\theta}_t) = \mathcal{E}(\boldsymbol{\theta}_0)$ for all $t \geq 0$. Hence $\boldsymbol{\theta}_{\lim}$ is at the intersection of the energy contour line $\mathcal{E} = \mathcal{E}_0$ with that of the set of critical points.*

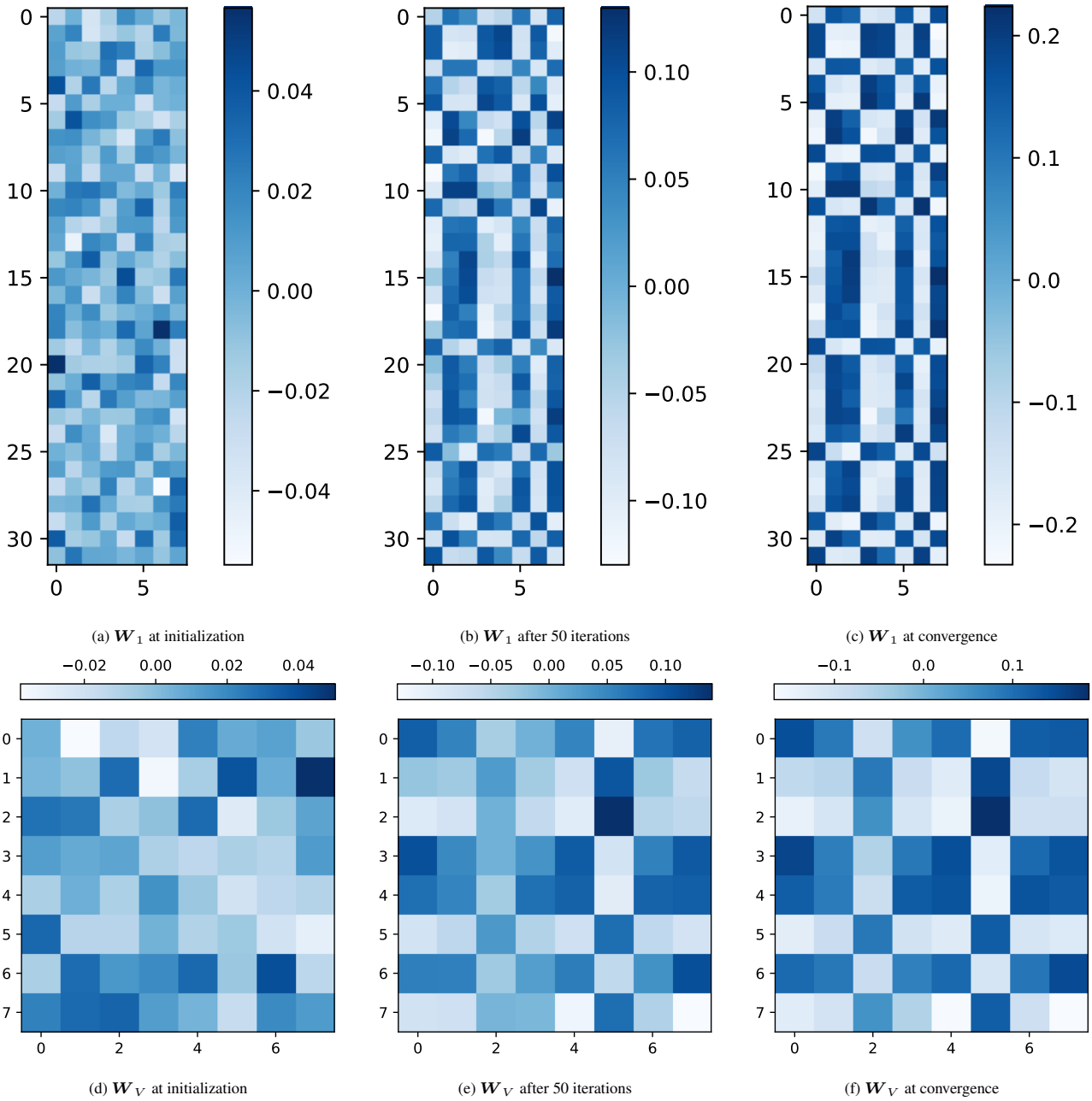
(ii) *if $\boldsymbol{\theta}_0 \in$ ea-plane, we have $\boldsymbol{\theta}_t \in$ ea-plane for all $t \geq 0$ and hence $\boldsymbol{\theta}_{\lim} \in$ ea-plane.*

Thm. 3 shows that the learning dynamics with attention closely resemble those without it (Thms. 2 and 8). While the set of all critical points of L, and thus the limit points of the flow, has a closed-form expression (Thm. 10), deriving the same for the initialization sets $\mathcal{I}_{\min}$ and $\mathcal{I}_\star$ to determine the basin of convergence is technically challenging (see discussion in App. H). Nonetheless, empirical observations with the standard Gaussian initialization around origin reveal a similar picture as in the two-dimensional setting for both the $p + q < 1$ and $p + q > 1$ cases (Fig. 2). We believe it's an interesting direction of future research to theoretically characterize this, analogous to Thms. 2 and 8. We refer to App. H for additional details and proofs.

A.3. Empirical results

We empirically validate the low-rank assumption behind our canonical parameterization and investigate our findings about local optima and initialization in the context of the full model.

A.3.1. LOW-RANK PARAMETERS



(a) $\boldsymbol{W}_1$ at initialization (b) $\boldsymbol{W}_1$ after 50 iterations (c) $\boldsymbol{W}_1$ at convergence

(d) $\boldsymbol{W}_V$ at initialization (e) $\boldsymbol{W}_V$ after 50 iterations (f) $\boldsymbol{W}_V$ at convergence

Figure 3: Evolution of parameters $\boldsymbol{W}_1$ and $\boldsymbol{W}_V$ across iterations, starting from a standard gaussian initialization. At convergence, all the parameter matrices are approximately rank-one.

Low-rank at convergence. We let the input Markov sequence to be $\{x_n\}_{n=1}^N \sim (\boldsymbol{\pi}(p,q), \boldsymbol{P}(p,q))$ for $p = 0.2, q = 0.3, N = 1024$ and consider the single-layer transformer as defined in Sec. 2 with embedding dimension $d = 8$. First, we initialize the parameters $\boldsymbol{\theta} = (\boldsymbol{e} = \boldsymbol{a}, \{\boldsymbol{p}_n\}_{n=1}^N, \ldots, \boldsymbol{W}_1, \boldsymbol{W}_2, b)$ using the standard Gaussian initialization with standard deviation 0.001 (Pagliardini, 2023) and train them using SGD on a batch size $B = 16$ and for $t = 800$ iterations. In Fig. 3, we track the value matrix $\boldsymbol{W}_V \in \mathbb{R}^{d\times d}$ and the weight matrix $\boldsymbol{W}_1 \in \mathbb{R}^{4d\times d}$ across iterations. We observe that at convergence both $\boldsymbol{W}_V$ and $\boldsymbol{W}_1$ are approximately rank-one with one of their components being same as the embedding vector (the row in $\boldsymbol{W}_V$ and column in $\boldsymbol{W}_1$). Further, the embedding vector has all entries in $\{\pm 1\}$ up to a scaling. We observe the same conclusion for other weight matrices $\boldsymbol{W}_{K,Q}, \boldsymbol{W}_2$ and for all values of $(p,q) \in (0,1)^2$.

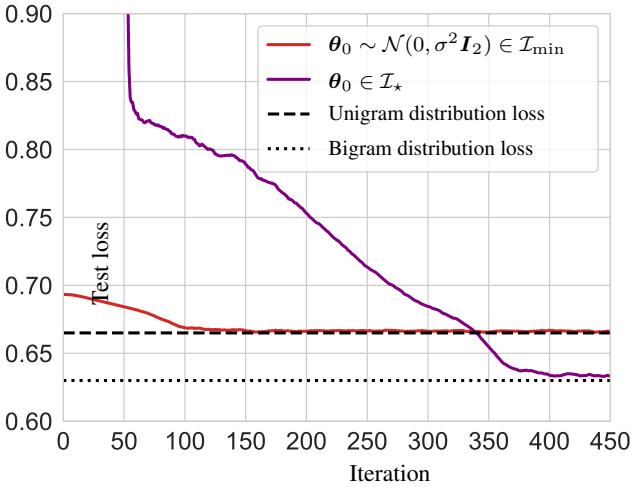


Figure 4: Comparison between the average loss curve for the standard gaussian initialization around 0 and our initialization, for $p = 0.5$ and $q = 0.8$. Starting from the standard initialization, the model converges to a local minimum corresponding to the unigram model. With our initialization, it converges to the global minimum corresponding to the bigram model.

Low-rank initialization remains low-rank during training. Inpsired by the low-rank structure obtained above, we randomly initialize the weight parameters as rank-one matrices and the embeddings on the hypercube $\{\pm 1\}^d$. After the initialization, we train them without any low-rank restrictions, and track them during the course of training. Interestingly, here we observe that the parameters still stay low-rank as illustrated in Fig. 6. A similar conclusion holds for the remaining weight matrices. Together these results provide the empirical basis for our canonical parameterization analysis in Sec. 3.

A.3.2. Effect of initialization: broader implications

Now we investigate the findings of Sec. 3 and Sec. 4, derived for the canonical low-rank model, more broadly in the context of the general single-layer transformer in Sec. 2. In particular, as shown in Thm. 2 and Fig. 1d for $p + q > 1$, any small initialization around zero would lead a local minima convergence. To test this hypothesis, we compare the standard initialization where all the transformer parameters $\boldsymbol{\theta} = (\boldsymbol{e} = \boldsymbol{a}, \{\boldsymbol{p}_n\}_{n=1}^N, \ldots, \boldsymbol{W}_1, \boldsymbol{W}_2, b)$ are randomly chosen around zero with small variance $\sigma = 0.02$, with a new initialization based on our results, where we initialize the embedding vector $\boldsymbol{e}$ such that all cordinates are equal to $e = 0.5$, $\boldsymbol{W}_1$ to be constant with the scalar $w_1 = 1$ and $\boldsymbol{W}_2$ constant with $w_2 = -1$ (corresponding to $\mathcal{I}_\star$ in Fig. 1d). We indeed observe that the final test loss matches the unigram loss for the standard initialization, while it converges to the optimal bigram loss for our initialization (see Fig. 4). Together these results indicate that though our analysis used canonical parameterization, the corresponding insights are more general and apply more broadly to the general architecture. In a similar spirit, analysis of initialization effects for deeper architectures is an interesting avenue of future research.

A.4. Related works

The recent success of transformer models in deep learning has sparked significant interest and active research in understanding them (Weiss et al., 2021; Oymak et al., 2023; Fu et al., 2023; Pérez et al., 2021; Elhage et al., 2021; Wei et al., 2022; Yun et al., 2020; Tarzanagh et al., 2023). In relation to our paper, they can be broadly classified into two topics: (i) **In-context learning (ICL):** ICL refers to the ability of transformers learn and reason from information present in their context (Chen et al., 2024; Dong et al., 2023; Akyürek et al., 2023; Von Oswald et al., 2023; Xie et al., 2021; Bai et al., 2023; Li et al., 2023a; Garg et al., 2022). Along this thread, the works most relevant to ours are (Bietti et al., 2023; Edelman et al., 2024; Nichani et al., 2024), which use Markovian input data to understand the ICL mechanism. (Bietti et al., 2023; Edelman et al., 2024) heuristically show how gradient-based updates can learn an induction-head mechanism using a simplified transformer architecture with frozen encodings, query matrices and linear activations. On the other hand, we consider the transformer in full generality including ReLU nonlinearity, capitalizng on inherent low-rank parameters, to provide a full characterization of the learning dynamics. (Nichani et al., 2024) demonstrates how two-layer transformers with GD learn induction head mechanism when the input has a causal tree dependency, such as in Markov chains. In

this work, we focus on the GF dynamics for single-layer transformers and show how they can also converge to local optima, further highlighting the role of initialization. (ii) **Training dynamics:** On the other hand, numerous works have investigated the training dynamics of transformers. For instance, (Chandra et al., 2024) examines the gradient flow in a simplified single-layer transformer, while (Tian et al., 2023) studies the process by which self-attention integrates input tokens, assuming the decoder learns faster than the attention layer. Unlike these settings, our focus is on understanding the training dynamics of the full transformer model without any simplifications. Other related works include (Snell et al., 2021), which analyzes gradient dynamics in LSTM Seq2seq models, (Jelassi et al., 2022), which shows how Vision Transformers learn spatial structures, and (Li et al., 2023b), which demonstrates that a single-layer transformer can learn a constrained topic model. A closely related work is (Ildiz et al., 2024), which shows that self-attention has a Markovian structure, but our focus is on self-attention's capability in modeling Markov chains and the associated training dynamics.

B. Single-layer transformer: architecture and results

We first describe the transformer architecture from Sec. 2:

$$\begin{aligned}
\boldsymbol{x}_n &= x_n\,\boldsymbol{e} + \boldsymbol{p}_n \in \mathbb{R}^d, && \text{(Embedding)}\\
\boldsymbol{y}_n &= \boldsymbol{x}_n + \sum_{i\in[n]} \mathrm{att}_{n,i}\cdot \boldsymbol{W}_V\,\boldsymbol{x}_i \in \mathbb{R}^d, && \text{(Attention)}\\
\boldsymbol{z}_n &= \boldsymbol{y}_n + \boldsymbol{W}_2\,\mathrm{ReLU}(\boldsymbol{W}_1\,\boldsymbol{y}_n) \in \mathbb{R}^d, && \text{(Feed-forward)}\\
\mathrm{logit}_n &= \langle \boldsymbol{a}, \boldsymbol{z}_n\rangle + b \qquad\qquad \in \mathbb{R}, && \text{(Linear)}\\
f_{\boldsymbol{\theta}}(x_1^n) &\triangleq \mathbb{P}_{\boldsymbol{\theta}}\left(x_{n+1}=1 \mid x_1^n\right) = \underbrace{\sigma(\mathrm{logit}_n)}_{\in[0,1]}. && \text{(Prediction)}
\end{aligned}$$

Here $\boldsymbol{\theta} \triangleq (\boldsymbol{e}, \{\boldsymbol{p}_n\}_{n=1}^N, \ldots, \boldsymbol{W}_1, \boldsymbol{W}_2, b, \boldsymbol{a}) \in \mathbb{R}^D$ denotes the full list of the transformer parameters from the embedding layer till the linear layer. In the attention layer, the weight assigned to each value, $\mathrm{att}_{n,i}$, is computed by a compatibility function of the query vector $\boldsymbol{q}_n \triangleq \boldsymbol{W}_Q\,\boldsymbol{x}_n$ and the corresponding key vectors $\boldsymbol{k}_i \triangleq \boldsymbol{W}_K\,\boldsymbol{x}_i$ for all $i \in [n]$. More precisely, $\mathrm{att}_{n,i} \triangleq \mathrm{softmax}((\langle \boldsymbol{q}_n, \boldsymbol{k}_1\rangle, \ldots, \langle \boldsymbol{q}_n, \boldsymbol{k}_n\rangle)/\sqrt{d})_i$. $\boldsymbol{W}_{K,Q,V} \in \mathbb{R}^{d\times d}$ are the respective key, query, and value matrices. For multi-headed attention, the same operation is performed on multiple parallel heads, whose outputs are additively combined.

Finally, the transformer parameters $\boldsymbol{\theta} \triangleq (\boldsymbol{e}, \{\boldsymbol{p}_n\}_{n=1}^N, \ldots, b, \boldsymbol{a})$ are trained via the cross-entropy loss on the next-token prediction:

$$L(\boldsymbol{\theta}) \triangleq -\frac{1}{N}\sum_{n\in[N]} \mathbb{E}_{x_1^{n+1}}[x_{n+1}\cdot \log f_{\boldsymbol{\theta}}(x_1^n) + (1-x_{n+1})\cdot \log(1-f_{\boldsymbol{\theta}}(x_1^n))]. \tag{11}$$

In this paper, we focus on the weight-tied scenario where $\boldsymbol{e} = \boldsymbol{a}$. Hence we let them be a single parameter with $\boldsymbol{\theta} = (\boldsymbol{e} = \boldsymbol{a}, \{\boldsymbol{p}_n\}_{n=1}^N, \ldots, b) \in \mathbb{R}^D$, where D is the total parameter dimensionality.

B.1. Loss landscape results

Now we recall the theoretical results from (Makkuva et al., 2024) about the loss landscape of L in the form of global and local minima.

Theorem 4 (Global minimum). *Let the input sequence be $\{x_n\}_{n=1}^N \sim (\boldsymbol{\pi}(p,q), \boldsymbol{P}(p,q))$ for some fixed $(p,q) \in (0,1)^2$. Then for all (p,q), there exists a $\boldsymbol{\theta}_\star \in \mathbb{R}^D$ with an explicit construction such that it is a global minimum for the population loss $L(\cdot)$ in Eq. (11), i.e.*

(i) *$L(\boldsymbol{\theta}) \geq L(\boldsymbol{\theta}_\star)$ for all $\boldsymbol{\theta} \in \mathbb{R}^D$.*

Further, $\boldsymbol{\theta}_\star$ satisfies:

(ii) *$\mathbb{P}_{\boldsymbol{\theta}_\star}\left(x_{n+1}=1 \mid x_1^n\right) = \mathbb{P}\left(x_{n+1}=1 \mid x_n\right)$, the Markov kernel.*
(iii) *$L(\boldsymbol{\theta}_\star) = H(x_{n+1}|x_n)$, the entropy rate of the Markov chain.*
(iv) *$\nabla L(\boldsymbol{\theta}_\star) = 0$, i.e. $\boldsymbol{\theta}_\star$ is a stationary point.*

Let $L_\star \triangleq L(\boldsymbol{\theta}_\star)$ be the global minimal loss from Thm. 4. Now we recall the result on the bad local minimum.

Theorem 5 (Bad local minimum). *Let the input sequence be $\{x_n\}_{n=1}^N \sim (\boldsymbol{\pi}(p,q), \boldsymbol{P}(p,q))$ for some fixed $(p,q) \in (0,1)^2$. If $p+q>1$, there exists an explicit $\boldsymbol{\theta}_{\min} \in \mathbb{R}^D$ such that it is a bad local minimum for the loss $L(\cdot)$, i.e.*

(i) *there exists a neighborhood $\mathcal{B}(\boldsymbol{\theta}_{\min}, r)$ with $r>0$ such that $L(\boldsymbol{\theta}) \geq L(\boldsymbol{\theta}_{\min})$ for all $\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_{\min}, r)$, with $L(\boldsymbol{\theta}_{\min}) > L_\star$.*

Further, $\boldsymbol{\theta}_{\min}$ satisfies:

(ii) *$\mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{\pi}}}\left(x_{n+1}=1 \mid x_1^n\right) = \mathbb{P}\left(x_{n+1}=1\right) = \pi_1$, the marginal distribution.*
(iii) *$L(\boldsymbol{\theta}_{\min}) = H(x_{n+1}) = H(\boldsymbol{\pi})$, the entropy of the marginal.*
(iv) *$\nabla L(\boldsymbol{\theta}_{\min}) = 0$, i.e. $\boldsymbol{\theta}_{\min}$ is a stationary point.*

C. Low-rank structure of the optima

Here we recall the low-rank structure for the global minima found by SGD consistently across multiple runs when $p+q<1$ (**?**)Appendix C.2]makkuva2024attention. In particular, it is observed that the token and positional encodings point in the same direction $\boldsymbol{\alpha}$, which is a low-rank factor for the weight matrices in the attention and the feedforward layers, which in turn are all rank-one. Mathematically,

Embedding. The embedding vector $\boldsymbol{e}$ obeys

$$\boldsymbol{e} = e \cdot \boldsymbol{\alpha}$$

for some $e>0$ and $\boldsymbol{\alpha} \in \{\pm 1\}^d$. Further, the positional embeddings $\boldsymbol{p}_n$ are constant across positions n pointing in the same direction albeit with a negative scalar, i.e.

$$\boldsymbol{p}_n = -p \cdot \boldsymbol{\alpha}$$

for $p>0$ and $p \approx \frac{e}{2}$ such that $e>p$. Thus from Embedding layer,

$$\boldsymbol{x}_n = (ex_n - p) \cdot \boldsymbol{\alpha}, \tag{12}$$

which ensures that the respective embeddings for the bit $x_n=0$ and $x_n=1$ are $\boldsymbol{x}_n = -p \cdot \boldsymbol{\alpha}$ and $\boldsymbol{x}_n = (e-p)\cdot\boldsymbol{\alpha}$, which are roughly anti-podal.

Attention. Recall from the Attention layer that the output $\boldsymbol{y}_n$ is given by $\boldsymbol{y}_n = \boldsymbol{x}_n + \boldsymbol{W}_O \sum_{i\in[n]} \mathrm{att}_{n,i} \cdot \boldsymbol{W}_V \boldsymbol{x}_i$, where the attention weights $\mathrm{att}_{n,i}$ are computed according to $\mathrm{att}_{n,i} = \exp\left(\langle \boldsymbol{q}_n, \boldsymbol{k}_i\rangle/\sqrt{d}\right) / \left(\sum_{j\in[n]} \exp\left(\langle \boldsymbol{q}_n, \boldsymbol{k}_j\rangle/\sqrt{d}\right)\right)$ with $\boldsymbol{q}_n = \boldsymbol{W}_Q \boldsymbol{x}_n$ and $\boldsymbol{k}_i = \boldsymbol{W}_K \boldsymbol{x}_i$. Here it is observed that the matrix products are all rank-one with $\boldsymbol{\alpha}$ being a factor, i.e.

$$\begin{aligned}
\boldsymbol{W}_O \boldsymbol{W}_V &= \boldsymbol{\alpha} \cdot \boldsymbol{v}^\top \in \mathbb{R}^{d\times d}, \quad \text{for some } \boldsymbol{v} \in \mathbb{R}^d, \\
\boldsymbol{W}_Q^\top \boldsymbol{W}_K &= (q^2 d)\, \boldsymbol{\alpha} \cdot \boldsymbol{\alpha}^\top \in \mathbb{R}^{d\times d}, \quad \text{for some } q \in \mathbb{R}.
\end{aligned}$$

Hence,

$$\boldsymbol{W}_V \boldsymbol{x}_i = \langle \boldsymbol{v}, \boldsymbol{\alpha}\rangle (ex_i - p)\, \boldsymbol{\alpha},$$

and

$$\begin{aligned}
\frac{\langle \boldsymbol{q}_n, \boldsymbol{k}_i\rangle}{\sqrt{d}} &= \frac{1}{\sqrt{d}} \cdot \boldsymbol{x}_n^\top \boldsymbol{W}_Q^\top \boldsymbol{W}_K \boldsymbol{x}_n = \frac{q^2 d}{\sqrt{d}} \cdot (\boldsymbol{x}_n^\top \boldsymbol{\alpha})(\boldsymbol{x}_i^\top \boldsymbol{\alpha}) \overset{(\|\boldsymbol{\alpha}\|^2 = d)}{=} \frac{q^2 d^3}{\sqrt{d}} \cdot (ex_n - p)(ex_i - p) \\
&= q^2 d^{5/2} \cdot (ex_n - p)(ex_i - p).
\end{aligned}$$

Thus,

$$\begin{aligned}
\boldsymbol{y}_n &= \boldsymbol{x}_n + \sum_{i\in[n]} \mathrm{att}_{n,i} \cdot \boldsymbol{W}_O \boldsymbol{W}_V \boldsymbol{x}_i \\
&= (ex_i - p)\, \boldsymbol{\alpha} + \sum_{i\in[n]} \mathrm{att}_{n,i} \cdot \langle \boldsymbol{v}, \boldsymbol{\alpha}\rangle (ex_i - p)\, \boldsymbol{\alpha} \\
&= \left((ex_n - p) + \langle \boldsymbol{v}, \boldsymbol{\alpha}\rangle \sum_{i\in[n]} \frac{\exp\left(q^2 d^{5/2}\,(ex_n - p)(ex_i - p)\right)}{\sum_{j\in[n]} \exp\left(q^2 d^{5/2}\,(ex_n - p)(ex_j - p)\right)} \cdot (ex_i - p) \right) \boldsymbol{\alpha}.
\end{aligned} \tag{13}$$

It is further noticed that $\langle \boldsymbol{v}, \boldsymbol{\alpha}\rangle \approx 0$ and hence $\boldsymbol{y}_n = (ex_n - p)\,\boldsymbol{\alpha} = \boldsymbol{x}_n$.

Feed-forward. For the Feed-forward layer, both the matrices $\boldsymbol{W}_1 \in \mathbb{R}^{r\times d}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{d\times r}$ exhibit rank-one structure with $\boldsymbol{\alpha}$ being one of the factors,

$$\boldsymbol{W}_1 = w \cdot \boldsymbol{w} \cdot \boldsymbol{\alpha}^\top, \quad \text{for some } \boldsymbol{w} \in \{\pm 1\}^r, w>0, \tag{14}$$

$$\boldsymbol{W}_2 = \boldsymbol{W}_1^\top. \tag{15}$$

Thus $\boldsymbol{W}_1\boldsymbol{y}_n = dw(ex_n - p)\,\boldsymbol{w}$. Since $-p < 0$ and $e - p > 0$, corresponding to $x_n = 0$ and $x_n = 1$ respectively, we obtain $\mathrm{ReLU}(\boldsymbol{W}_1\boldsymbol{y}_n) = dw\left((1-x_n)p \cdot \mathrm{ReLU}(-\boldsymbol{w}) + x_n(e-p)\cdot \mathrm{ReLU}(\boldsymbol{w})\right)$. Denoting the number of ones in $\boldsymbol{w}$ as β, i.e. $\beta = \sum_{i=1}^{r} \mathbb{1}(w_i = 1)$, we further simplify:

$$\begin{aligned}
\boldsymbol{W}_2\mathrm{ReLU}(\boldsymbol{W}_1\boldsymbol{y}_n) &= \boldsymbol{W}_1^\top \mathrm{ReLU}(\boldsymbol{W}_1\boldsymbol{y}_n) \\
&= w^2 d\left((1-x_n)p \cdot \langle \boldsymbol{w}, \mathrm{ReLU}(-\boldsymbol{w})\rangle + x_n(e-p)\cdot \langle \boldsymbol{w}, \mathrm{ReLU}(\boldsymbol{w})\rangle\right)\boldsymbol{\alpha} \\
&= w^2 d\left((1-x_n)p \cdot (\beta - r) + x_n(e-p)\cdot \beta\right)\boldsymbol{\alpha} \\
&= w^2 d(ex_n - p)\left((2\beta - r)x_n + r - \beta\right)\boldsymbol{\alpha}.
\end{aligned}$$

Hence

$$\boldsymbol{z}_n = \boldsymbol{y}_n + \boldsymbol{W}_2\mathrm{ReLU}(\boldsymbol{W}_1\boldsymbol{y}_n) = (ex_n - p)\left(1 + w^2 d\left((2\beta - r)x_n + r - \beta\right)\right)\boldsymbol{\alpha}. \tag{16}$$

Linear. Using the fact that $\boldsymbol{e} = \boldsymbol{a} = e\cdot\boldsymbol{\alpha}$ due to weight-tying, we obtain from Linear layer that

$$\mathrm{logit}_n = \langle \boldsymbol{e}, \boldsymbol{z}_n\rangle + b = ed(ex_n - p)\left(1 + w^2 d\left((2\beta - r)x_n + r - \beta\right)\right) + b. \tag{17}$$

Prediction. We finally obtain that the prediction probability

$$f_{\boldsymbol{\theta}}(x_1^n) = \sigma(\mathrm{logit}_n) = x_n \cdot \sigma\left(ed(e-p)\left(1 + \beta w^2 d\right) + b\right) + (1 - x_n)\cdot \sigma\left(-edp\left(1 + (r-\beta)w^2 d\right) + b\right).$$

Thus we see that the prediction probability and hence the loss function $L(\cdot)$ in Eq. (11) is influenced only by the scalars e, p, w, b and β.

D. Canonical reparameterization

Building on the low-rank strucutre of the transformer parameters described above, we consider a special parameterization for them. A key property of this parameterization is that it covers both the global and local minima from Thm. 4 and Thm. 5 for all $(p, q) \in (0, 1)^2$. Recall that Thm. 5 characterizes local minima only for $p + q > 1$ whereas our special parameterization allows to discover local minima even for $p + q < 1$. Our construction follows the same outline as in Eqs. (12)-(17). First we start with the embedding layer.

Embedding. We let $\boldsymbol{e} = e\cdot\boldsymbol{\alpha}$ and $\boldsymbol{p}_n = -p\cdot\boldsymbol{\alpha}$ for all n where $e > 0, p = \frac{e}{2}$ and $\boldsymbol{\alpha} \in \{\pm 1\}^d/\sqrt{d}$. Thus the embedding $\boldsymbol{x}_n$ from Eq. (12) simplifies to

$$\boldsymbol{x}_n = e\left(x_n - \frac{1}{2}\right)\boldsymbol{\alpha} \in \{\pm\frac{e}{2}\}\,\boldsymbol{\alpha}. \tag{18}$$

Attention. Substituting this $\boldsymbol{x}_n$ in Eq. (13), we have

$$\boldsymbol{y}_n = e\left(x_n - \frac{1}{2}\right)\boldsymbol{\alpha} + \langle \boldsymbol{v}, \boldsymbol{\alpha}\rangle\left(\sum_{i\in[n]} \mathrm{att}_{n,i}\cdot e\left(x_i - \frac{1}{2}\right)\right)\boldsymbol{\alpha}, \tag{19}$$

where the attention weights $\mathrm{att}_{n,i} = \frac{\exp\left(e^2 q^2 d^{5/2}\,(x_n - \frac{1}{2})(x_i - \frac{1}{2})\right)}{\sum_{j\in[n]}\exp\left(e^2 q^2 d^{5/2}\,(x_n - \frac{1}{2})(x_j - \frac{1}{2})\right)} \in (0, 1)$ for some $q \in \mathbb{R}$. Since $\langle \boldsymbol{v}, \boldsymbol{\alpha}\rangle \approx 0$, we let $\boldsymbol{v} = 0$ and obtain

$$\boldsymbol{y}_n = \boldsymbol{x}_n = e\left(x_n - \frac{1}{2}\right)\boldsymbol{\alpha}. \tag{20}$$

Feed-forward. For the feed-forward layer, we observe from Eq. (14) and Eq. (16) that for any $\boldsymbol{w} \in \{\pm 1\}^r$, only the number of 1's in $\boldsymbol{w}$, β, matters for the final vector $\boldsymbol{z}_n$ which further interacts with the weight scalar w. Hence without loss of generality, we set $\boldsymbol{w}$ to be the all-ones vector: $\boldsymbol{w} = \mathbf{1} \in \mathbb{R}^r$ and hence $\beta = r = 4d$. While we observe from Eq. (14) that $\boldsymbol{W}_2 = \boldsymbol{W}_1^\top$ for $p + q < 1$, we observe from the proof of the Thm. 4 for $p + q > 1$ in (Makkuva et al., 2024)-Appendix

B.2 that we need $\boldsymbol{W}_2 = -\boldsymbol{W}_1^\top$ in this scenario. Hence we consider the following parameterization that covers both these scenarios:

$$\boldsymbol{W}_1 = \frac{|w|}{\sqrt{d}}\,\mathbf{1}\cdot\boldsymbol{\alpha}^\top \in \mathbb{R}^{4d\times d}, \quad \boldsymbol{W}_2 = \frac{w}{\sqrt{d}}\,\boldsymbol{\alpha}\cdot\mathbf{1}^\top \in \mathbb{R}^{d\times 4d}.$$

Here $w > 0$ ensures $\boldsymbol{W}_2 = \boldsymbol{W}_1^\top$ whereas $w < 0$, $\boldsymbol{W}_2 = -\boldsymbol{W}_1^\top$. Using this parameterization, substituting $\beta = r = 4d$ and $w \mapsto \frac{w}{d}$ in Eq. (16), we get

$$\boldsymbol{z}_n = e\left(x_n - \frac{1}{2}\right)(1 + 4w|w|x_n)\,\boldsymbol{\alpha}. \tag{21}$$

Linear. Since $\boldsymbol{e} = \boldsymbol{a} = e\cdot\boldsymbol{\alpha}$ due to weight-tying, Eq. (17) simplifies to

$$\text{logit}_n = \langle \boldsymbol{e}, \boldsymbol{z}_n\rangle + b = e^2\left(x_n - \frac{1}{2}\right)(1 + 4w|w|x_n) + b \tag{22}$$

$$\overset{(x_n = x_n^2)}{=} e^2\left(x_n + 4w|w|x_n - \frac{1}{2} - 2w|w|x_n\right) + b \tag{23}$$

$$= e^2(1 + 2w|w|)\,x_n + b - \frac{e^2}{2}. \tag{24}$$

Prediction. The next-token prediction probability is

$$f_{(\boldsymbol{\theta},b)}(x_1^n) = \sigma\left(e^2(1 + 2w|w|)\,x_n + b - \frac{e^2}{2}\right), \quad \boldsymbol{\theta} \triangleq (e, w) \in \mathbb{R}^2. \tag{25}$$

Loss. While we assumed $e > 0$ in the beginning, in view of Eq. (25) and the fact that $\boldsymbol{\alpha} \in \{\pm 1\}^d/\sqrt{d}$, we see that $e \in \mathbb{R}$ gives us the same expression for probability. Thus the final probability depends on just the three scalars $(e, w, b) \in \mathbb{R}^3$. Defining $\boldsymbol{\theta} = (e, w) \in \mathbb{R}^2$, we recall the cross-entropy loss $L(\cdot)$ from Eq. (8) in Sec. 2 for this canonical model:

$$L(\boldsymbol{\theta}, b) = -\frac{1}{N}\sum_{n\in[N]} \mathbb{E}_{x_1^{n+1}}[x_{n+1}\cdot\log f_{(\boldsymbol{\theta},b)}(x_1^n) + (1 - x_{n+1})\cdot\log(1 - f_{(\boldsymbol{\theta},b)}(x_1^n))]. \tag{26}$$

It turns out that we can further remove the bias b by minimizing the loss over it which we discuss in App. F. For now in the next section, we analyze when it's present as in Eq. (26).

E. Analysis of the loss with the bias, $L(\boldsymbol{\theta}, b)$, in Eq. (8) and Eq. (26)

In this section, we analyze the loss function with the bias, $L(\boldsymbol{\theta}, b)$, from Eq. (8) and Eq. (26), which will later be useful for studying $L(\boldsymbol{\theta})$. First we characterize the set of its critical points in $\mathbb{R}^3$. To this end, we define the following sets of points

$$\boldsymbol{\Gamma}_\star(p,q) \triangleq \left\{(e,w,b) \in \mathbb{R}^3 : e^2(1+2w|w|) = \log\frac{(1-p)(1-q)}{pq},\ b - \frac{e^2}{2} = \log\frac{p}{1-p}\right\}, \tag{27}$$

$$\boldsymbol{\Gamma}_{\min}(p,q) \triangleq \left\{(e,w,b) \in \mathbb{R}^3 : e = 0, (p+q-1)(1+2w|w|) > 0, b = \log\frac{p}{q}\right\}, \tag{28}$$

$$\boldsymbol{\Gamma}_{\text{sad}}(p,q) \triangleq \left\{(e,w,b) \in \mathbb{R}^3 : e = 0, (p+q-1)(1+2w|w|) \leq 0, b = \log\frac{p}{q}\right\}. \tag{29}$$

The following result establishes that these sets exhaust all the critical points.

Theorem 6 (All critical points)**.** *Let the input sequence be $\{x_n\}_{n=1}^N \sim (\boldsymbol{\pi}, \boldsymbol{P})$, the transformer parameters $(\boldsymbol{\theta}, b) = (e, w, b) \in \mathbb{R}^3$, and the next-token prediction loss $L(\cdot)$ be as in Eq. (26). Then all the stationary points of L are either in $\boldsymbol{\Gamma}_\star$, $\boldsymbol{\Gamma}_{\min}$, or $\boldsymbol{\Gamma}_{\text{sad}}$, i.e.*

$$\left\{(\boldsymbol{\theta}, b) \in \mathbb{R}^3 : \nabla L(\boldsymbol{\theta}, b) = 0\right\} = \boldsymbol{\Gamma}_\star \cup \boldsymbol{\Gamma}_{\min} \cup \boldsymbol{\Gamma}_{\text{sad}}. \tag{30}$$

Proof. We refer to App. I.1. □

Recall the definitions of local minima & maxima, global minima, and that of all the saddle points from Sec. 3.1. We are now ready to present the main result about the loss landscape of $L(\cdot)$.

Theorem 7 (Loss landscape with bias)**.** *Let the input sequence be $\{x_n\}_{n=1}^N \sim (\boldsymbol{\pi}, \boldsymbol{P})$, the transformer parameters $(e, w, b) \in \mathbb{R}^3$, and the next-token prediction loss $L(\cdot)$ be as in Eq. (26). Then for any $(p, q) \in (0, 1)^2$ with $p + q \neq 1$ and $N \in \mathbb{N}$,*

(i) *the set of all global minima of L is given by $\boldsymbol{\Gamma}_\star(p, q)$,*
(ii) *the set of all bad local minima of L is given by $\boldsymbol{\Gamma}_{\min}(p, q)$,*
(iii) *and the set of all saddle points of L is $\boldsymbol{\Gamma}_{\text{sad}}(p, q)$.*

Furthermore, for any $\boldsymbol{\gamma}_\star \in \boldsymbol{\Gamma}_\star, \boldsymbol{\gamma}_{\min} \in \boldsymbol{\Gamma}_{\min}$, and $\boldsymbol{\gamma}_{\text{sad}} \in \boldsymbol{\Gamma}_{\text{sad}}$, the losses are ordered as

$$H(x_{n+1} \mid x_n) = L(\boldsymbol{\gamma}_\star) < L(\boldsymbol{\gamma}_{\min}) = L(\boldsymbol{\gamma}_{\text{sad}}) = H(x_{n+1}).$$

Remark 1. Note that a bad local minimum is a local minimum whose loss value is strictly less than that of the global minimum, as is the case here. Interestingly, Thm. 7 highlights that all local minima for the loss L are indeed bad local minima.

Proof. We refer to App. I.2. □

E.1. Technical lemmas

The proofs of both Thm. 6 and Thm. 7 rely on few key lemmas that we present below. First we start with the result that rewrites the loss $L(\boldsymbol{\theta}, b)$ from Eq. (26) in a compact manner using the logistic function $\ell_{\log}(\cdot)$.

Lemma 2 (Loss as a logistic function)**.** *The next-token prediction loss $L(\cdot)$ in Eq. (26) can be written as*

$$\begin{aligned} L(\boldsymbol{\theta}, b) &= \frac{1}{N}\sum_{n\in[N]} \mathbb{E}[\ell_{\log}\left((2x_{n+1}-1)\cdot \text{logit}_n\right)] \\ &= \mathbb{E}_{X,Y}\left[\ell_{\log}\left((2Y-1)\left(e^2(1+2w|w|)X + b - \frac{e^2}{2}\right)\right)\right], \end{aligned} \tag{31}$$

where $(X, Y) \in \{0, 1\}^2$ are distributed according to $(X, Y) \sim (\boldsymbol{\pi}, \boldsymbol{P})$, i.e. X is a Bernoulli random variable with $X \sim \boldsymbol{\pi} \equiv \text{Bern}(p/(p+q))$ and $Y|X \sim \boldsymbol{P}(p, q)$, the Markov kernel.

The following lemma establishes the gradients of the loss function with respect to the parameters e, w, and b.

Lemma 3 (Gradient computation)**.** *For any $(e, w, b) \in \mathbb{R}^3$ and the next-token prediction loss $L(\cdot)$ in Eq. (26), the gradients are given by*

$$\frac{\partial L}{\partial e} = \mathbb{E}_X \left[(f_1 X + f_2)(2X(1 + 2w|w|) - 1) \right] \cdot e,$$
$$\frac{\partial L}{\partial w} = \mathbb{E}_X \left[(f_1 X + f_2) X \right] \cdot 4e^2 |w|,$$
$$\frac{\partial L}{\partial b} = \mathbb{E}_X \left[f_1 X + f_2 \right],$$

where $X \in \{0, 1\}$ is a Bernoulli random variable with $X \sim \mathrm{Bern}(p/(p+q))$, $f_1 = \sigma\left(2e^2 w|w| + b + \frac{e^2}{2}\right) + q - 1 - \sigma\left(b - \frac{e^2}{2}\right) + p$, and $f_2 = \sigma\left(b - \frac{e^2}{2}\right) - p$.

Remark 2. It is interesting to note that the gradients for both e and w are product of an expectation term and an e factor. Also, except for scaling factors in terms of (e, w, b), all the gradients are governed by the two expectation terms $\mathbb{E}[(f_1 X + f_2)X]$ and $\mathbb{E}[f_1 X + f_2]$. This observation plays a key role in obtaining an ordinary differential equation which yields the energy function $\mathcal{E}$, defined in Eq. (4).

Now we characterize the Hessian at both local-minima and saddle points.

Lemma 4 (Hessian at local-minima and saddle points)**.** *For the canonical parameterization $\boldsymbol{\gamma} = (b, e, w) \in \mathbb{R}^3$ and the next-token prediction loss $L(\cdot)$ in Eq. (26), the Hessian at any $\boldsymbol{\gamma}_{\min} \in \boldsymbol{\Gamma}_{\min}$ or $\boldsymbol{\gamma}_{\mathrm{sad}} \in \boldsymbol{\Gamma}_{\mathrm{sad}}$ is given by*

$$\nabla^2 L(\boldsymbol{\gamma})\Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_{\min}, \boldsymbol{\gamma}_{\mathrm{sad}}} = \pi_0 \pi_1 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2(p+q-1)(1+2w|w|) & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where $\pi_0 = \frac{q}{p+q}$ and $\pi_1 = \frac{p}{p+q}$.

Remark 3. We note that the Hessian is computed with the parameter ordering (b, e, w).

The proofs of the lemmas are presented in App. J.

F. Analysis of the loss without bias, $L(\boldsymbol{\theta})$, and proof of Thm. 1

The proof of Thm. 1, concerning the loss $L(\boldsymbol{\theta})$ in Eq. (9), is similar to that of Thm. 7 which studies the loss $L(\boldsymbol{\theta}, b)$ with the bias present. The main idea is to establish the analogous set of lemmas, as in App. E, when the bias is substituted with its optimal choice. First we recall the loss function

$$L(\boldsymbol{\theta}) \triangleq L(\boldsymbol{\theta}, b_\star) = -\frac{1}{N} \sum_{n \in [N]} \mathbb{E}_{x_1^{n+1}}[x_{n+1} \cdot \log f_{(\boldsymbol{\theta}, b_\star)}(x_1^n) + (1 - x_{n+1}) \cdot \log(1 - f_{(\boldsymbol{\theta}, b_\star)}(x_1^n))], \tag{32}$$
$$b_\star = \operatorname*{argmin}_{b \in \mathbb{R}} L(\boldsymbol{\theta}, b).$$

We start with the result that establishes a closed form expression for $b_\star$.

Lemma 5 (Optimal bias). *For $\boldsymbol{\theta} = (e, w) \in \mathbb{R}^2$ and $b \in \mathbb{R}$, let $L(\boldsymbol{\theta}, b)$ be the next-token prediction loss defined in Eq. (26). Then, for any $\boldsymbol{\theta} \in \mathbb{R}^2$, $L(\boldsymbol{\theta}, b)$ is convex in b and the minimizer $b_\star \triangleq \operatorname{argmin}_{b \in \mathbb{R}} L(\boldsymbol{\theta}, b)$ is given by*

$$\exp\left(b_\star - \frac{e^2}{2}\right) = \frac{1}{2A}\left[\frac{p}{q} - 1 + \sqrt{\left(\frac{p}{q} - 1\right)^2 + 4 \cdot \frac{p}{q} \cdot A}\right], \quad A \triangleq \exp(e^2(1 + 2w|w|)). \tag{33}$$

Consequently, if $e^2(1 + 2w|w|) = \log \frac{(1-p)(1-q)}{pq}$, then $b_\star - \frac{e^2}{2} = \log \frac{p}{1-p}$. If $e = 0$, then $b_\star = \log \frac{p}{q}$.

Now we rewrite the loss in terms of the logistic function.

Lemma 6 (Loss as a logistic function). *For any $\boldsymbol{\theta} \in \mathbb{R}^2$, the next-token prediction loss $L(\boldsymbol{\theta})$ in Eq. (32) can be written as*

$$\begin{aligned} L(\boldsymbol{\theta}) &= \frac{1}{N} \sum_{n \in [N]} \mathbb{E}[\ell_{\log}((2x_{n+1} - 1) \cdot \text{logit}_n)] \\ &= \mathbb{E}_{X,Y}\left[\ell_{\log}\left((2Y - 1)\left(e^2(1 + 2w|w|)X + b_\star - \frac{e^2}{2}\right)\right)\right]. \end{aligned} \tag{34}$$

where $b_\star$ follows from Eq. (33), $(X, Y) \in \{0, 1\}^2$ are distributed according to $(X, Y) \sim (\boldsymbol{\pi}, \boldsymbol{P})$, i.e. X is a Bernoulli random variable with $X \sim \boldsymbol{\pi} \equiv \text{Bern}(p/(p + q))$ and $Y|X \sim \boldsymbol{P}(p, q)$, the Markov kernel.

The following lemma establishes the gradients of the loss.

Lemma 7 (Gradient computation). *For any $\boldsymbol{\theta} = (e, w) \in \mathbb{R}^2$ and the next-token prediction loss $L(\boldsymbol{\theta})$ in Eq. (32), the gradients are given by*

$$\begin{aligned} \frac{\partial L}{\partial e} &= \mathbb{E}_X[(f_1 X + f_2)X] \cdot 2(1 + 2w|w|)e, \\ \frac{\partial L}{\partial w} &= \mathbb{E}_X[(f_1 X + f_2)X] \cdot 4e^2|w|, \end{aligned}$$

where $X \in \{0, 1\}$ is a Bernoulli random variable with $X \sim \text{Bern}(p/(p + q))$, $f_1 = \sigma\left(2e^2 w|w| + b_\star + \frac{e^2}{2}\right) + q - 1 - \sigma\left(b_\star - \frac{e^2}{2}\right) + p$, and $f_2 = \sigma\left(b_\star - \frac{e^2}{2}\right) - p$. Further, $\pi_1 f_1 + f_2 = 0$.

Remark 4. We observe above that the gradients for both e and w are proportional to each other, except for the scaling factors in terms of e and w. This forms the basis for the derivation of the energy function discussed in App. G.

The following lemma characterizes the Hessian.

Lemma 8 (Hessian computation). *Let $\boldsymbol{\gamma} = (b, \boldsymbol{\theta}) \in \mathbb{R}^3$ with $\boldsymbol{\theta} = (e, w) \in \mathbb{R}^2$, and $L(\boldsymbol{\gamma})$ be the next-token prediction loss in Eq. (26) and $L(\boldsymbol{\theta})$ be the one in Eq. (32). Let the Hessian of L at $\boldsymbol{\gamma}$ be*

$$H(\boldsymbol{\gamma}) \triangleq \nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}} L = \begin{bmatrix} H_{bb} & H_{b\boldsymbol{\theta}} \\ H_{b\boldsymbol{\theta}}^\top & H_{\boldsymbol{\theta}\boldsymbol{\theta}} \end{bmatrix} = \begin{bmatrix} \nabla^2_{bb} L & \nabla^2_{b\boldsymbol{\theta}} L \\ (\nabla^2_{b\boldsymbol{\theta}} L)^\top & \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} L \end{bmatrix} \in \mathbb{R}^{3 \times 3}.$$

Then the Hessian of L at $\boldsymbol{\theta} \in \mathbb{R}^2$ is given by

$$H(\boldsymbol{\theta}) \triangleq \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} L = H_{\boldsymbol{\theta}\boldsymbol{\theta}} - H_{b\boldsymbol{\theta}}^\top \cdot H_{bb}^{-1} \cdot H_{b\boldsymbol{\theta}}. \tag{35}$$

Consequently, for any $\boldsymbol{\gamma} = (b, e, w) \in \boldsymbol{\Gamma}_{\min} \cup \boldsymbol{\Gamma}_{\text{sad}}$, the Hessian $H(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = (e, w)$ is given by

$$H(\boldsymbol{\theta}) = \pi_0 \pi_1 \begin{bmatrix} 2(p+q-1)(1+2w|w|) & 0 \\ 0 & 0 \end{bmatrix}, \tag{36}$$

where $\pi_0 = \frac{q}{p+q}$ and $\pi_1 = \frac{p}{p+q}$.

The proofs of the above lemmas are deferred to App. K. We are now ready to present the proof of Thm. 1.

F.1. Proof of Thm. 1

Proof. Let $\boldsymbol{\theta} \in \mathbb{R}^2$ and $\boldsymbol{\gamma}(\boldsymbol{\theta}) = (\boldsymbol{\theta}, b_\star(\boldsymbol{\theta})) \in \mathbb{R}^3$ be its embedding in $\mathbb{R}^3$ with the optimal bias $b_\star(\boldsymbol{\theta}) = \operatorname{argmin}_{b \in \mathbb{R}} L(\boldsymbol{\theta}, b)$ from Lemma 5. Define the following four sets of points:

$$\begin{aligned}
\boldsymbol{\Theta}_\star(p,q) &\triangleq \left\{ (e,w) \in \mathbb{R}^2 : e^2(1+2w|w|) = \log \frac{(1-p)(1-q)}{pq} \right\}, \\
\boldsymbol{\Theta}_{\min}(p,q) &\triangleq \left\{ (e,w) \in \mathbb{R}^2 : e = 0,\ (p+q-1)(1+2w|w|) > 0 \right\}, \\
\boldsymbol{\Theta}_{\max}(p,q) &\triangleq \left\{ (e,w) \in \mathbb{R}^2 : e = 0,\ (p+q-1)(1+2w|w|) < 0 \right\}, \\
\boldsymbol{\Theta}_{\text{sad}}(p,q) &\triangleq \left\{ (e,w) : e = 0,\ w = -1/\sqrt{2} \right\}.
\end{aligned}$$

First we show that any critical point of $L : \mathbb{R}^2 \to \mathbb{R}$ has to lie in one of these sets. Then we characterize that they correspond to the set of all global minima, local minima & maxima, and saddle points respectively.

(i) Set of all critical points: Recall from Thm. 7 that for any critical point $\boldsymbol{\gamma} = (\boldsymbol{\theta}, b) = (e, w, b) \in \mathbb{R}^3$ of L, $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_\star \cup \boldsymbol{\Gamma}_{\min} \cup \boldsymbol{\Gamma}_{\text{sad}}$. Here the main observation is that all these critical points are of the form $(\boldsymbol{\theta}, b_\star(\boldsymbol{\theta}))$ where $\boldsymbol{\theta} \in \boldsymbol{\Theta}_\star \cup \boldsymbol{\Theta}_{\min} \cup \boldsymbol{\Theta}_{\max} \cup \boldsymbol{\Theta}_{\text{sad}}$. To see this, let $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_\star$. Here we have $e^2(1+2w|w|) = \log \frac{(1-p)(1-q)}{pq}$ from Eq. (27) and hence $\boldsymbol{\theta} \in \boldsymbol{\Theta}_\star$. Further, by Lemma 5, we have that the optimal bias for this $\boldsymbol{\theta}$ satisfies $b_\star - \frac{e^2}{2} = \log \frac{p}{1-p}$, which is precisely the characterization of the bias b for $\boldsymbol{\gamma} = (e, w, b)$ in Eq. (27). Likewise, if $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_{\min} \cup \boldsymbol{\Gamma}_{\text{sad}}$, we have $e = 0$ and hence $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\min} \cup \boldsymbol{\Theta}_{\max} \cup \boldsymbol{\Theta}_{\text{sad}}$. Hence by Lemma 5, $b_\star = \log \frac{p}{q}$, matching that of Eq. (28) and Eq. (29). Thus the set of all critical points of L in $\mathbb{R}^3$ are of the form $(\boldsymbol{\theta}, b_\star(\boldsymbol{\theta}))$ with where $\boldsymbol{\theta} \in \boldsymbol{\Theta}_\star \cup \boldsymbol{\Theta}_{\min} \cup \boldsymbol{\Theta}_{\max} \cup \boldsymbol{\Theta}_{\text{sad}}$. Since $\boldsymbol{\Gamma}_\star \cup \boldsymbol{\Gamma}_{\min} \cup \boldsymbol{\Gamma}_{\text{sad}}$ covers the entirety of stationary points of L in $\mathbb{R}^3$, it follows that the set of all stationary points in $\mathbb{R}^2$ is precisely $\boldsymbol{\Theta}_\star \cup \boldsymbol{\Theta}_{\min} \cup \boldsymbol{\Theta}_{\max} \cup \boldsymbol{\Theta}_{\text{sad}}$. Also, the ordering of losses directly follows from the aforementioned observation.

Now we characterize these critical points in terms of the extrema.

(ii) Set of global and local minima: From Eq. (27), for any global minimum $\boldsymbol{\gamma}_\star = (\boldsymbol{\theta}_\star, b_\star(\boldsymbol{\theta}_\star))$ of L in $\mathbb{R}^3$, we have $\boldsymbol{\theta}_\star \in \boldsymbol{\Theta}_\star \subseteq \mathbb{R}^2$. Hence by definition, $\boldsymbol{\Theta}_\star$ is the set of all global minima in $\mathbb{R}^2$. A similar argument holds for $\boldsymbol{\Theta}_{\min}$, which establishes that it is a set of all local minima.

(iii) Set of local maxima and saddle points: From Eq. (29), for any saddle point $\boldsymbol{\gamma} = (e, w, b_\star(e, w))$ of L in $\mathbb{R}^3$, we have that $e = 0$ and $(p+q-1)(1+2w|w|) \le 0$. Hence $\boldsymbol{\theta} = (e, w) \in \boldsymbol{\Theta}_{\max} \cup \boldsymbol{\Theta}_{\text{sad}}$. Suppose $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\max}$ which implies $e = 0, (p+q-1)(1+2w|w|) < 0$. By Lemma 8, the Hessian at $\boldsymbol{\theta}$ (upto a positive scale) is a diagonal matrix with the entries $(p+q-1)(1+2w|w|) < 0$ and 0, corresponding to the directions of e and w respectively. Though one of the eigenvalue here is zero, using a continuity argument as in the proof of Thm. 7 for local minima, we can establish that $\boldsymbol{\theta}$ is indeed a local maximum. Thus $\boldsymbol{\Theta}_{\max}$ is a set of local minima.

Now suppose $(e, w) \in \boldsymbol{\Theta}_{\text{sad}}$. Thus $e = 0$ and $w = -\frac{1}{\sqrt{2}}$. Since it lies at the intersection of $\boldsymbol{\Theta}_{\min}$ and $\boldsymbol{\Theta}_{\max}$, using a neighborhood argument, it's straightforward to see that $\boldsymbol{\Theta}_{\text{sad}}$ is indeed a set of saddle points.

Finally it follows that $\boldsymbol{\Theta}_{\min}, \boldsymbol{\Theta}_{\max}, \boldsymbol{\Theta}_{\text{sad}}$ are the only set of local minima, maxima, and saddle points from the above fact about the characterization of the set of all critical points in terms of these sets and $\boldsymbol{\Theta}_\star$, the ordering of the losses, and using the same argument as in the final steps of the proof of Thm. 7 with the bias. This concludes the proof.

□

G. Proofs of Thm. 2 and Thm. 8

Before we present the proofs of Thm. 2 and Thm. 8, we present few important lemmas that characterize the dynamics of the gradient flow trajectories. Recall from Sec. 4 that the trajectory $(\boldsymbol{\theta}_t)_{t\geq 0}$ is governed by

$$\frac{\mathrm{d}\boldsymbol{\theta}_t}{\mathrm{d}t} = -\nabla L(\boldsymbol{\theta}_t), \quad \boldsymbol{\theta}_t = (e_t, w_t) \in \mathbb{R}^2,\, t \geq 0, \tag{GF}$$

starting with a randomly initalized $\boldsymbol{\theta}_0$. The *energy function* $\mathcal{E}(\cdot,\cdot)$ is defined as

$$\mathcal{E}(e,w) \triangleq e^2 - (w^2 + \mathrm{sign}(w)\cdot \log|w|), \quad \forall (e,w) \in \mathbb{R}^2 \setminus \text{e-axis}, \tag{37}$$

where e-axis $\triangleq \{(e, w=0)\}$ and w-axis $\triangleq \{(e=0, w)\}$. Note that $\mathcal{E}_{\text{sad}} = \mathcal{E}(0, -\frac{1}{\sqrt{2}}) = -\frac{1+\log 2}{2}$. We re-present the Lemma 1 from Sec. 4 below for the sake of completeness.

Lemma 9 (Constant energy along the flow)**.** *For any $(p,q) \in (0,1)^2$ and initialization $\boldsymbol{\theta}_0 = (e_0, w_0) \in \mathbb{R}^2$, let $(\boldsymbol{\theta}_t)_{t\geq 0}$ be the corresponding GF trajectory starting from $\boldsymbol{\theta}_0$. If $w_0 \neq 0$, then the energy stays constant along the trajectory, i.e.*

$$\mathcal{E}(\boldsymbol{\theta}_t) = e_t^2 - (w_t^2 + \mathrm{sign}(w_t)\cdot \log|w_t|) = \mathcal{E}(\boldsymbol{\theta}_0), \quad \forall t \geq 0. \tag{38}$$

On the other hand, if $w_0 = 0$, $w_t = 0$ for all $t \geq 0$. Hence, if we initialize on e-axis *the trajectory always stays on the* e-axis.

Now we establish that the GF trajectories always converge.

Lemma 10 (GF convergence)**.** *Let $(\boldsymbol{\theta}_t)_{t\geq 0}$ be a continuously diferentiable GF trajectory starting from $\boldsymbol{\theta}_0$. Then for all initializations $\boldsymbol{\theta}_0 \in \mathbb{R}^2$,*

(i) *$(\boldsymbol{\theta}_t)_{t\geq 0}$ is bounded,*
(ii) *there exists a $\boldsymbol{\theta}_{\lim} \in \mathbb{R}^2$ such that $\lim_{t\to\infty} \boldsymbol{\theta}_t = \boldsymbol{\theta}_{\lim}$ and*
(iii) *$\lim_{t\to\infty} \|\nabla L(\boldsymbol{\theta}_t)\| = \|\nabla L(\boldsymbol{\theta}_{\lim})\| = 0$.*

Hence $\boldsymbol{\theta}_{\lim}$ is a critical point of L.

The following result characterizes the energy of the limit point.

Lemma 11 (Energy at the limit point)**.** *Consider the same setting as in Lemma 10. If $\boldsymbol{\theta}_0 \in \mathbb{R}^2 \setminus$ e-axis, then $\mathcal{E}(\boldsymbol{\theta}_{\lim}) = \mathcal{E}(\boldsymbol{\theta}_0)$. Hence $\boldsymbol{\theta}_{\lim}$ lies at the intersection of the contour line $\mathcal{E}(e,w) = \mathcal{E}_0$ with the set of critical points of L in $\mathbb{R}^2$.*

On the other hand, if $\boldsymbol{\theta}_0 \in$ e-axis, then $\boldsymbol{\theta}_{\lim} \in$ e-axis.

We now study the energy function on the w-axis which plays a key role in the GF analysis.

Lemma 12 (Analysis of the energy function)**.** *Let $\mathcal{E}(\cdot,\cdot)$ be the energy function defined in Eq. (51) and $f(w) \triangleq \mathcal{E}(e=0,w) = -(w^2 + \mathrm{sign}(w)\cdot\log|w|)$ be the energy evaluated on* w-axis *for $w \in \mathbb{R}\setminus\{0\}$. Then*

(i) *$f : (-\infty, -1/\sqrt{2}] \to (-\infty, \mathcal{E}_{\text{sad}}]$ is monotonically increasing with $\lim_{w\to-\infty} f(w) = -\infty$ and the maximum being $f(-1/\sqrt{2}) = \mathcal{E}_{\text{sad}}$,*
(ii) *$f : [-1/\sqrt{2}, 0) \to [\mathcal{E}_{\text{sad}}, -\infty)$ is monotonically decreasing with $\lim_{w\to 0^-} f(w) = -\infty$,*
(iii) *$f'(-\frac{1}{\sqrt{2}}) = 0$, and*
(iv) *$f : (0,\infty) \to (-\infty, \infty)$ is monotonically decreasing with $\lim_{w\to 0^+} f(w) = \infty$ and $\lim_{w\to\infty} f(w) = -\infty$.*

We are now ready to prove Thm. 2 corresponding to $p + q > 1$.

G.1. Proof of Thm. 2

Proof. Let $\boldsymbol{\theta}_0 = (e_0, w_0) \in \mathbb{R}^2$ be the initialization for the GF trajectory $(\boldsymbol{\theta}_t)_{t\geq 0}$. Recall that

$$\mathcal{I}_{\min} \triangleq \left\{(e,w) : w \in (-1/\sqrt{2}, 0),\, e \in (-g(w), g(w)),\, g(w) = \sqrt{w^2 - \log(-w) + \mathcal{E}_{\text{sad}}}\right\}$$
$$\cup \{(e,w) : w > 0\} \cup \{(e,w) : w = 0\},$$

$$\mathcal{I}_{\text{sad}} \triangleq \left\{(e,w) : w \in [-1/\sqrt{2}, 0),\, e = \pm\sqrt{w^2 - \log(-w) + \mathcal{E}_{\text{sad}}}\right\},$$
$$\mathcal{I}_{\max} \triangleq \left\{(e,w) : e = 0,\, w < -1/\sqrt{2}\right\},$$
$$\mathcal{I}_{\star} \triangleq \mathbb{R}^2 \setminus (\mathcal{I}_{\min} \cup \mathcal{I}_{\text{sad}} \cup \mathcal{I}_{\max}).$$

We consider the cases $\boldsymbol{\theta}_0 \in$ e-axis and $\boldsymbol{\theta}_0 \in \mathbb{R}^2 \setminus$ e-axis separately. First recall from Thm. 1 and Eq. (**??**) that for $p+q > 1$, the loci of the global minima, $e^2(1+2w|w|) = \log\frac{(1-p)(1-q)}{pq} < 0$, lies entirely in the negative half-plane corresponding to $w < -\frac{1}{\sqrt{2}}$. On the other hand, all the local minima, maxima and the saddle points span the w-axis corresponding to $e = 0$.

(i) $\boldsymbol{\theta}_0 \in \mathbb{R}^2 \setminus$ e-axis: Let $\mathcal{E}_0 = \mathcal{E}(\boldsymbol{\theta}_0) \in \mathbb{R}$. By Lemmas. (9), (10), and (11), we have that the trajectory $(\boldsymbol{\theta}_t)_{t\geq 0}$ always stays on the contour line $\mathcal{E}(e,w) = \mathcal{E}_0$ and converges to the limit $\boldsymbol{\theta}_{\lim}$ which is an intersection of this contour line with the set of critical points of L. Hence the crux of the proof is to establish where these intersections occur based on the initialization $\boldsymbol{\theta}_0$ and the initial energy $\mathcal{E}_0$. This gives rise to the set of initializations $\mathcal{I}_{\min}, \mathcal{I}_{\max}, \mathcal{I}_{\text{sad}}$, and $\mathcal{I}_{\star}$ that correspond to the limit being a local minimum/maximum, a saddle point, or a global minimum.

We characterize them individually below starting with $\mathcal{I}_{\min}$.

Initializations for local minima, $\mathcal{I}_{\min}$. For $\boldsymbol{\theta}_0 = (e_0, w_0) \in \mathbb{R}^2 \setminus$ e-axis, assume that $w_0 > 0$. Since $\mathcal{E}_0 \in \mathbb{R}$, there exists an unique $w_\star > 0$ such that $f(w_\star) = \mathcal{E}(0, w_\star) = \mathcal{E}_0$ by Lemma 12, (iv). Further using the fact that the energy contour lines do not cross each other (by definition of a contour line) and the fact they do not intersect the e-axis (it's an energy barrier as discussed in Sec. 4), it follows that the contour line $\mathcal{E}(e,w) = \mathcal{E}_0$ stays entirely in the positive half-plane corresponding to $w > 0$ and $w_\star > 0$ is the unique (and only) intersection of this line with the w-axis, and hence the set of critical points. Since the w-axis corresponding to $w > 0$ is a set of a local minima (Eq. (**??**)), it follows that any initialization (e_0, w_0) with $w_0 > 0$ converges to a local minimum.

Now suppose $-\frac{1}{\sqrt{2}} < w_0 < 0$ and $e_0 \in (-g(w_0), g(w_0))$, where $g(w_0) = \sqrt{w^2 - \log(-w) + \mathcal{E}_{\text{sad}}}$. Thus $|e_0| < g(w_0)$ and hence $e_0^2 - (w_0^2 - \log(-w_0)) = \mathcal{E}(e_0, w_0) = \mathcal{E}_0 < \mathcal{E}_{\text{sad}}$. Hence by Lemma 12, (iii), there is a unique intersection of the contour line $\mathcal{E}(e,w) = \mathcal{E}_0$ with the w-axis, which lies in the region $\left(-\frac{1}{\sqrt{2}}, 0\right)$. Further note that this contour line cannot intersect with the global minima loci as it lies in the half-plane $w < -\frac{1}{\sqrt{2}}$, and hence its only intersection with the set of critical points is this segment of w-axis, which is precisely the set of local minima the GF initialized on this line would converge to.

Thus we have shown that any initialization in $\mathcal{I}_{\min} \setminus \cup\{(e,w) : w = 0\}$ converges to a local minimum, the set of which exhausts all the set of local minima $\boldsymbol{\Theta}_{\min}$ except for the origin. Below we will estbalish that any initialization on e-axis $= \{(e,w) : w = 0\}$ converges to the origin, implying $\mathcal{I}_{\min}$ is the full set of initializations for which the limit is a local minimum.

Initializations for saddle points, $\mathcal{I}_{\text{sad}}$. It's straightforward to see that for any $\boldsymbol{\theta}_0 \in \mathcal{I}_{\text{sad}}$, $e_0^2 - (w_0^2 - \log(-w_0)) = \mathcal{E}(0, -\frac{1}{\sqrt{2}}) = \mathcal{E}_{\text{sad}}$. Since $-\frac{1}{\sqrt{2}} \leq w_0 < 0$, the point $(w,e) = (-\frac{1}{\sqrt{2}}, 0)$ is the only intersection of the contour line with the set of critical points, any initialization in $\mathcal{I}_{\text{sad}}$ converges to the saddle point. On the other hand, there also exists a contour line $e_0^2 - (w_0^2 - \log(-w_0)) = \mathcal{E}_{\text{sad}}$ for $w_0 < -\frac{1}{\sqrt{2}}$ that passes through $(-\frac{1}{\sqrt{2}}, 0) \in \mathbb{R}^2$ and further intersecting with the global minima loci $\boldsymbol{\Theta}_\star$. However, if we initialize on this line the flow escapes away from the saddle point and converges instead to a global minimum. To show this, it suffices to prove that $\frac{\mathrm{d}e_t}{\mathrm{d}t} > 0$ and $\frac{\mathrm{d}w_t}{\mathrm{d}t} < 0$ if $e_0 > 0$ and $w_0 < -\frac{1}{\sqrt{2}}$, such that (w_0, e_0) is close to the saddle point $(-\frac{1}{\sqrt{2}}, 0)$ (the case for $e_0 < 0$ is similar as the flow is symmetric in $e \in \mathbb{R}$). From Lemma 7 and the definition of the GF, we have that

$$\frac{\mathrm{d}e_t}{\mathrm{d}t} = -\frac{\partial L}{\partial e}(e_0, w_0) = 2\mathbb{E}_X\left[(f_1 X + f_2)X\right] \cdot (1 - 2w_0^2))e_0$$
$$\frac{\mathrm{d}w_t}{\mathrm{d}t} = -\frac{\partial L}{\partial w}(e_0, w_0) = 4\mathbb{E}_X\left[(f_1 X + f_2)X\right] \cdot (-e_0^2 w_0).$$

So it suffices to show that $\mathbb{E}_X\left[(f_1 X + f_2)X\right] > 0$. To establish this, we have from Lemma 5 that

$$\mathbb{E}_X\left[(f_1 X + f_2)X\right] = \mathbb{E}[X](f_1 + f_2) = \pi_1\left(-\frac{f_2}{\pi_1} + f_2\right) = -\pi_0 \cdot f_2.$$

From the defintion of f_2 and the optimal bias $b_\star$ in Lemma 7 and Lemma 5 respectively, we obtain

$$f_2 = \sigma\left(b_\star - \frac{e_0^2}{2}\right) - p = \left(1 + \exp\left(-b_\star + \frac{e_0^2}{2}\right)\right)^{-1} - p$$
$$= \left(1 + \frac{2A}{\frac{p}{q} - 1 + \sqrt{\left(\frac{p}{q} - 1\right)^2 + 4 \cdot \frac{p}{q} \cdot A}}\right)^{-1} - p, \quad A \triangleq \exp(e_0^2(1 - 2w_0^2)).$$

When $e_0 = 0$, we have $A = 1$ and hence

$$f_2 = \left(1 + \frac{q}{p}\right)^{-1} - p = \frac{p}{p+q} - p = -\frac{p}{p+q}(p + q - 1) < 0, \tag{39}$$

where we used the fact that $p + q > 1$. Hence by continuity of f_2 in e_0, for e_0 sufficiently close to 0, $f_2 < 0$ which proves our claim about the direction of the flow close to the saddle point. By using the continuity of the flow, it follows that GF cannot converge to saddle point when initialized on this contour line for $w_0 < -\frac{1}{\sqrt{2}}$. Thus $\mathcal{I}_{\text{sad}}$ is the only set of initializations for convergence to $\mathbf{\Theta}_{\text{sad}}$.

Initializations for local maxima, $\mathcal{I}_{\max}$. If $p + q > 1$, we have from Thm. 1 that $\mathbf{\Theta}_{\max} = \left\{(e, w) \in \mathbb{R}^2 : e = 0,\ (1 + 2w|w|) < 0\right\} = \left\{(e, w) \in \mathbb{R}^2 : e = 0,\ w < -\frac{1}{\sqrt{2}}\right\}$. Thus for any $\boldsymbol{\theta}_0 \in \mathbf{\Theta}_{\max}$, $\frac{\mathrm{d}\boldsymbol{\theta}_t}{\mathrm{d}dt} = 0$ for all $t \geq 0$ and hence $\boldsymbol{\theta}_{\lim} = \boldsymbol{\theta}_0$. Further if we slightly perturb away from this set, from Eq. (39) it follows that the flow diverges and hence it's an unstable set of critical points (they are local maxima indeed). Thus the only set of initializations leading to local maxima are $\mathcal{I}_{\max} = \mathbf{\Theta}_{\max}$.

Initializations for the global minima, $\mathcal{I}_\star$. Since the set of all critical points of L is $\mathbf{\Theta}_\star \cup \mathbf{\Theta}_{\min} \cup \mathbf{\Theta}_{\max} \cup \mathbf{\Theta}_{\text{sad}}$, and the initializations in $\mathcal{I}_{\min}$, $\mathcal{I}_{\text{sad}}$, and $\mathcal{I}_{\max}$ converge to $\mathbf{\Theta}_{\min}$, $\mathbf{\Theta}_{\text{sad}}$, and $\mathbf{\Theta}_{\max}$ respectively, it follows that the set of initializations for which the GF converges to global minima is $\mathcal{I}_\star = \mathbb{R}^2 \setminus (\mathcal{I}_{\min} \cup \mathcal{I}_{\text{sad}} \cup \mathcal{I}_{\max})$.

In fact, since the loci of the global minima lies in the half-plane correspondint to $w < -\frac{1}{\sqrt{2}}$ when $p + q > 1$, we can precisely determine the location of the global minimum for which the intersection occurs for any $\boldsymbol{\theta}_0 \in \mathcal{I}_\star$. Specifically, we can solve the pair of equations $\mathcal{E}(e, w) = e^2 - w^2 + \log(-w) = \mathcal{E}_0$ and $e^2(1 - 2w^2) = \log\frac{(1-p)(1-q)}{pq}$ which has a unique solution for $w < 0$ (upto a sign flip in e).

(ii) $\boldsymbol{\theta}_0 \in$ e-axis $\Rightarrow \boldsymbol{\theta}_0 \in \mathcal{I}_{\min}$: If $\boldsymbol{\theta}_0 = (e_0, w_0) \in$ e-axis, we have that $w_0 = 0$ and hence $w_t = 0$ for all $t \geq 0$ (Lemma 9). Lemma 10-(i) also establishes that the iterates $(\boldsymbol{\theta}_t = (e_t, 0))_{t \geq 0}$ stay bounded on the e-axis and monotonically decrease. Since the origin is the only critical point of L on the e-axis, and $\lim_{t\to\infty} \boldsymbol{\theta}_t = \boldsymbol{\theta}_{\lim}$ exists, it follows that $\boldsymbol{\theta}_{\lim} = (0, 0)$, a local minima. Thus $\boldsymbol{\theta}_0 \in \mathcal{I}_{\min}$.

This concludes the proof for all the initializations $\boldsymbol{\theta}_0 \in \mathbb{R}^2$.

Gaussian initialization $\mathcal{N}(0, \sigma^2 \boldsymbol{I}_2)$. When $\boldsymbol{\theta}_0$ is initialized according to the standard Gaussian distribution $\mathcal{N}(0, \sigma^2 \boldsymbol{I}_2)$ with $\sigma^2 \ll \frac{1}{\sqrt{2}}$, we note that $\boldsymbol{\theta}_0$ lands in the set $\mathcal{I}_{\min}$ with high probability. In fact, this probability can be made arbitrarily close to 1 depending on σ^2. Thus this initialization will lead to a local minimum convergence on the w-axis. □

G.2. GF dynamics for $p + q < 1$

Theorem 8 (GF dynamics for $p + q < 1$)**.** *Under the same setting as in Thm. 2 with $p + q < 1$, and any initialization $\boldsymbol{\theta}_0 \in \mathbb{R}^2$, the GF trajectory always converges to a $\boldsymbol{\theta}_{\lim} \in \mathbb{R}^2$ which is a critical point of the loss L. More specifically, $\boldsymbol{\theta}_{\lim}$ is*

(i) *a local minimum if*

$$\boldsymbol{\theta}_0 \in \mathcal{I}_{\min} \triangleq \left\{(e, w) : w < -1/\sqrt{2},\ e \in (-g(w), g(w)),\ g(w) = \sqrt{w^2 - \log(-w) + \mathcal{E}_{\text{sad}}}\right\},$$

(ii) *a saddle point if* $\boldsymbol{\theta}_0 \in \mathcal{I}_{\text{sad}} \triangleq \left\{(e, w) : w \leq -1/\sqrt{2},\ e = \pm\sqrt{w^2 - \log(-w) + \mathcal{E}_{\text{sad}}}\right\}$,

(iii) *a local maximum if* $\boldsymbol{\theta}_0 \in \mathcal{I}_{\max} \triangleq \left\{(e, w) : e = 0,\ w > -1/\sqrt{2}\right\}$,

(iv) *and a global minimum if* $\boldsymbol{\theta}_0 \in \mathbb{R}^2 \setminus (\mathcal{I}_{\min} \cup \mathcal{I}_{\text{sad}} \cup \mathcal{I}_{\max})$.

Consequentely, if we use the standard initialization $\boldsymbol{\theta}_0 \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_2)$ *with* $\sigma^2 \ll 1/\sqrt{2}$, $\boldsymbol{\theta}_{\lim}$ *will be a global minimum.*

Proof. The proof for the case of $p + q < 1$ essentially follows the same steps as that of $p + q > 1$. If the initialization is not on the e-axis we use the energy equation to establish the convergence to the critical point at the intersection of the energy contour line with the critical set and if it starts on the e-axis, the only change is that it now converges to the global minimum instead of the origin as in the earlier case. This is due to the fact that origin turns out to be a local maximum when $p + q < 1$ and hence it's an unstable critical point (which can be established as in the proof of Thm. 2 for $\mathcal{I}_{\max}$). □

H. Gradient flow analysis with attention

In this section, we analyze the learning dynamics of the transformer parameters $\boldsymbol{\theta} \in \mathbb{R}^3$ with the attention scalar $a \in \mathbb{R}$, i.e. $\boldsymbol{\theta} = (e, w, a) \in \mathbb{R}^3$. Similar to the analysis for $\boldsymbol{\theta} = (e, w) \in \mathbb{R}^2$, we first introduce the canonical parameterization including $a \in \mathbb{R}$, then analyze the corresponding loss function $L(\cdot)$ in terms of its gradients and critical points, and capitalize on it to study the gradient flow dynamics using the energy align*. We first start with the parameterization.

H.1. Canonical parameterization with attention

Embedding. Recall from App. D that we let $\boldsymbol{e} = e \cdot \boldsymbol{\alpha}$ and $\boldsymbol{p}_n = -p \cdot \boldsymbol{\alpha}$ for all n where $e > 0, p = \frac{e}{2}$ and $\boldsymbol{\alpha} \in \{\pm 1\}^d/\sqrt{d}$. This results in the embedding

$$\boldsymbol{x}_n = e\left(x_n - \frac{1}{2}\right)\boldsymbol{\alpha}.$$

Attention. Similarly, we recall from Eq. (19) that the attention output $\boldsymbol{y}_n$ is given by

$$\boldsymbol{y}_n = e\left(x_n - \frac{1}{2}\right)\boldsymbol{\alpha} + \langle \boldsymbol{v}, \boldsymbol{\alpha}\rangle \left(\sum_{i\in[n]} \mathrm{att}_{n,i} \cdot e\left(x_i - \frac{1}{2}\right)\right)\boldsymbol{\alpha}, \tag{40}$$

where

$$\mathrm{att}_{n,i} \triangleq \exp\left(\langle \boldsymbol{q}_n, \boldsymbol{k}_i\rangle/\sqrt{d}\right) / \left(\sum_{j\in[n]} \exp\left(\langle \boldsymbol{q}_n, \boldsymbol{k}_j\rangle/\sqrt{d}\right)\right),\ \boldsymbol{q}_n = \boldsymbol{W}_Q\,\boldsymbol{x}_n,\ \boldsymbol{k}_i = \boldsymbol{W}_K\,\boldsymbol{x}_i,$$
$$\boldsymbol{W}_Q^\top \boldsymbol{W}_K = (q^2 d)\,\boldsymbol{\alpha}\cdot\boldsymbol{\alpha}^\top \in \mathbb{R}^{d\times d}, \quad \text{for some } q \in \mathbb{R}.$$

Instead of the softmax, now we assume that the attention weights are linear in the scaled dot product, i.e.

$$\begin{aligned}
\mathrm{att}_{n,i} = \frac{\langle \boldsymbol{q}_n, \boldsymbol{k}_i\rangle}{n\sqrt{d}} &= \frac{1}{\sqrt{d}} \cdot \boldsymbol{x}_n^\top \boldsymbol{W}_Q^\top \boldsymbol{W}_K \boldsymbol{x}_n = \frac{q^2 d}{n\sqrt{d}} \cdot (\boldsymbol{x}_n^\top \boldsymbol{\alpha})(\boldsymbol{x}_i^\top \boldsymbol{\alpha}) \\
&\overset{(\|\boldsymbol{\alpha}\|^2 = d)}{=} \frac{q^2 d^3}{n\sqrt{d}} \cdot (ex_n - p)(ex_i - p) \\
&= \frac{q^2 d^{5/2}}{n} \cdot (ex_n - p)(ex_i - p) \\
&= \frac{q^2 d^{5/2} e^2}{n} \cdot \left(x_n - \frac{1}{2}\right)\left(x_i - \frac{1}{2}\right).
\end{aligned} \tag{41}$$

Note that the $1/n$ factor is to ensure normalization for the attention weights in Eq. (40). Now substituting Eq. (41) in Eq. (40), we obtain

$$\begin{aligned}
\boldsymbol{y}_n &= e\left(x_n - \frac{1}{2}\right)\boldsymbol{\alpha} + \langle \boldsymbol{v}, \boldsymbol{\alpha}\rangle \left(\sum_{i\in[n]} \mathrm{att}_{n,i} \cdot e\left(x_i - \frac{1}{2}\right)\right)\boldsymbol{\alpha} \\
&= \left[e\left(x_n - \frac{1}{2}\right) + \langle \boldsymbol{v}, \boldsymbol{\alpha}\rangle \left(\sum_{i\in[n]} \frac{1}{n} q^2 d^{5/2} e^2 (x_n - \frac{1}{2})(x_i - \frac{1}{2})\right) \cdot e\left(x_i - \frac{1}{2}\right)\right]\boldsymbol{\alpha} \\
&= \left[e\left(x_n - \frac{1}{2}\right)\left(1 + \langle \boldsymbol{v}, \boldsymbol{\alpha}\rangle q^2 d^{5/2} e^2 \left(x_i - \frac{1}{2}\right)^2\right)\right]\boldsymbol{\alpha} \\
&= \left[e\left(x_n - \frac{1}{2}\right)\left(1 + \underbrace{\langle \boldsymbol{v}, \boldsymbol{\alpha}\rangle q^2 d^{5/2} \frac{1}{4}}_{a} \cdot e^2 \cdot\right)\right]\boldsymbol{\alpha} \\
&= e\left(x_n - \frac{1}{2}\right)\left(1 + ae^2\right)\boldsymbol{\alpha},
\end{aligned}$$

where we used the fact that $(x_i - \frac{1}{2})^2 = \frac{1}{4}$ since $x_i \in \{0, 1\}$, and

$$a \triangleq \frac{\langle \boldsymbol{v}, \boldsymbol{\alpha} \rangle q^2 d^{5/2}}{4} \tag{42}$$

is the attention scalar. Note that this includes the scaling $\langle \boldsymbol{v}, \boldsymbol{\alpha} \rangle$ from the value matrix $\boldsymbol{W}_V$ and q^2 from the query-key dot product. Thus we succinctly have

$$\boldsymbol{y}_n = e \left(x_n - \frac{1}{2} \right) \left(1 + a e^2 \right) \boldsymbol{\alpha}. \tag{43}$$

Feed-forward. For the feed-forward layer, we have that $\boldsymbol{W}_1 = \frac{|w|}{\sqrt{d}} \mathbf{1} \cdot \boldsymbol{\alpha}^\top \in \mathbb{R}^{4d \times d}$, $\boldsymbol{W}_2 = \frac{w}{\sqrt{d}} \boldsymbol{\alpha} \cdot \mathbf{1}^\top \in \mathbb{R}^{d \times 4d}$. Hence Eq. (43) implies

$$\boldsymbol{W}_1 \boldsymbol{y}_n = \frac{|w|}{\sqrt{d}} \mathbf{1} \cdot \boldsymbol{\alpha}^\top \left[e \left(x_n - \frac{1}{2} \right) \left(1 + a e^2 \right) \right] \boldsymbol{\alpha} = \frac{|w|}{\sqrt{d}} \left[e \left(x_n - \frac{1}{2} \right) \left(1 + a e^2 \right) \right] \mathbf{1}.$$

Thus,

$$\begin{aligned}
\mathrm{ReLU}(\boldsymbol{W}_1 \boldsymbol{y}_n) &= \frac{|w|}{\sqrt{d}} \mathbf{1} \cdot \mathrm{ReLU} \left(\left[e \left(x_n - \frac{1}{2} \right) \left(1 + a e^2 \right) \right] \right) \\
&= \frac{|w|}{\sqrt{d}} \mathbf{1} \cdot e \, \mathrm{ReLU} \left(\left[\left(x_n - \frac{1}{2} \right) \left(1 + a e^2 \right) \right] \right) \\
&= \frac{|w|}{\sqrt{d}} \mathbf{1} \cdot e \left(\frac{x_n}{2} \mathrm{ReLU}\big(1 + a e^2\big) + \frac{1 - x_n}{2} \mathrm{ReLU}\big(-1 - a e^2\big) \right) \\
&= \frac{|w|}{2\sqrt{d}} \mathbf{1} \cdot e \left(x_n \left[\mathrm{ReLU}\big(1 + a e^2\big) - \mathrm{ReLU}\big(-1 - a e^2\big) \right] + \mathrm{ReLU}\left(-1 - a e^2\right) \right).
\end{aligned}$$

Using $\mathrm{ReLU}(x) - \mathrm{ReLU}(-x) = x$ above,

$$\mathrm{ReLU}(\boldsymbol{W}_1 \boldsymbol{y}_n) = \frac{|w|}{2\sqrt{d}} \mathbf{1} \cdot e \left(x_n \left(1 + a e^2 \right) + \mathrm{ReLU}\left(-1 - a e^2 \right) \right).$$

Hence,

$$\begin{aligned}
\boldsymbol{W}_2 \mathrm{ReLU}(\boldsymbol{W}_1 \boldsymbol{y}_n) &= \frac{w}{\sqrt{d}} \boldsymbol{\alpha} \cdot \mathbf{1}^\top \frac{|w|}{2\sqrt{d}} \mathbf{1} \cdot e \left(x_n \left(1 + a e^2 \right) + \mathrm{ReLU}\left(-1 - a e^2 \right) \right) \\
&= 2w|w| e \left(x_n \left(1 + a e^2 \right) + \mathrm{ReLU}\left(-1 - a e^2 \right) \right) \boldsymbol{\alpha} \\
&= 2w|w| e \left(x_n \left(1 + a e^2 \right) + \frac{\left(-1 - a e^2 \right)}{2} + \frac{|1 + a e^2|}{2} \right) \boldsymbol{\alpha} \\
&= 2w|w| e \left(\left(x_n - \frac{1}{2} \right) \left(1 + a e^2 \right) + \frac{|1 + a e^2|}{2} \right) \boldsymbol{\alpha}.
\end{aligned}$$

Thus the embedding $\boldsymbol{z}_n$ is given by

$$\begin{aligned}
\boldsymbol{z}_n = \boldsymbol{y}_n + \boldsymbol{W}_2 \mathrm{ReLU}(\boldsymbol{W}_1 \boldsymbol{y}_n) &= \left[e \left(x_n - \frac{1}{2} \right) \left(1 + a e^2 \right) \right] \boldsymbol{\alpha} + 2w|w| e \left(\left(x_n - \frac{1}{2} \right) \left(1 + a e^2 \right) + \frac{|1 + a e^2|}{2} \right) \boldsymbol{\alpha} \\
&= e \left[\left(x_n - \frac{1}{2} \right) \left(1 + a e^2 \right) \left(1 + 2w|w| \right) + w|w(1 + a e^2)| \right] \boldsymbol{\alpha}.
\end{aligned}$$

Linear. Since $\boldsymbol{e} = \boldsymbol{a} = e \cdot \boldsymbol{\alpha}$ due to weight-tying, the logits are given by

$$\mathrm{logit}_n(e, w, a, b) = \langle \boldsymbol{a}, \boldsymbol{z}_n \rangle + b = e^2 \left[\left(x_n - \frac{1}{2} \right) \left(1 + a e^2 \right) \left(1 + 2w|w| \right) + w|w(1 + a e^2)| \right] + b.$$

Loss. Denote $\boldsymbol{\theta} \triangleq (e, w, a) \in \mathbb{R}^3$. Similar to the case without a (Eq. (11) and Lemma 2), the cross-entropy loss in our setting can be compactly written as

$$L(\boldsymbol{\theta}, b) = \frac{1}{N} \sum_{n \in [N]} \mathbb{E}[\ell_{\log}\left((2x_{n+1} - 1) \cdot \mathrm{logit}_n(\boldsymbol{\theta}, b)\right)] = \mathbb{E}_{X,Y}\left[\ell_{\log}\left((2Y - 1) \cdot \mathrm{logit}_X(\boldsymbol{\theta}, b)\right)\right], \tag{44}$$

where $\mathrm{logit}_X(\boldsymbol{\theta}, b) \triangleq e^2\left[\left(X - \frac{1}{2}\right)\left(1 + ae^2\right)(1 + 2w|w|) + w|w(1 + ae^2)|\right] + b$, $(X, Y) \in \{0, 1\}^2$ are distributed according to $(X, Y) \sim (\boldsymbol{\pi}, \boldsymbol{P})$, i.e. X is a Bernoulli random variable with $X \sim \boldsymbol{\pi} \equiv \mathrm{Bern}(p/(p + q))$ and $Y|X \sim \boldsymbol{P}(p, q)$, the Markov kernel. Further, using the convexity of b in $L(\cdot, b)$, we can consider the optimal bias $b_\star(\boldsymbol{\theta}) = \mathrm{argmin}_{b \in \mathbb{R}} L(\boldsymbol{\theta}, b)$ in Eq. (44) to obtain the loss $L(\boldsymbol{\theta})$:

$$\begin{aligned} L(\boldsymbol{\theta}) \triangleq L(\boldsymbol{\theta}, b_\star) &= \mathbb{E}_{X,Y}\left[\ell_{\log}\left((2Y - 1) \cdot \mathrm{logit}_X(\boldsymbol{\theta}, b_\star)\right)\right] \\ &= \mathbb{E}_{X,Y}\left[\ell_{\log}\left((2Y - 1) \cdot \left(e^2\left[\left(X - \frac{1}{2}\right)\left(1 + ae^2\right)(1 + 2w|w|) + w|w(1 + ae^2)|\right] + b_\star\right)\right)\right]. \end{aligned} \tag{45}$$

H.2. Analysis of the loss function $L(\boldsymbol{\theta})$ from Eq. (45)

Now we establish the gradients of the loss function.

Lemma 13 (Gradient computation and optimal bias)**.** *For any $\boldsymbol{\theta} = (e, w, a) \in \mathbb{R}^3$ and the next-token prediction loss $L(\boldsymbol{\theta})$ in Eq. (45), the gradients are given by*

$$\begin{aligned} \frac{\partial L}{\partial e} &= -\mathbb{E}\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] \cdot 2e\left(1 + ae^2\right)(1 + 2w|w|) \\ &\quad - \mathbb{E}\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] \cdot 2e^3 a\left(1 + 2w|w|\right), \\ \frac{\partial L}{\partial w} &= -\mathbb{E}\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] \cdot 2e^2\left(1 + ae^2\right)\left(|w| + \mathrm{sign}\left(w\right) w\right), \\ \frac{\partial L}{\partial a} &= -\mathbb{E}\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] \cdot e^4\left(1 + 2w|w|\right), \end{aligned}$$

where $X \in \{0, 1\}$ is a Bernoulli random variable with $X \sim \mathrm{Bern}(p/(p + q))$, and

$$\begin{aligned} &f_1 \triangleq 1 - p - q - \phi_1 + \phi_0, \quad f_2 \triangleq p - \phi_0, \\ &\phi_1 \triangleq \sigma\left(e^2\left(\frac{1}{2}\left(1 + ae^2\right)(1 + 2w|w|) + w|w(1 + ae^2)|\right) + b_\star\right), \\ &\phi_0 \triangleq \sigma\left(e^2\left(\frac{-1}{2}\left(1 + ae^2\right)(1 + 2w|w|) + w|w(1 + ae^2)|\right) + b_\star\right), \end{aligned}$$

where the optimal bias $b_\star$ is obtained by solving $\pi_1 f_1 + f_2 = 0$.

Proof. We defer to App. M. □

Theorem 9 (All critical points for linear attention in $\mathbb{R}^4$)**.** *Let the input sequence be $\{x_n\}_{n=1}^N \sim (\boldsymbol{\pi}, \boldsymbol{P})$, the transformer parameters $\boldsymbol{\theta} = (e, w, b, a) \in \mathbb{R}^4$, and the next-token prediction loss $L(\cdot)$ be as in Eq. (44). Then for any $(p, q) \in (0, 1)^2$ with $p + q \neq 1$ and $N \in \mathbb{N}$,*

(i) *the set of all global minima is given by*

$$\boldsymbol{\Gamma}_\star(p, q) \triangleq \{(e, w, b, a) \in \mathbb{R}^4 : e^2 w|w(1 + ae^2)| + b = \frac{1}{2}\log\frac{p(1 - q)}{q(1 - p)}, \tag{46}$$

$$e^2\left(1 + ae^2\right)(1 + 2w|w|) = \log\frac{(1 - q)(1 - p)}{pq}\} \tag{47}$$

(ii) *a set of local minima is given by*

$$\mathbf{\Gamma}_{\min}(p,q) \triangleq \left\{ \boldsymbol{\gamma}_{\min} = (e,w,b,a) \in \mathbb{R}^4 : e = 0, (p+q-1)(1+2w|w|) > 0, b = \log\frac{p}{q} \right\},$$

(iii) *a set of saddle points is*

$$\mathbf{\Gamma}_{\text{sad}}(p,q) \triangleq \left\{ \boldsymbol{\gamma}_{\text{sad}} = (e,w,b,a) \in \mathbb{R}^4 : e = 0, (p+q-1)(1+2w|w|) \leq 0, b = \log\frac{p}{q} \right\}.$$

(iv) *a set of stationary points is*

$$\mathbf{\Gamma}_{\text{station}}(p,q) \triangleq \left\{ \boldsymbol{\gamma}_{\text{station}} = (e,w,b,a) \in \mathbb{R}^4 : e \neq 0, 1+ae^2 = 0, 1+2w|w| = 0, b = \log\frac{p}{q} \right\},$$

Thus the set of all critical points is

$$\left\{ \boldsymbol{\theta} \in \mathbb{R}^2 : \nabla L(\boldsymbol{\theta}) = 0 \right\} = \mathbf{\Gamma}_\star \cup \mathbf{\Gamma}_{\min} \cup \mathbf{\Gamma}_{\text{sad}} \cup \mathbf{\Gamma}_{\text{station}}. \tag{48}$$

In addition, for any $\boldsymbol{\theta}_\star \in \mathbf{\Gamma}_\star, \boldsymbol{\theta}_{\min} \in \mathbf{\Gamma}_{\min}$ and $\boldsymbol{\theta}_{\text{sad}} \in \mathbf{\Gamma}_{\text{sad}}$, the loss values satisfy

$$H(x_{n+1} \mid x_n) = L(\boldsymbol{\theta}_\star) < L(\boldsymbol{\theta}_{\min}) = L(\boldsymbol{\theta}_{\max}) = L(\boldsymbol{\theta}_{\text{sad}}) = H(x_{n+1}).$$

Theorem 10 (All critical points in $\mathbb{R}^3$)**.** *Let the input sequence be $\{x_n\}_{n=1}^N \sim (\boldsymbol{\pi}, \boldsymbol{P})$, the transformer parameters $\boldsymbol{\theta} = (e,w,a) \in \mathbb{R}^3$, and the next-token prediction loss $L(\cdot)$ be as in Eq. (45). Then for any $(p,q) \in (0,1)^2$ with $p+q \neq 1$ and $N \in \mathbb{N}$,*

(i) *the set of all global minima is given by*

$$\mathbf{\Theta}_\star(p,q) \triangleq \{(e,w,a) \in \mathbb{R}^3 : e^2\left(1+ae^2\right)(1+2w|w|) = \log\frac{(1-q)(1-p)}{pq}\} \tag{49}$$

(ii) *a set of local minima is given by*

$$\mathbf{\Theta}_{\min}(p,q) \triangleq \left\{(e,w,a) \in \mathbb{R}^3 : e = 0, (p+q-1)(1+2w|w|) > 0\right\},$$

(iii) *a set of local maxima is given by*

$$\mathbf{\Theta}_{\min}(p,q) \triangleq \left\{(e,w,a) \in \mathbb{R}^3 : e = 0, (p+q-1)(1+2w|w|) < 0\right\},$$

(iv) *a set of saddle points is*

$$\mathbf{\Theta}_{\text{sad}}(p,q) \triangleq \left\{(e,w,a) \in \mathbb{R}^3 : \left(0, -1/\sqrt{2}, a\right)\right\}.$$

Defining a set of stationary points $\mathbf{\Theta}_{\text{station}}(p,q) \triangleq \left\{(e,w,a) \in \mathbb{R}^3 : e \neq 0, 1+ae^2 = 0, 1+2w|w| = 0\right\}$, the set of all critical points is

$$\left\{ \boldsymbol{\theta} \in \mathbb{R}^2 : \nabla L(\boldsymbol{\theta}) = 0 \right\} = \mathbf{\Theta}_\star \cup \mathbf{\Theta}_{\min} \cup \mathbf{\Theta}_{\max} \cup \mathbf{\Theta}_{\text{sad}} \cup \mathbf{\Theta}_{\text{station}}. \tag{50}$$

In addition, for any $\boldsymbol{\theta}_\star \in \mathbf{\Theta}_\star, \boldsymbol{\theta}_{\min} \in \mathbf{\Theta}_{\min}$, and $\boldsymbol{\theta}_{\text{sad}} \in \mathbf{\Theta}_{\text{sad}}$, the loss values satisfy

$$H(x_{n+1} \mid x_n) = L(\boldsymbol{\theta}_\star) < L(\boldsymbol{\theta}_{\min}) = L(\boldsymbol{\theta}_{\max}) = L(\boldsymbol{\theta}_{\text{sad}}) = H(x_{n+1}).$$

We defer the proofs of the theorems to App. M.2.

H.3. Gradient flow analysis

Analogous to the gradient flow analysis for $\boldsymbol{\theta} = (e, w) \in \mathbb{R}^2$ in G, we now study its countepart toegether with the attention scalar, i.e. $\boldsymbol{\theta} = (e, w, a) \in \mathbb{R}^3$. To this end, let $(\boldsymbol{\theta}_t)_{t\geq 0}$ be a C^1 curve in $\mathbb{R}^3$ governed by

$$\frac{\mathrm{d}\boldsymbol{\theta}_t}{\mathrm{d}t} = -\nabla L(\boldsymbol{\theta}_t), \quad \boldsymbol{\theta}_t = (e_t, w_t, a_t) \in \mathbb{R}^3,\ t \geq 0, \tag{GF-Attention}$$

starting with a randomly initalized $\boldsymbol{\theta}_0$. We define the *energy function* $\mathcal{E}(\cdot,\cdot,\cdot)$ as

$$\mathcal{E}(e, w, a) \triangleq e^2 - (w^2 + \operatorname{sign}(w) \cdot \log|w|) - 2a^2, \quad \forall (e, w, a) \in \mathbb{R}^3 \setminus \text{ea-plane}, \tag{51}$$

where ea-plane $\triangleq \{(e, w = 0, a)\}$. The following lemma presents the crucial result that the energy is constant along the flow in GF-Attention.

Lemma 14 (Constant energy along the flow)**.** *For any $(p, q) \in (0, 1)^2$ and initialization $\boldsymbol{\theta}_0 = (e_0, w_0, a_0) \in \mathbb{R}^3$, let $(\boldsymbol{\theta}_t)_{t\geq 0}$ be the corresponding GF-Attention trajectory starting from $\boldsymbol{\theta}_0$. If $w_0 \neq 0$, then the energy stays constant along the trajectory, i.e.*

$$\mathcal{E}(\boldsymbol{\theta}_t) = e_t^2 - (w_t^2 + \operatorname{sign}(w_t) \cdot \log|w_t|) - 2a_t^2 = \mathcal{E}(\boldsymbol{\theta}_0), \quad \forall t \geq 0. \tag{52}$$

On the other hand, if $w_0 = 0$, $w_t = 0$ for all $t \geq 0$. Hence, if we initialize on ea-plane *the trajectory always stays on the* ea-plane.

Now we characterize the convergence of the gradient flow.

Lemma 15 (GF convergence)**.** *Let $(\boldsymbol{\theta}_t)_{t\geq 0}$ be a continuously diferentiable GF-Attention trajectory starting from $\boldsymbol{\theta}_0$. Then for all initializations $\boldsymbol{\theta}_0 \in \mathbb{R}^3$,*

(i) *$(\boldsymbol{\theta}_t)_{t\geq 0}$ is bounded,*
(ii) *there exists a $\boldsymbol{\theta}_{\lim} \in \mathbb{R}^3$ such that $\lim_{t\to\infty} \boldsymbol{\theta}_t = \boldsymbol{\theta}_{\lim}$ and*
(iii) *$\lim_{t\to\infty} \|\nabla L(\boldsymbol{\theta}_t)\| = \|\nabla L(\boldsymbol{\theta}_{\lim})\| = 0$.*

Hence $\boldsymbol{\theta}_{\lim}$ is a critical point of L.

The following result characterizes the energy of the limit point.

Lemma 16 (Energy at the limit point)**.** *Consider the same setting as in Lemma 15. If $\boldsymbol{\theta}_0 \in \mathbb{R}^3 \setminus$ ea-plane, then $\mathcal{E}(\boldsymbol{\theta}_{\lim}) = \mathcal{E}(\boldsymbol{\theta}_0)$. Hence $\boldsymbol{\theta}_{\lim}$ lies at the intersection of the contour line $\mathcal{E}(e, w) = \mathcal{E}_0$ with the set of critical points of L in $\mathbb{R}^3$.*

On the other hand, if $\boldsymbol{\theta}_0 \in$ ea-plane, then $\boldsymbol{\theta}_{\lim} \in$ ea-plane.

We defer the proofs of the lemmas to App. M.

H.4. Role of Standard Initialization

Theorem 11 ([Informal] Role of standard initialization for $p + q - 1 > 0$)**.** *If we use the standard initialization $\boldsymbol{\theta}_0 \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_2)$ with $\sigma^2 \ll 1/\sqrt{2}$, $\boldsymbol{\theta}_{\lim}$ will be a local minimum with high probability.*

I. Proofs of theorems in App. E

I.1. Proof of Thm. 7

Proof. We characterize the set of global minima, local minima, and that of the saddle points individually.

(i) Set of all global minima. Let $\boldsymbol{\gamma}_\star \in \mathbb{R}^3$ be arbitrary. From (**?**)Lemma 1]makkuva2024attention, we have that $\boldsymbol{\gamma}_\star$ is a global minimum for the loss $L(\cdot)$ in Eq. (26) if and only if its prediction probability satisfies $f_{\boldsymbol{\gamma}_\star}(x_1^n) = \mathbb{P}\left(x_{n+1} = 1 \mid x_n\right)$, the Markov kernel. Since the input $\{x_n\}_{n=1}^N \sim (\boldsymbol{\pi}(p,q), \boldsymbol{P}(p,q))$, we have that

$$\mathbb{P}\left(x_{n+1} = 1 \mid x_n\right) = (1 - x_n)p + x_n(1-q) = (1-p-q)x_n + p. \tag{53}$$

On the other hand, by definition, from Eq. (7), $f_{\boldsymbol{\gamma}_\star}(x_1^n) = \sigma\left(e^2(1+2w|w|)\, x_n + b - \frac{e^2}{2}\right)$, where $\boldsymbol{\gamma}_\star = (e, w, b)$. Since $x_n \in \{0,1\}$, this can be further simplified to

$$\begin{aligned} f_{\boldsymbol{\gamma}_\star}(x_1^n) &= \sigma\left(e^2(1+2w|w|)\, x_n + b - \frac{e^2}{2}\right) \\ &= x_n \cdot \sigma\left(e^2(1+2w|w|) + b - \frac{e^2}{2}\right) + (1-x_n)\cdot\sigma\left(b - \frac{e^2}{2}\right) \\ &= x_n\left(\sigma\left(2e^2 w|w|) + b + \frac{e^2}{2}\right) - \sigma\left(b - \frac{e^2}{2}\right)\right) + \sigma\left(b - \frac{e^2}{2}\right). \end{aligned} \tag{54}$$

Since both $f_{\boldsymbol{\gamma}_\star}(x_1^n)$ and $\mathbb{P}\left(x_{n+1} = 1 \mid x_n\right)$ are linear functions of x_n, equating them for all vallues of $x_n \in \{0,1\}$ implies that the respective coeffecients in these functions in Eq. (53) and Eq. (54) are also equal, i.e.

$$\begin{aligned} \sigma\left(b - \frac{e^2}{2}\right) &= p, \\ \sigma\left(2e^2 w|w| + b + \frac{e^2}{2}\right) - \sigma\left(b - \frac{e^2}{2}\right) &= 1 - p - q, \end{aligned}$$

and hence

$$\sigma\left(b - \frac{e^2}{2}\right) = p, \quad \sigma\left(2e^2 w|w|) + b + \frac{e^2}{2}\right) = 1 - q. \tag{55}$$

Since $\sigma(z) = y$ for $y \in (0,1)$ implies $z = \log\frac{y}{1-y}$, Eq. (55) can be rewritten as

$$b - \frac{e^2}{2} = \log\frac{p}{1-p}, \quad 2e^2 w|w| + b + \frac{e^2}{2} = \log\frac{1-q}{q}.$$

Using $2e^2 w|w| + b + \frac{e^2}{2} = e^2(1+2w|w|) + b - \frac{e^2}{2} = e^2(1+2w|w|) + \log\frac{p}{1-p}$ in the second equality above, we obtain

$$\begin{aligned} b - \frac{e^2}{2} &= \log\frac{p}{1-p}, \\ e^2(1+2w|w|) &= \log\frac{1-q}{q} + \log\frac{1-p}{p} = \log\frac{(1-p)(1-q)}{pq}. \end{aligned} \tag{56}$$

Thus $\boldsymbol{\gamma}_\star \in \mathbb{R}^3$ is a global minimum for $L(\cdot)$ if and only if it satisfies Eq. (56) (note that it's already a critical point, as established in Thm. 6). Thus, the set of all global minimum $\boldsymbol{\Gamma}_\star(p,q)$ is given by

$$\boldsymbol{\Gamma}_\star(p,q) \triangleq \left\{\boldsymbol{\gamma}_\star = (e,w,b) \in \mathbb{R}^3 : e^2(1+2w|w|) = \log\frac{(1-p)(1-q)}{pq},\ b - \frac{e^2}{2} = \log\frac{p}{1-p}\right\}.$$

Since the prediction $f_{\boldsymbol{\gamma}_\star}(\cdot)$ equals the Markov kernel for any $\boldsymbol{\gamma}_\star \in \boldsymbol{\Gamma}_\star$, it follows from Thm. 4 (or (**?**)Lemma 1]makkuva2024attention) that $L(\boldsymbol{\gamma}_\star) = H(x_{n+1} \mid x_n)$, the entropy rate of the Markov chain.

(ii) Set of local minima and saddle points.

Define $\mathbf{\Gamma}_{\min}(p,q) \subseteq \mathbb{R}^3$ and $\mathbf{\Gamma}_{\text{sad}} \subseteq \mathbb{R}^3$ as follows:

$$\mathbf{\Gamma}_{\min}(p,q) \triangleq \left\{ \boldsymbol{\gamma}_{\min} = (e,w,b) \in \mathbb{R}^3 : e = 0, (p+q-1)(1+2w|w|) > 0, b = \log\frac{p}{q} \right\},$$
$$\mathbf{\Gamma}_{\text{sad}}(p,q) \triangleq \left\{ \boldsymbol{\gamma}_{\text{sad}} = (e,w,b) \in \mathbb{R}^3 : e = 0, (p+q-1)(1+2w|w|) \leq 0, b = \log\frac{p}{q} \right\}.$$

To show that $\mathbf{\Gamma}_{\min}$ is the set of all bad local minima for $L(\cdot)$, we first show that any $\boldsymbol{\gamma}_{\min} \in \mathbf{\Gamma}_{\min}$ is a bad local minimum and then show that every bad local minimum should belong to $\mathbf{\Gamma}_{\min}$. Similarly for $\mathbf{\Gamma}_{\text{sad}}$. We start with the local minima.

Let $\boldsymbol{\gamma}_{\min} = (e,w,b) \in \mathbf{\Gamma}_{\min}$. Recall that $\boldsymbol{\gamma}_{\min}$ is a stationary point (Thm. 6), i.e.

$$\nabla L(\boldsymbol{\gamma}_{\min}) = 0.$$

Rearragning the order of scalars and writing $\boldsymbol{\gamma}_{\min} = (b,e,w)$, from Lemma 4, the Hessian of the loss at $\boldsymbol{\gamma}_{\min}$ is

$$\nabla^2 L(\boldsymbol{\gamma}_{\min}) = \pi_0\pi_1 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2(p+q-1)(1+2w|w|) & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{57}$$

By definition, $\boldsymbol{\gamma}_{\min} = (b,e,w)$ satisfies $(p+q-1)(1+2w|w|) > 0$. Thus its Hessian in Eq. (57) has a block diagonal structure of the form $\begin{bmatrix} \boldsymbol{H}_{b,e} & 0 \\ 0 & 0 \end{bmatrix}$ where $\boldsymbol{H}_{b,e}$ has both the eigen values positive, and hence positive-definite. In other words, $\boldsymbol{\gamma}_{\min}$ is a local minimum for $L(\cdot)$ in the $(b,e) \in \mathbb{R}^2$ space for any fixed w in the set. Interestingly, using the continuity argument and the fact that $L(b = \log\frac{p}{q}, e = 0, w)$ is constant in $w \in \mathbb{R}$, we can essentially follow the same steps as in proof of Theorem 2 in (**?**)Appendix B.3]makkuva2024attention (Thm. 5 above) to show that $\boldsymbol{\gamma}_{\min} = (b,e,w)$ is a also local minimum for $L(\cdot)$ in the full parameter space $\mathbb{R}^3$. This establishes that $\boldsymbol{\gamma}_{\min}$ is a local minimum for $L(\cdot)$.

For the saddle points, let $\boldsymbol{\gamma}_{\text{sad}} = (e,w,b) \in \mathbf{\Gamma}_{\text{sad}}$. We have that $\boldsymbol{\gamma}_{\text{sad}}$ is a stationary point (Thm. 6) and Lemma 4 implies its Hessian (after rearraging the order of scalars as above with $\boldsymbol{\gamma}_{\text{sad}} = (b,e,w)$) is:

$$\nabla^2 L(\boldsymbol{\gamma}_{\text{sad}}) = \pi_0\pi_1 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2(p+q-1)(1+2w|w|) & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{58}$$

If $w \neq -\frac{1}{\sqrt{2}}$, $(p+q-1)(1+2w|w|) < 0$ for any $\boldsymbol{\gamma}_{\text{sad}} \in \mathbf{\Gamma}_{\text{sad}}$, and hence the Hessian $\nabla^2 L(\boldsymbol{\gamma}_{\text{sad}})$ in Eq. (58) as both positive, negative, and zero eigen values. Thus $\boldsymbol{\gamma}_{\text{sad}}$ is a saddle point for $L(\cdot)$. Using a neighborhood argument, we can similarly argue for $w = \frac{1}{\sqrt{2}}$ to establish that it's also a saddle point. Now we compute the loss value.

For any $\boldsymbol{\gamma}_{\min} = (e,w,b) \in \mathbf{\Gamma}_{\min}$ or $\boldsymbol{\gamma}_{\text{sad}} = (e,w,b) \in \mathbf{\Gamma}_{\text{sad}}$, we have that $e = 0$ and $b = \log\frac{p}{q}$. Thus for $\boldsymbol{\gamma} = \boldsymbol{\gamma}_{\min}$ or $\mathbf{\Gamma}_{\text{sad}}$, the prediction probability in view of Eq. (7) is

$$\mathbb{P}_{\boldsymbol{\gamma}}(x_{n+1} = 1 \mid x_1^n) = \sigma\left(e^2(1+2w|w|)\, x_n + b - \frac{e^2}{2}\right) = \sigma(b) = \frac{p}{p+q} = \mathbb{P}\left(x_{n+1} = 1\right),$$

the marginal distribution. Substituting this equality in the definition of cross-entropy loss $L(\cdot)$ in Eq. (1) and the fact that $\mathbb{P}\left(x_{n+1} = 1\right) = \frac{p}{p+q} = \pi_1$, following the same steps as in (**?**)Appendix B.3]makkuva2024attention, we obtain

$$\begin{aligned} L(\boldsymbol{\gamma}) &= -\frac{1}{N}\sum_{n\in[N]} \mathbb{E}_{x_1^{n+1}}[x_{n+1}\cdot\log f_{\boldsymbol{\gamma}}(x_1^n) + (1-x_{n+1})\cdot\log(1-f_{\boldsymbol{\gamma}}(x_1^n))] \\ &= -\frac{1}{N}\sum_{n\in[N]} \mathbb{E}_{x_1^{n+1}}[x_{n+1}\cdot\log\pi_1 + (1-x_{n+1})\cdot\log\pi_0] \\ &= \frac{1}{N}\sum_{n\in[N]} [-\pi_1\log\pi_1 - \pi_0\log\pi_0] \end{aligned}$$

$$= H(\boldsymbol{\pi}) = H(x_{n+1}).$$

Thus $L(\boldsymbol{\gamma}_{\min}) = L(\boldsymbol{\gamma}_{\text{sad}}) = H(x_{n+1})$. To see that $H(x_{n+1} \mid x_n) = L(\boldsymbol{\gamma}_\star) < L(\boldsymbol{\gamma}_{\min}) = L(\boldsymbol{\gamma}_{\text{sad}}) = H(x_{n+1})$ for any global minimum $\boldsymbol{\gamma}_\star$, observe that the gap

$$L(\boldsymbol{\gamma}_{\min}) - L_\star = H(x_{n+1}) - H(x_{n+1}|x_n) = I(x_n; x_{n+1}) \geq 0,$$

where $I(x_n; x_{n+1})$ is the mutual information between x_n and x_{n+1} (Cover and Thomas, 2006). Hence the optimality gap equals zero if and only if the mutual information equals zero, which happens when x_n and x_{n+1} are independent, i.e. $\mathbb{P}(x_{n+1} = 1 \mid x_n)$ doesn't depend on x_n. Since $\mathbb{P}(x_{n+1} = 1 \mid x_n) = (1-p-q)x_n + p$ from Eq. (53), this happens only when $p + q = 1$ which contradicts the theorem assumption that $p + q \neq 1$. Hence $H(x_{n+1} \mid x_n) = L(\boldsymbol{\gamma}_\star) < L(\boldsymbol{\gamma}_{\min}) = L(\boldsymbol{\gamma}_{\text{sad}}) = H(x_{n+1})$.

Now we finally show that $\boldsymbol{\Gamma}_{\min}$ and $\boldsymbol{\Gamma}_{\text{sad}}$ are the only set of bad local minima and saddle points respectively. Let $\boldsymbol{\gamma}$ is a bad local minimum for $L(\cdot)$. By definition, it's also a critical point. Recall from Thm. 6 that any stationary point $\boldsymbol{\gamma} = (e, w, b)$ for the loss $L(\cdot)$ satisfies that either $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_\star$, $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_{\min}$, or $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_{\text{sad}}$. Clearly $\boldsymbol{\gamma} \notin \boldsymbol{\Gamma}_\star$, as $\boldsymbol{\Gamma}_\star$ is the set of all global minima. Similarly, $\boldsymbol{\gamma} \notin \boldsymbol{\Gamma}_{\text{sad}}$ as every point in $\boldsymbol{\Gamma}_{\text{sad}}$ is a saddle point for the loss $L(\cdot)$ as established above. Hence $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_{\min}$. Thus every bad local minimum in $\mathbb{R}^3$ belongs to $\boldsymbol{\Gamma}_{\min}$. This coupled with the fact above that $\boldsymbol{\Gamma}_{\min}$ is a set of bad local minima implies $\boldsymbol{\Gamma}_{\min}$ is indeed the set of all bad local minima. The proof for $\boldsymbol{\Gamma}_{\text{sad}}$ is similar.

□

I.2. Proof of Thm. 6

Proof. Let $\boldsymbol{\gamma} = (e, w, b) \in \mathbb{R}^3$ be such that $\nabla L(\boldsymbol{\gamma}) = \left(\frac{\partial L}{\partial e}, \frac{\partial L}{\partial w}, \frac{\partial L}{\partial b}\right)^\top = 0$. By Lemma 3, we have

$$\frac{\partial L}{\partial e} = \mathbb{E}_X\left[(f_1 X + f_2)(2X(1 + 2w|w|) - 1))\right] \cdot e = 0, \tag{59}$$

$$\frac{\partial L}{\partial w} = \mathbb{E}_X\left[(f_1 X + f_2)X\right] \cdot 4e^2|w| = 0, \tag{60}$$

$$\frac{\partial L}{\partial b} = \mathbb{E}_X\left[f_1 X + f_2\right] = 0, \tag{61}$$

where $X \sim \text{Bern}(p/(p+q))$, $f_1 = \sigma\left(2e^2 w|w| + b + \frac{e^2}{2}\right) + q - 1 - \sigma\left(b - \frac{e^2}{2}\right) + p$, and $f_2 = \sigma\left(b - \frac{e^2}{2}\right) - p$. Our goal is to now show that Eqs. (59)-(61) hold only if either $(e = 0, b = \log\frac{p}{q})$ or $(f_1 = 0, f_2 = 0)$. We consider two cases corresponding to $e = 0$ and $e \neq 0$.

(i): $e = 0$. If $e = 0$, we readily see that $\frac{\partial L}{\partial e} = \frac{\partial L}{\partial e} = 0$. Further, $f_1 = p + q - 1$ and $f_2 = \sigma(b) - p$. Hence, Eq. (61) implies that

$$\mathbb{E}_X\left[f_1 X + f_2\right] = (p+q-1)\mathbb{E}[X] + \sigma(b) - p = (p+q-1)\frac{p}{p+q} + \sigma(b) - p = \sigma(b) - \frac{p}{p+q} = 0,$$

which implies that $b = \log\frac{p}{q}$. Since $w \in \mathbb{R}$ is arbitrary, we see in this case that

$$\begin{aligned}\boldsymbol{\gamma} \in &\left\{(e, w, b) \in \mathbb{R}^3 : e = 0, (p+q-1)(1+2w|w|) > 0, b = \log\frac{p}{q}\right\} \bigcup \\ &\left\{(e, w, b) \in \mathbb{R}^3 : e = 0, (p+q-1)(1+2w|w|) < 0, b = \log\frac{p}{q}\right\}, \\ = &\boldsymbol{\Gamma}_{\min} \cup \boldsymbol{\Gamma}_{\text{sad}}.\end{aligned}$$

(ii): $e \neq 0$. Suppose $e \neq 0$. Here we show that $f_1 = f_2 = 0$ and hence $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_\star$. We consider two cases corresponding to $w \neq 0$ and $w = 0$. Let $w \neq 0$. Since both $e \neq 0$ and $w \neq 0$, and $X = X^2$ in Eq. (60) with $\mathbb{E}[X] = \pi_1 = \frac{p}{p+q} > 0$, we obtain that

$$\mathbb{E}_X[f_1 X^2 + f_2 X] = (f_1 + f_2)\mathbb{E}[X] = (f_1 + f_2)\pi_1 = 0,$$

and hence $f_1 + f_2 = 0$. Further Eq. (61) implies that $f_1\pi_1 + f_2 = 0$. Together, this implies $f_1 = f_2 = 0$.

Now suppose $w = 0$. From Eq. (61), we have that $\mathbb{E}[f_1 X + f_2] = f_1\pi_1 + f_2 = 0$. Since $e \neq 0$, Eq. (59) yields

$$\mathbb{E}_X\left[(f_1 X + f_2)(2X-1))\right] = 2\mathbb{E}_X\left[(f_1 X + f_2)X\right] = 2\pi_1(f_1+f_2) = 0.$$

So $f_1 = f_2 = 0$ in this case too. Thus we have showed that whenever $e \neq 0$, we have $f_1 = f_2 = 0$. Recalling the expressions for f_1 and f_2,

$$\begin{aligned} f_2 &= \sigma\left(b - \frac{e^2}{2}\right) - p = 0 \Rightarrow b - \frac{e^2}{2} = \log\frac{p}{1-p}, \\ f_1 &= \sigma\left(2e^2 w|w| + b + \frac{e^2}{2}\right) + q - 1 - \sigma\left(b - \frac{e^2}{2}\right) + p = \sigma\left(2e^2 w|w| + b + \frac{e^2}{2}\right) + q - 1 = 0, \end{aligned} \tag{62}$$

and hence,

$$2e^2 w|w| + b + \frac{e^2}{2} = \log\frac{1-q}{q}.$$

Substituting $b - \frac{e^2}{2} = \log\frac{p}{1-p}$ in the above equation,

$$2e^2 w|w| + e^2 + b - \frac{e^2}{2} = 2e^2 w|w| + e^2 + \log\frac{p}{1-p} = \log\frac{1-q}{q},$$

and thus,

$$e^2(1 + 2w|w|) = \log\frac{(1-p)(1-q)}{pq}. \tag{63}$$

In view of Eq. (62) and Eq. (63), we have that $\boldsymbol{\gamma} = (e, w, b) \in \boldsymbol{\Gamma}_\star$.

Together, we have shown that whenever $\nabla L(\boldsymbol{\gamma}) = 0$, we have $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_\star \cup \boldsymbol{\Gamma}_{\min} \cup \boldsymbol{\Gamma}_{\text{sad}}$. Since $\boldsymbol{\Gamma}_\star \cup \boldsymbol{\Gamma}_{\min} \cup \boldsymbol{\Gamma}_{\text{sad}} \subseteq \{\boldsymbol{\gamma} : \nabla L(\boldsymbol{\gamma}) = 0\}$, we are done. □

J. Proofs of technical lemmas in App. E

J.1. Proof of Lemma 2

Proof. Recall from Eq. (11) that the cross-entropy loss $L(\cdot)$ is defined as

$$L(\boldsymbol{\gamma}) = -\frac{1}{N}\sum_{n\in[N]} \mathbb{E}_{x_1^{n+1}}[x_{n+1}\cdot\log f_{\boldsymbol{\gamma}}(x_1^n) + (1-x_{n+1})\cdot\log(1-f_{\boldsymbol{\gamma}}(x_1^n))], \tag{64}$$

where $f_{\boldsymbol{\gamma}}(x_1^n) = \sigma(\text{logit}_n) = \sigma\left(e^2(1+2w|w|)\,x_n + b - \frac{e^2}{2}\right)$ from Eq. (7). For any $Y \in \{0,1\}, Z \in \mathbb{R}$, using the fact that $1 - \sigma(Z) = \sigma(-Z)$, we have

$$\begin{aligned} -\left[Y\log\sigma(Z) + (1-Y)\log(1-\sigma(Z))\right] &= -\left[Y\log\sigma(Z) + (1-Y)\log\sigma(-Z)\right] \\ &\overset{(a)}{=} Y\log(1+\exp(-Z)) + (1-Y)\log(1+\exp(Z)) \\ &= \log\left(1+\exp(-(2Y-1)Z)\right) \\ &\overset{(b)}{=} \ell_{\log}\left((2Y-1)Z\right), \end{aligned} \tag{65}$$

where (a) and (b) follow from the definitions of sigmoid and the logistic functions: $\sigma(z) = \frac{1}{1+\exp(-z)}$, $\ell_{\log}(z) = \log(1+\exp(-z))$ for $z \in \mathbb{R}$. Substituting $Y = x_{n+1} \in \{0,1\}$ and $Z = \text{logit}_n \in \mathbb{R}$ in Eq. (65), Eq. (64) simplifies to

$$L(\boldsymbol{\gamma}) = \frac{1}{N}\sum_{n\in[N]} \mathbb{E}[\ell_{\log}\left((2x_{n+1}-1)\cdot\text{logit}_n\right)].$$

Since logit_n is only a function of x_n, the above expectation is over the distribution of the pairs (x_n, x_{n+1}), which for all $n \in [N]$ have the same law as a pair of random variables (X, Y) with $X \sim \boldsymbol{\pi} \equiv \text{Bern}(p/(p+q))$ and $Y|X \sim \boldsymbol{P}(p,q)$, the Markov kernel. Hence the above equality can be rewritten using the definition of logit_n as

$$L(\boldsymbol{\gamma}) = \frac{1}{N}\sum_{n\in[N]} \mathbb{E}[\ell_{\log}\left((2x_{n+1}-1)\cdot \text{logit}_n\right)] = \mathbb{E}_{X,Y}\left[\ell_{\log}\left((2Y-1)\left(e^2(1+2w|w|)X + b - \frac{e^2}{2}\right)\right)\right].$$

□

J.2. Proof of Lemma 3

Proof. With $\boldsymbol{\gamma} = (e, w, b)$ and θ denoting either of the scalars e, w, or b, we have from (**?**)Lemma 2]makkuva2024attention that the gradient of the loss $L(\cdot)$ is given by

$$\nabla_\theta L(\boldsymbol{\gamma}) = -\frac{1}{N}\sum_{n\in[N]} \mathbb{E}_{x_1^{n+1}}\left[(x_{n+1} - f_{\boldsymbol{\theta}}(x_1^n))\cdot \nabla_\theta \,\text{logit}_n\right], \tag{66}$$

where $f_{\boldsymbol{\gamma}}(x_1^n) = \sigma(\text{logit}_n) = \sigma\left(e^2(1+2w|w|)\,x_n + b - \frac{e^2}{2}\right)$. Using the same argument as in the proof of Lemma 6, we can replace the expecatations in Eq. (66) with that of a pair of random variables (X, Y) with $X \sim \boldsymbol{\pi} \equiv \text{Bern}(p/(p+q))$ and $Y|X \sim \boldsymbol{P}(p,q)$, the Markov kernel. That is,

$$\nabla_\theta L(\boldsymbol{\gamma}) = -\mathbb{E}_{X,Y}\left[\left(Y - \sigma\left(e^2(1+2w|w|)X + b - \frac{e^2}{2}\right)\right)\cdot \nabla_\theta\left(e^2(1+2w|w|)X + b - \frac{e^2}{2}\right)\right]. \tag{67}$$

Now we define the error term $\mathcal{E}(X,Y) \triangleq -\left(Y - \sigma\left(e^2(1+2w|w|)X + b - \frac{e^2}{2}\right)\right)$. Our goal is to show that $\mathbb{E}[\mathcal{E}(X,Y) \mid X] = f_1 X + f_2$, where $f_1 \triangleq \sigma\left(2e^2w|w| + b + \frac{e^2}{2}\right) + q - 1 - \sigma\left(b - \frac{e^2}{2}\right) + p$, and $f_2 \triangleq \sigma\left(b - \frac{e^2}{2}\right) - p$, which suffices to prove the lemma. To this end, using the fact that $X \in \{0,1\}$, we have

$$\begin{aligned}
\mathcal{E}(X,Y) &= -\left(Y - \sigma\left(e^2(1+2w|w|)X + b - \frac{e^2}{2}\right)\right)\\
&= -\left(Y - X\cdot\sigma\left(2e^2w|w| + b + \frac{e^2}{2}\right) - (1-X)\cdot\sigma\left(b - \frac{e^2}{2}\right)\right)\\
&= -Y + X\left(\sigma\left(2e^2w|w| + b + \frac{e^2}{2}\right) - \sigma\left(b - \frac{e^2}{2}\right)\right) + \sigma\left(b - \frac{e^2}{2}\right).
\end{aligned}$$

Now taking the conditional expectation with respect to X and using the fact that $\mathbb{E}[Y|X] = \mathbb{P}(Y=1 \mid X) = (1-p-q)X + p$ (since $Y|X \sim \boldsymbol{P}(p,q)$), we have

$$\begin{aligned}
\mathbb{E}[\mathcal{E}(X,Y) \mid X] &= -(1-p-q)X - p + X\left(\sigma\left(2e^2w|w| + b + \frac{e^2}{2}\right) - \sigma\left(b - \frac{e^2}{2}\right)\right) + \sigma\left(b - \frac{e^2}{2}\right)\\
&= X\left(\sigma\left(2e^2w|w| + b + \frac{e^2}{2}\right) + q - 1 - \sigma\left(b - \frac{e^2}{2}\right) + p\right) + \sigma\left(b - \frac{e^2}{2}\right) - p\\
&\overset{(a)}{=} f_1 X + f_2,
\end{aligned}$$

where (a) follows from the definition of f_1 and f_2 above. Thus Eq. (67) simplifies to

$$\nabla_\theta L(\boldsymbol{\gamma}) = \mathbb{E}_X\left[(f_1X + f_2)\cdot\nabla_\theta\left(e^2(1+2w|w|)X + b - \frac{e^2}{2}\right)\right].$$

Letting $\theta = e, w$, and b in the above equation, we finally obtain the individual gradients:

$$\frac{\partial L}{\partial e} = \mathbb{E}_X\left[(f_1X + f_2)(2X(1+2w|w| - 1))\right]\cdot e,$$

$$\frac{\partial L}{\partial w} = \mathbb{E}_X\left[(f_1 X + f_2)X\right] \cdot 4e^2|w|,$$
$$\frac{\partial L}{\partial b} = \mathbb{E}_X\left[f_1 X + f_2\right].$$

□

J.3. Proof of Lemma 4

Proof. Slightly changing the variable order, for any $\boldsymbol{\gamma} = (b, e, w) \in \mathbb{R}^3$, we define

$$\boldsymbol{H}(\boldsymbol{\gamma}) \triangleq \nabla^2 L(\boldsymbol{\gamma}) = \begin{bmatrix} \frac{\partial^2 L}{\partial b^2} & \frac{\partial^2 L}{\partial b \partial e} & \frac{\partial^2 L}{\partial b \partial w} \\ \frac{\partial^2 L}{\partial e \partial b} & \frac{\partial^2 L}{\partial e^2} & \frac{\partial^2 L}{\partial e \partial w} \\ \frac{\partial^2 L}{\partial w \partial b} & \frac{\partial^2 L}{\partial w \partial e} & \frac{\partial^2 L}{\partial w^2} \end{bmatrix} \in \mathbb{R}^{3\times 3}. \tag{68}$$

Recall that for any $\boldsymbol{\gamma}_{\min} \in \boldsymbol{\Gamma}_{\min}$ and $\boldsymbol{\gamma}_{\text{sad}} \in \boldsymbol{\Gamma}_{\text{sad}}$, we have $e = 0$ and $b = \log \frac{p}{q}$. Now we compute the second derivatives of L with respect to any $\boldsymbol{\gamma} = (b = \log \frac{p}{q}, e = 0, w)$. We start with the first derivatives. By Lemma 7, the gradients are

$$\begin{aligned} \frac{\partial L}{\partial b} &= \mathbb{E}_X\left[f_1 X + f_2\right], \\ \frac{\partial L}{\partial e} &= \mathbb{E}_X\left[(f_1 X + f_2)(2X(1 + 2w|w|) - 1))\right] \cdot e, \\ \frac{\partial L}{\partial w} &= \mathbb{E}_X\left[(f_1 X + f_2)X\right] \cdot 4e^2|w|, \end{aligned} \tag{69}$$

where

$$f_1 \triangleq \sigma\left(2e^2 w|w| + b + \frac{e^2}{2}\right) + q - 1 - \sigma\left(b - \frac{e^2}{2}\right) + p,$$
$$f_2 \triangleq \sigma\left(b - \frac{e^2}{2}\right) - p.$$

From Eq. (69), we see that the second derivaties of L depend on the first-derivatives of f_1 and f_2, which we now compute. Recall that the derivative of the sigmoid function obeys $\sigma'(z) = \sigma(z)(1 - \sigma(z)) = \sigma(z)\sigma(-z)$ for any $z \in \mathbb{R}$. Now the gradients of f_1 and f_2 with respect to b, e, and w are

$$\begin{aligned} \frac{\partial f_1}{\partial b} &= \sigma\left(2e^2 w|w| + b + \frac{e^2}{2}\right)\sigma\left(-2e^2 w|w| - b - \frac{e^2}{2}\right) - \sigma\left(b - \frac{e^2}{2}\right)\sigma\left(-b + \frac{e^2}{2}\right), \\ \frac{\partial f_2}{\partial b} &= \sigma\left(b - \frac{e^2}{2}\right)\sigma\left(-b + \frac{e^2}{2}\right), \\ \frac{\partial f_1}{\partial e} &= (4ew|w| + e)\,\sigma\left(2e^2 w|w| + b + \frac{e^2}{2}\right)\sigma\left(-2e^2 w|w| - b - \frac{e^2}{2}\right) + e\,\sigma\left(b - \frac{e^2}{2}\right)\sigma\left(-b + \frac{e^2}{2}\right), \\ \frac{\partial f_2}{\partial e} &= (-e)\,\sigma\left(b - \frac{e^2}{2}\right)\sigma\left(-b + \frac{e^2}{2}\right), \\ \frac{\partial f_1}{\partial w} &= (4e^2 \cdot w\text{sign}(w))\,\sigma\left(2e^2 w|w| + b + \frac{e^2}{2}\right)\sigma\left(-2e^2 w|w| - b - \frac{e^2}{2}\right), \\ \frac{\partial f_2}{\partial w} &= 0. \end{aligned}$$

Using the fact that $\sigma\left(\log\frac{p}{q}\right) = \frac{p}{p+q} = \pi_1$ and $\sigma\left(-\log\frac{p}{q}\right) = \frac{q}{p+q} = \pi_0$, the above gradients evaluated for any $\boldsymbol{\gamma} = (b = \log\frac{p}{q}, e = 0, w)$ further reduce to

$$\begin{aligned}
\frac{\partial f_1}{\partial b}\bigg|_{\boldsymbol{\gamma}} = 0, \quad \frac{\partial f_2}{\partial b}\bigg|_{\boldsymbol{\gamma}} = \pi_0\pi_1, \\
\frac{\partial f_1}{\partial e}\bigg|_{\boldsymbol{\gamma}} = 0, \quad \frac{\partial f_2}{\partial e}\bigg|_{\boldsymbol{\gamma}} = 0, \\
\frac{\partial f_1}{\partial w}\bigg|_{\boldsymbol{\gamma}} = 0, \quad \frac{\partial f_2}{\partial w}\bigg|_{\boldsymbol{\gamma}} = 0.
\end{aligned} \tag{70}$$

Now substituting Eq. (70) when computing the second-derivatives of L in Eq. (69), we obtain

$$\begin{aligned}
\frac{\partial^2 L}{\partial b^2}\bigg|_{\boldsymbol{\gamma}} &= \mathbb{E}_X\left[\frac{\partial f_1}{\partial b}\bigg|_{\boldsymbol{\gamma}} X + \frac{\partial f_2}{\partial b}\bigg|_{\boldsymbol{\gamma}}\right] = \pi_0\pi_1, \\
\frac{\partial^2 L}{\partial b \partial e}\bigg|_{\boldsymbol{\gamma}} &= \mathbb{E}_X\left[\frac{\partial f_1}{\partial e}\bigg|_{\boldsymbol{\gamma}} X + \frac{\partial f_2}{\partial e}\bigg|_{\boldsymbol{\gamma}}\right] = 0, \\
\frac{\partial^2 L}{\partial b \partial w}\bigg|_{\boldsymbol{\gamma}} &= \mathbb{E}_X\left[\frac{\partial f_1}{\partial w}\bigg|_{\boldsymbol{\gamma}} X + \frac{\partial f_2}{\partial w}\bigg|_{\boldsymbol{\gamma}}\right] = 0, \\
\frac{\partial^2 L}{\partial e^2}\bigg|_{\boldsymbol{\gamma}} &= \mathbb{E}_X\left[(f_1 X + f_2)(2X(1+2w|w|) - 1))\right]\Big|_{\boldsymbol{\gamma}} \\
&= \mathbb{E}_X\left[(2f_1(1+2w|w|) - f_1 + 2f_2(1+2w|w|)X - f_2\right]\Big|_{\boldsymbol{\gamma}} \\
&= \mathbb{E}_X\left[(f_1(1+4w|w|) + f_2(2+4w|w|)X - f_2\right]\Big|_{\boldsymbol{\gamma}} \\
&= (f_1(1+4w|w|) + f_2(2+4w|w|))\pi_1 - f_2\Big|_{\boldsymbol{\gamma}} \\
&\overset{(a)}{=} ((p+q-1)(1+4w|w|) - \pi_1(p+q-1)(2+4w|w|))\pi_1 + \pi_1(p+q-1) \\
&= \pi_1(p+q-1)\left(1+4w|w| - \pi_1(2+4w|w|) + 1\right) \\
&\overset{(b)}{=} 2\pi_1\pi_0(p+q-1)(1+2w|w|),
\end{aligned} \tag{71}$$

where (a) follows from the fact that $f_1|_{\boldsymbol{\gamma}} = p+q-1$, $f_2|_{\boldsymbol{\gamma}} = \sigma(b) - p = \frac{p}{p+q} - p = \frac{-p}{p+q}(p+q-1) = -\pi_1(p+q-1)$ and (b) from $1-\pi_1 = \pi_0$. Returning to the remaining second derivatives,

$$
\begin{aligned}
\left.\frac{\partial^2 L}{\partial e \partial w}\right|_{\boldsymbol{\gamma}} &= \left.\frac{\partial}{\partial e}\left(\mathbb{E}[(f_1 X + f_2)X]\cdot 4e^2|w|\right)\right|_{\boldsymbol{\gamma}} \\
&= \left.\frac{\partial}{\partial e}\left(\mathbb{E}[(f_1 + f_2)X]\cdot 4e^2|w|\right)\right|_{\boldsymbol{\gamma}} \\
&= \left.\frac{\partial}{\partial e}\left((f_1+f_2)\cdot 4\pi_1 e^2|w|\right)\right|_{\boldsymbol{\gamma}} \\
&= \left.\frac{\partial}{\partial e}\left(\left(\sigma\left(2e^2 w|w| + b + \frac{e^2}{2}\right) + q - 1\right)\cdot 4\pi_1 e^2|w|\right)\right|_{\boldsymbol{\gamma}} \\
&= \left.\left(\frac{\partial}{\partial e}\sigma\left(2e^2 w|w| + b + \frac{e^2}{2}\right)\right) 4\pi_1 e^2|w|\right|_{\boldsymbol{\gamma}} \\
&\qquad + \left.\left(\sigma\left(2e^2 w|w| + b + \frac{e^2}{2}\right) + q - 1\right)\cdot \frac{\partial}{\partial e}(4\pi_1 e^2|w|)\right|_{\boldsymbol{\gamma}} \\
&= 0, \\
\left.\frac{\partial^2 L}{\partial w^2}\right|_{\boldsymbol{\gamma}} &= \left.\frac{\partial}{\partial w}\left(\mathbb{E}[(f_1 X + f_2)X]\cdot 4e^2|w|\right)\right. \\
&= \left.\left(\frac{\partial}{\partial w}\mathbb{E}[(f_1 X + f_2)X]\cdot 4|w|\right) e^2\right|_{\boldsymbol{\gamma}} \\
&= 0.
\end{aligned}
\tag{72}
$$

Congregating all the second derivatives from Eq. (71) and Eq. (72) into the Hessian $\boldsymbol{H}(\boldsymbol{\gamma})$ in Eq. (68), we finally obtain

$$
\boldsymbol{H}(\boldsymbol{\gamma}) = \pi_0\pi_1 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2(p+q-1)(1+2w|w|) & 0 \\ 0 & 0 & 0 \end{bmatrix}.
$$

□

K. Proofs of lemmas in App. F

K.1. Proof of Lemma 5

Proof. Recall from Lemma 2 that for any $\boldsymbol{\theta} = (e, w) \in \mathbb{R}^2$ and $b \in \mathbb{R}$, we have

$$L(\boldsymbol{\theta}, b) = \mathbb{E}_{X,Y}\left[\ell_{\log}\left((2Y-1)\left(e^2(1+2w|w|)X + b - \frac{e^2}{2}\right)\right)\right].$$

Since $\ell_{\log}(\cdot)$ is a convex function, $Y \in \{0,1\}$ and thus $2Y-1 \in \{\pm 1\}$, the convexity of L in b follows from the following fact:

$$\frac{\partial^2 L}{\partial b^2} = \mathbb{E}_{X,Y}\left[\ell''_{\log}\left((2Y-1)\left(e^2(1+2w|w|)X + b - \frac{e^2}{2}\right)\right)\right] \geq 0.$$

To find the optimal $b_\star$, we set the gradient $\frac{\partial L}{\partial b} = 0$. Thus from Lemma 3, we obtain

$$\frac{\partial L}{\partial b} = \mathbb{E}_X\left[f_1 X + f_2\right] = 0,$$

$$f_1 = \sigma\left(2e^2 w|w| + b + \frac{e^2}{2}\right) + q - 1 - \sigma\left(b - \frac{e^2}{2}\right) + p, f_2 = \sigma\left(b - \frac{e^2}{2}\right) - p.$$

Substituting $E \triangleq e^2(1+2w|w|)$, $B \triangleq b - \frac{e^2}{2}$, and $\mathbb{E}[X] = \pi_1 = p/(p+q)$ in the above equations,

$$\pi_1\left(\sigma(E+B) + q - 1 - \sigma(B) + p\right) + \sigma(B) - p = \pi_1 \cdot \sigma(E+B) + \pi_0 \cdot \sigma(B) + \frac{p(p+q-1)}{p+q} - p = 0.$$

Further simplifying,

$$\pi_0 \cdot \sigma(B) = \pi_1 \cdot (1 - \sigma(E+B)) = \pi_1 \cdot \sigma(-E-B).$$

In other words,

$$\frac{(1+\exp(-B))^{-1}}{(1+\exp(E+B))^{-1}} = \frac{\pi_1}{\pi_0} = \frac{p}{q} \Rightarrow \frac{1+\exp(E)\exp(B)}{1+\exp(-B)} = \frac{p}{q}.$$

Defining $x \triangleq \exp(B)$ and $A \triangleq \exp(E)$, we thus obtain the following quadratic equation in x and its corresponding roots:

$$Ax^2 - x\left(\frac{p}{q} - 1\right) - \frac{p}{q} = 0 \Rightarrow x = \frac{1}{2A}\left[\frac{p}{q} - 1 \pm \sqrt{\left(\frac{p}{q} - 1\right)^2 + 4 \cdot \frac{p}{q} \cdot A}\right].$$

Since $x > 0$, we take the root corresponding to the addition choice above and resubstituting $x = \exp(b - \frac{e^2}{2})$ and $A = \exp(e^2(1+2w|w|))$, we obtain the final expression for $b_\star$. In particular, if $e = 0$, it is easy to see that $A = 1$ and hence $x = \exp(b_\star) = \frac{p}{q}$, implying $b_\star = \log \frac{p}{q}$. Similarly, if $A = \frac{(1-p)(1-q)}{pq}$, it's straightforward to see that $\exp(b_\star - e^2/2) = \frac{p}{1-p}$ and hence $b_\star - e^2/2 = \log \frac{p}{1-p}$. □

K.2. Proof of Lemma 6

Proof. The proof directly follows from Lemma 2 by substituting $b = b_\star$. □

K.3. Proof of Lemma 7

Proof. By Danskin's theorem (Danskin, 1966), it follows that for $b_\star = \operatorname{argmin}_{b\in\mathbb{R}} L(\boldsymbol{\theta}, b)$ and $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, b_\star)$, we have $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, b_\star)$. Using the gradient expressions of $L(\boldsymbol{\theta}, b)$ w.r.t $\boldsymbol{\theta}$ from Lemma 3, and using the fact that $\frac{\partial L}{\partial b} = \mathbb{E}[f_1 X + f_2]$ at $b = b_\star$, the claim follows. □

K.4. Proof of Lemma 8

Proof. Since $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, b_\star)$ where $b_\star = \operatorname{argmin}_{b\in\mathbb{R}} L(\boldsymbol{\theta}, b)$, the identity in Eq. (35) about the Hessian of the loss L with respect to $\boldsymbol{\theta}$ follows from the classical result of (**?**)Lemma 2.2]shapiro1985second about second-derivatives of extremal-value functions. Finally Eq. (36) follows from substituting the full Hessian in $\mathbb{R}^{3\times 3}$ from Lemma 4 in this identity. □

L. Proofs of lemmas in App. G

L.1. Proof of Lemma 9

Proof. Denote $(e_t, w_t) = (e, w)$ with the dependence on time implicitly assumed. Then by the definition of GF and the gradient expressions in Lemma 7, we have that

$$\frac{\mathrm{d}e}{\mathrm{d}t} = -\frac{\partial L}{\partial e}(\boldsymbol{\theta}_t) = -2\mathbb{E}[(f_1 X + f_2)X] \cdot (1 + 2w|w|)e, \tag{73}$$

$$\frac{\mathrm{d}w}{\mathrm{d}t} = -\frac{\partial L}{\partial w}(\boldsymbol{\theta}_t) = -\mathbb{E}[(f_1 X + f_2)X] \cdot 4e^2|w|. \tag{74}$$

Dividing Eq. (73) by $(1 + 2w|w|)e$ and Eq. (74) by $4e^2|w|$, we have

$$\frac{1}{1 + 2w|w|}\frac{\mathrm{d}e}{\mathrm{d}t} = \frac{1}{2e|w|}\frac{\mathrm{d}w}{\mathrm{d}t} \Rightarrow e\frac{\mathrm{d}e}{\mathrm{d}t} = \left(w + \frac{1}{2|w|}\right)\frac{\mathrm{d}w}{\mathrm{d}t}.$$

Noting that $\frac{\mathrm{d}}{\mathrm{d}t}(\mathrm{sign}(w) \cdot \log|w|) = \frac{1}{|w|}\frac{\mathrm{d}w}{\mathrm{d}t}$, the above equation can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^2 - w^2 - \mathrm{sign}(w) \cdot \log|w|\right) = 0.$$

Thus defining $\mathcal{E}(e, w) = e^2 - w^2 - \mathrm{sign}(w) \cdot \log|w|$ for $w \neq 0$, the above equation implies $\mathcal{E}(\boldsymbol{\theta}_t) = \mathcal{E}(\boldsymbol{\theta}_0)$ for $\boldsymbol{\theta}_0 = (e_0, w_0)$ with $w_0 \neq 0$. On the other hand, it's easy to see that if $w_0 = 0$, Eq. (74) implies $\frac{\mathrm{d}w}{\mathrm{d}t} = 0$ at $t = 0$ and hence $w_t = 0$ for all $t \geq 0$. □

L.2. Proof of Lemma 10

Proof. To prove the convergence of the trajectory $(\boldsymbol{\theta}_t)_{t\geq 0}$, we use the classical result due to Łojasiewicz (**?**)Theorem 2.2]absil2005convergence which gurantees the convergence of gradient flow for real analytic functions, as long as the trajectory is bounded. Hence we first show the boundedness of the trajectory.

(i) $(\boldsymbol{\theta}_t)_{t\geq 0}$ is bounded.

We consider the cases $\boldsymbol{\theta}_0 \in$ e-axis and $\boldsymbol{\theta}_0 \in \mathbb{R}^2 \setminus$ e-axis separately.

Let's suppose $\boldsymbol{\theta}_0 = (e_0, w_0) \in$ e-axis, i.e. $w_0 = 0$. Thus it follows from Lemma 9 that $w_t = 0$ for all $t \geq 0$. That is, the trajectory always stays on the e-axis and it suffices to track $(e_t)_{t\geq 0}$ and show that they are bounded. To this end, we show that if $e_0 > 0$, we have $\frac{\mathrm{d}e}{\mathrm{d}t} < 0$ and if $e_0 < 0$, we have $\frac{\mathrm{d}e}{\mathrm{d}t} > 0$ for all $t \geq 0$, which establishes our claim. We have from the GF and Lemma 7 that

$$\frac{\mathrm{d}e_t}{\mathrm{d}t} = -\frac{\partial L}{\partial e}(e_t, w_t = 0) = -2\mathbb{E}_X\left[(f_1 X + f_2)X\right]e_t = -2\pi_1(f_1 + f_2)e_t, \tag{75}$$

and

$$\begin{aligned} f_1 + f_2 &= -\frac{f_2}{\pi_1} + f_2 = -\frac{\pi_0}{\pi_1}f_2 = -\frac{\pi_0}{\pi_1}\left[\sigma\left((b_\star)_t - \frac{e_t^2}{2}\right) - p\right] \\ &= -\frac{\pi_0}{\pi_1}\left[\left(1 + \exp\left(-(b_\star)_t + \frac{e_t^2}{2}\right)\right)^{-1} - p\right] \\ &= -\frac{\pi_0}{\pi_1}\left[\left(1 + \frac{2x_t}{\frac{p}{q} - 1 + \sqrt{\left(\frac{p}{q} - 1\right)^2 + 4 \cdot \frac{p}{q} \cdot x_t}}\right)^{-1} - p\right], \quad x_t \triangleq \exp(e_t^2). \end{aligned} \tag{76}$$

Defining

$$g(x) \triangleq \frac{2x}{\frac{p}{q} - 1 + \sqrt{\left(\frac{p}{q} - 1\right)^2 + 4 \cdot \frac{p}{q} \cdot x}}, \tag{77}$$

and substituting Eq. (77) and Eq. (76) in Eq. (75), we obtain

$$\frac{\mathrm{d}e_t}{\mathrm{d}t} = 2\pi_0 \left(\frac{1}{1+g(x_t)} - p \right) \cdot e_t. \tag{78}$$

Since $x_t = \exp(e_t^2) = \exp(-e_t^2)$, in view of Eq. (78), with out loss of generality, we can assume that $e_0 > 0$ and show that $\frac{\mathrm{d}e_t}{\mathrm{d}t} < 0$ for all $t \geq 0$. That is, the RHS Eq. (78) is negative. Note that $x_t \geq 1$ since $x_t = \exp(e_t^2)$ and $g(x_t) > 0$ since the denominator $\frac{p}{q} - 1 + \sqrt{\left(\frac{p}{q}-1\right)^2 + 4 \cdot \frac{p}{q} \cdot x_t} > \frac{p}{q} - 1 + \frac{p}{q} + 1 = 2 \cdot \frac{p}{q} > 0$. Further $g(1) = \frac{q}{p}$ and $\lim_{x\to\infty} g(x) = \infty$. If we show that $g(x)$ is increasing in x for $x \geq 1$, it implies $\frac{1}{1+g(x)} - p < \frac{1}{1+g(1)} - p = \frac{1}{1+\frac{q}{p}} - p = -\frac{p}{p+q}(p+q-1) < 0$. Thus the gradient in Eq. (78) remains negative starting at $t = 0$ and hence the sequence $(e_t)_{t\geq 0}$ will be bounded. Now it remains to show $g(\cdot)$ is increasing, i.e. $g'(\cdot) > 0$. Defining $C = \left(\frac{p}{q} - 1\right) / \left(2\sqrt{\frac{p}{q}}\right)$ and $D = C^2$, we have that $g(x)$ upto a postive scaling is

$$g(x) = \frac{x}{C + \sqrt{x+D}}.$$

Hence

$$g'(x) = \frac{C + \sqrt{x+D} - \frac{x}{2\sqrt{x+D}}}{(C+\sqrt{x+D})^2}.$$

Thus it suffices to show that $h_1(x) \triangleq C + \sqrt{x+D} > h_2(x) \triangleq \frac{x}{2\sqrt{x+D}}$ for $x \geq 1$. Note that $h_1(1) - h_2(1)$ is given by

$$\begin{aligned} h_1(1) - h_2(1) &= \frac{\frac{p}{q}-1}{2\sqrt{\frac{p}{q}}} + \sqrt{1 + \left(\frac{\frac{p}{q}-1}{2\sqrt{\frac{p}{q}}}\right)^2} - \frac{1}{2\sqrt{1 + \left(\frac{\frac{p}{q}-1}{2\sqrt{\frac{p}{q}}}\right)^2}} \\ &= \sqrt{\frac{p}{q}} - \frac{\sqrt{\frac{p}{q}}}{1+\frac{p}{q}} \\ &> 0. \end{aligned}$$

Now we show that $h'(x) > h_2'(x)$ for all $x \geq 1$ which implies that $h_1(x) > h_2(x)$ for all $x \geq 1$, thus establishing our claim. To this end, we have that

$$\begin{aligned} h_1'(x) - h_2'(x) &= \frac{1}{2\sqrt{x+D}} - \frac{\sqrt{x+D} - \frac{x}{2\sqrt{x+D}}}{2(x+D)} \\ &= \frac{x}{2\sqrt{x+D}(x+D)} > 0. \end{aligned}$$

This proves our claim that $g(\cdot)$ is increasing and hence $(e_t)_{t\geq 0}$, and consequently $(\boldsymbol{\theta}_t)_{t\geq 0}$, is bounded when $\boldsymbol{\theta}_0 \in$ e-axis.

Now let's assume that $\boldsymbol{\theta}_0 = (e_0, w_0) \in \mathbb{R}^2 \setminus$ e-axis. Since $(\boldsymbol{\theta}_t)_{t\geq 0} \subseteq \mathbb{R}^2 \setminus$ e-axis, it follows that the loss $L(\cdot)$ is analytic on the trajectory (since the logistic function is analytic), and hence by (**?**)Theorem 2.2]absil2005convergence, it follows that $\lim_{t\to\infty} \|\boldsymbol{\theta}_t\|$ exists. Now we show that $\lim_{t\to\infty} \|\boldsymbol{\theta}_t\| \neq \infty$, which implies the desired result about boundedness. To show $\lim_{t\to\infty} \|\boldsymbol{\theta}_t\| \neq \infty$, we show that there exists a large $B > 0$ such that for any $\boldsymbol{\theta}_t = (e, w) \in \mathbb{R}^2$ with $\|(e,w)\| \geq B$, the velocity vector $\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}$ points inwards into the ball of radius B and thus the trajectory always stays inside this ball, and hence bounded. To establish this, let's denote $(e_t, w_t) = (e, w)$ with the dependence on time implicitly assumed. Then by the definition of GF and the gradient expressions in Lemma 7, we have that

$$\frac{\mathrm{d}e}{\mathrm{d}t} = -\frac{\partial L}{\partial e}(\boldsymbol{\theta}_t) = -2\mathbb{E}[(f_1 X + f_2)X] \cdot (1 + 2w|w|)e \tag{79}$$

$$\frac{\mathrm{d}w}{\mathrm{d}t} = -\frac{\partial L}{\partial w}(\boldsymbol{\theta}_t) = -\mathbb{E}[(f_1 X + f_2)X] \cdot 4e^2|w|, \tag{80}$$

where $f_1 = \sigma\left(2e^2w|w| + b_\star + \frac{e^2}{2}\right) + q - 1 - \sigma\left(b_\star - \frac{e^2}{2}\right) + p$, and $f_2 = \sigma\left(b_\star - \frac{e^2}{2}\right) - p$ with $\pi_1 f_1 + f_2 = 0$. Given that only $\frac{\mathrm{d}e}{\mathrm{d}t}$ flips in sign under the transformation $(e, w) \mapsto (-e, w)$, with out loss of generality we can assume $e > 0$. Now let's also assume $w > 0$. Thus, in view of Eq. (79) and GF, to show that the derivative points inwards, it suffices to show that $\mathbb{E}[(f_1 X + f_2)X] > 0$ for reasonably large B with $\|(e, w)\| = B$. Similar to Eq. (76) and Eq. (77) above, using the relation $\pi_1 f_1 + f_2 = 0$, we obtain

$$\mathbb{E}[(f_1 X + f_2)X] = \pi_1(f_1 + f_2) = -\pi_0\left(\frac{1}{1 + g(x)} - p\right), \quad x \triangleq \exp(e^2(1 + 2w|w|)). \tag{81}$$

Using the fact that $g(x)$ is increasing for $x \geq 1$ with $\lim_{x\to\infty} g(x) = \infty$, and $|w| = w > 0$, we can chose a $B > 0$ such that for any $\|(e, w)\| \geq B$, in Eq. (81) we have $1/(1 + g(x)) < p$ and hence $\mathbb{E}[(f_1 X + f_2)X] > 0$. This finishes the proof of our claim. The proof for $w < 0$ is similar, where we make use of the fact that $\lim_{x\to 0} g(x) = 0$ to show $\mathbb{E}[(f_1 X + f_2)X] < 0$ for e, w reasonably large.

(ii) $\lim_{t\to\infty} \boldsymbol{\theta}_t = \boldsymbol{\theta}_{\lim}$. Since the logistic function $\ell_{\log}(\cdot)$ is analytic, it follows from Lemma 6 that the loss $L(\boldsymbol{\theta})$ is analytic too whenever $w \neq 0$. On the other hand, when $w = 0$, it's easy to see that L is an analytic function of $e \in \mathbb{R}$. By Lemma 9, we know that if $w_0 \neq 0$, $w_t \neq 0$ and if $w_0 = 0$, $w_t = 0$ for all $t \geq 0$. Thus the loss is analytic on the trajectory for all $t \geq 0$. Since the trajectory is bounded, it follows from Łojasiewicz's theorem (**?**)Theorem 2.2]absil2005convergence that there exists a $\boldsymbol{\theta}_{\lim} \in \mathbb{R}^2$ such that $\lim_{t\to\infty} \boldsymbol{\theta}_t = \boldsymbol{\theta}_{\lim}$.

(iii) $\lim_{t\to\infty} \|\nabla L(\boldsymbol{\theta}_t)\| = \|\nabla L(\boldsymbol{\theta}_{\lim})\| = 0$. Since the trajectory is bounded, it follows from (**?**)Theorem 2]ahmadova2023convergence that the gradient converges to zero, i.e. $\lim_{t\to\infty} \|\nabla L(\boldsymbol{\theta}_t)\| = 0$. Since $\nabla L(\cdot)$ is a continuous function and $\lim_{t\to\infty} \boldsymbol{\theta}_t = \boldsymbol{\theta}_{\lim}$, we have $\lim_{t\to\infty} \|\nabla L(\boldsymbol{\theta}_t)\| = \|\nabla L(\boldsymbol{\theta}_{\lim})\| = 0$.

□

L.3. Proof of Lemma 11

Proof. Since the energy function $\mathcal{E}(\cdot, \cdot)$ in Eq. (51) is a continuous function in $\mathbb{R}^2 \setminus$ e-axis, and any trajectory $(\boldsymbol{\theta}_t)_{t\geq 0}$ with intialization $\theta \in \mathbb{R}^2 \setminus$ e-axis stays in $\mathbb{R}^2 \setminus$ e-axis for all $t \geq 0$ (Lemma 9), it follows that $\lim_{t\to\infty} \mathcal{E}(\boldsymbol{\theta}_t) = \mathcal{E}(\boldsymbol{\theta}_{\lim}) = \mathcal{E}(\boldsymbol{\theta}_0)$. As $\nabla L(\boldsymbol{\theta}_{\lim}) = 0$ from Lemma 10, it follows that $\boldsymbol{\theta}_{\lim}$ lies at the intersection of the contour line $\mathcal{E}(e, w) = \mathcal{E}_0$ with the set of critical points of L in $\mathbb{R}^2$.

On the other hand, if $\boldsymbol{\theta}_0 \in$ e-axis, we have $\boldsymbol{\theta}_t \in$ e-axis from Lemma 9 for all $t \geq 0$. Hence $\boldsymbol{\theta}_{\lim} \in$ e-axis. □

L.4. Proof of Lemma 12

Proof. Recall that $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$, defined as $f(w) \triangleq \mathcal{E}(e = 0, w) = -(w^2 + \mathrm{sign}(w) \cdot \log|w|)$. If $w < 0$, we have

$$f(w) = -(w^2 - \log(-w)), \quad f'(w) = -2w + \frac{1}{w}.$$

Hence $f'(w) \geq 0$ for $w \in (-\infty, -1/\sqrt{2}]$ and $f'(w) \leq 0$ for $w \in [-1/\sqrt{2}, 0)$ with $f'(-\frac{1}{\sqrt{2}}) = 0$. It's also straightforward to see that $\lim_{w\to-\infty} f(w) = -\infty$, $\lim_{w\to 0^-} f(w) = -\infty$, and $f(-1/\sqrt{2}) = \mathcal{E}_{\mathrm{sad}}$ (by the definition of f). This establishes (i), (ii), and (iii).

On the other hand, for $w > 0$, we have $f(w) = -(w^2 + \log w)$ and $f'(w) = -(2w + 1/w)$. Hence f is monotonically decreasing for $w > 0$ with $\lim_{w\to 0^+} f(w) = \infty$ and $\lim_{w\to\infty} f(w) = -\infty$. Note that $w = 0$ acts as an energy barrier since $\lim_{w\to 0^-} f(w) = -\infty$ whereas $\lim_{w\to 0^+} f(w) = \infty$. □

M. Proofs of lemmas in App. H

M.1. Proof of Lemma 13

Proof. First we consider the loss with the bias $L(\boldsymbol{\theta}, b)$ from 44:

$$\begin{aligned} L(\boldsymbol{\theta}, b) &= \mathbb{E}_{X,Y}\left[\ell_{\log}\left((2Y-1)\cdot \mathrm{logit}_X(\boldsymbol{\theta}, b)\right)\right] \\ \mathrm{logit}_X(\boldsymbol{\theta}, b) &= e^2\left[\left(X-\frac{1}{2}\right)\left(1+ae^2\right)\left(1+2w|w|\right)+w|w(1+ae^2)|\right]+b, \quad \boldsymbol{\theta}=(e,w,a). \end{aligned}$$

Using the fact that $\ell'_{\log}(z) = \sigma(z) - 1$ and $2Y - 1 \in \{\pm 1\}$, we have for any $\theta \in \{e, w, a, b\}$ that

$$\nabla_\theta L = \mathbb{E}\left[\left(\sigma((2Y-1)\cdot \mathrm{logit}_X)-1\right)(2Y-1)\cdot \nabla_\theta \mathrm{logit}_X\right] = \mathbb{E}\left[\left(\sigma(\mathrm{logit}_X)-Y\right)\cdot \nabla_\theta \mathrm{logit}_X\right]. \tag{82}$$

Now we simplify $\sigma(\mathrm{logit}_X)$ using the fact that $X \in \{0, 1\}$:

$$\begin{aligned} \sigma(\mathrm{logit}_X) &= \sigma\left(e^2\left[\left(X-\frac{1}{2}\right)\left(1+ae^2\right)\left(1+2w|w|\right)+w|w(1+ae^2)|\right]+b\right) \\ &= X\underbrace{\sigma\left(e^2\left[\frac{1}{2}\left(1+ae^2\right)\left(1+2w|w|\right)+w|w(1+ae^2)|\right]+b\right)}_{\triangleq \phi_1} \\ &\quad + (1-X)\underbrace{\sigma\left(e^2\left[\frac{-1}{2}\left(1+ae^2\right)\left(1+2w|w|\right)+w|w(1+ae^2)|\right]+b\right)}_{\triangleq \phi_0} \\ &= X\phi_1 + (1-X)\phi_0 \\ &= X(\phi_1-\phi_0)+\phi_0. \end{aligned}$$

Thus the gradients in 82 are given by

$$\begin{aligned} \nabla_\theta L &= -\mathbb{E}\left[\left(Y - X(\phi_1-\phi_0)-\phi_0\right)\nabla_\theta \mathrm{logit}_X\right] \\ &= -\mathbb{E}_X\left[\mathbb{E}_{x_1^{n+1}}\left[\left(\mathbb{E}[Y\mid X] - X(\phi_1-\phi_0)-\phi_0\right)\nabla_\theta \mathrm{logit}_X\right]\right] \\ &= -\mathbb{E}_X\left[\left((1-p-q)X + p - X(\phi_1-\phi_0)-\phi_0\right)\nabla_\theta \mathrm{logit}_X\right] \\ &= -\mathbb{E}_X\left[\left(\left(\underbrace{1-p-q-\phi_1+\phi_0}_{f_1}\right)X + \underbrace{p-\phi_0}_{f_2}\right)\nabla_\theta \mathrm{logit}_X\right] \\ &= -\mathbb{E}_X\left[\left(f_1X+f_2\right)\nabla_\theta \mathrm{logit}_X\right]. \end{aligned} \tag{83}$$

Now we compute the individual gradients with respect to e, w, a and b. Recall that

$$\mathrm{logit}_X = e^2\left[\left(X-\frac{1}{2}\right)\left(1+ae^2\right)\left(1+2w|w|\right)+w|w(1+ae^2)|\right]+b.$$

Thus,

$$\begin{aligned}\nabla_e \text{logit}_X &= 2e\left[\left(X-\frac{1}{2}\right)\left(1+ae^2\right)\left(1+2w|w|\right)+w|w(1+ae^2)|\right]\\&+e^2\left[2ae\left(X-\frac{1}{2}\right)\left(1+2w|w|\right)+w\text{sign}\left(w\left(1+ae^2\right)\right)\left(2ae\right)\right]\\&=2e\left(X-\frac{1}{2}\right)\left(1+ae^2\right)\left(1+2w|w|\right)+2ew|w(1+ae^2)|\\&+2e^3a\left(X-\frac{1}{2}\right)\left(1+2w|w|\right)+2e^3aw\text{sign}\left(w\left(1+ae^2\right)\right)\\&=2e\left(X-\frac{1}{2}\right)\left(1+ae^2\right)\left(1+2w|w|\right)+2e^3a\left(X-\frac{1}{2}\right)\left(1+2w|w|\right)\\&+2ew|w(1+ae^2)|+2e^3aw\text{sign}\left(w\left(1+ae^2\right)\right).\end{aligned}$$

Substituting the above equation in Eq. (83), we obtain

$$\begin{aligned}\nabla_e L =&-\left(\mathbb{E}\left[\left(f_1X+f_2\right)\left(X-\frac{1}{2}\right)\right]\right)\cdot 2e\left(1+ae^2\right)\left(1+2w|w|\right)\\&-\left(\mathbb{E}\left[\left(f_1X+f_2\right)\left(X-\frac{1}{2}\right)\right]\right)\cdot 2e^3a\left(1+2w|w|\right)\\&-\left(\mathbb{E}\left[\left(f_1X+f_2\right)\right]\right)\cdot 2ew|w(1+ae^2)|\\&-\left(\mathbb{E}\left[\left(f_1X+f_2\right)\right]\right)\cdot 2e^3aw\,\text{sign}\left(w\left(1+ae^2\right)\right).\end{aligned}\tag{84}$$

Now we compute the derivative with respect to w.

$$\begin{aligned}\text{logit}_X &= e^2\left[\left(X-\frac{1}{2}\right)\left(1+ae^2\right)\left(1+2w|w|\right)+w|w(1+ae^2)|\right]+b\\\Rightarrow\nabla_w\text{logit}_X &= 2e^2\left(X-\frac{1}{2}\right)\left(1+ae^2\right)\left(|w|+\text{sign}\left(w\right)w\right)\\&+e^2\left[|w(1+ae^2)|+w\left(1+ae^2\right)\text{sign}\left(w\left(1+ae^2\right)\right)\right]\\\Rightarrow\nabla_w L =&-\left(\mathbb{E}\left[\left(f_1X+f_2\right)\left(X-\frac{1}{2}\right)\right]\right)2e^2\left(1+ae^2\right)\left(|w|+\text{sign}\left(w\right)w\right)\\&-\left(\mathbb{E}\left[\left(f_1X+f_2\right)\right]\right)e^2\left[|w(1+ae^2)|+w\left(1+ae^2\right)\text{sign}\left(w\left(1+ae^2\right)\right)\right].\end{aligned}$$

Similarly, for a:

$$\begin{aligned}\text{logit}_X &= e^2\left[\left(X-\frac{1}{2}\right)\left(1+2w|w|\right)+w|w(1+ae^2)|\right]+b\\\Rightarrow\nabla_a\text{logit}_X &= e^4\left(X-\frac{1}{2}\right)\left(1+ae^2\right)\left(1+2w|w|\right)\\&+e^4w^2\text{sign}\left(w(1+ae^2)\right)\\\Rightarrow\nabla_a L =&-\left(\mathbb{E}\left[\left(f_1X+f_2\right)\left(X-\frac{1}{2}\right)\right]\right)e^4\left(1+2w|w|\right)\\&-\left(\mathbb{E}\left[\left(f_1X+f_2\right)\right]\right)e^4w^2\text{sign}\left(w(1+ae^2)\right).\end{aligned}$$

Finally, since $\nabla_b\text{logit}_X=1$, it follows from Eq. (83) that

$$\nabla_b L=-\mathbb{E}\left[f_1X+f_2\right].\tag{85}$$

For the optimal $b_\star$, we have $\nabla_b L=0$ and hence $\mathbb{E}\left[f_1X+f_2\right]=0$, simplifying the expressions for the gradients of e, w, and a above. This concludes the proof.

□

M.2. Optimality conditions for linear self- attention

We prove Thm. 9 and Thm. 10 below. Note that Thm. 10 directly follows from the former by removing the bias b since for critical points, the bias b is already the optimal one, similar to the proof of Thm. 1.

We characterize the set of global minima first.

Set of all global minima. Let $\boldsymbol{\gamma}_\star \in \mathbb{R}^4$ be arbitrary. From (Makkuva et al., 2024)-Lemma 1, we have that $\boldsymbol{\gamma}_\star$ is a global minimum for the loss $L(\cdot)$ in Eq. (26) if and only if its prediction probability satisfies $f_{\boldsymbol{\gamma}_\star}(x_1^n) = \mathbb{P}\left(x_{n+1} = 1 \mid x_n\right)$, the Markov kernel. Since the input $\{x_n\}_{n=1}^N \sim (\boldsymbol{\pi}(p,q), \boldsymbol{P}(p,q))$, we have that

$$\mathbb{P}\left(x_{n+1} = 1 \mid x_n\right) = (1-x_n)p + x_n(1-q) = (1-p-q)x_n + p. \tag{86}$$

On the other hand, by definition, from Eq. (7), $f_{\boldsymbol{\gamma}_\star}(x_1^n) = \sigma\left(e^2\left[\left(x_n - \frac{1}{2}\right)\left(1+ae^2\right)\left(1+2w|w|\right) + w|w(1+ae^2)|\right] + b\right)$, where $\boldsymbol{\gamma}_\star = (e, w, b)$. Since $x_n \in \{0,1\}$, this can be further simplified to

$$\begin{aligned}
f_{\boldsymbol{\gamma}_\star}(x_1^n) &= \sigma\left(e^2\left[\left(x_n - \frac{1}{2}\right)\left(1+ae^2\right)\left(1+2w|w|\right) + w|w(1+ae^2)|\right] + b\right) && (87)\\
&= x_n\sigma\left(e^2\left[\frac{1}{2}\left(1+ae^2\right)\left(1+2w|w|\right) + w|w(1+ae^2)|\right] + b\right) \\
&\quad + (1-x_n)\sigma\left(e^2\left[\frac{-1}{2}\left(1+ae^2\right)\left(1+2w|w|\right) + w|w(1+ae^2)|\right] + b\right) && (88)\\
&= x_n\sigma\left(e^2\left[\frac{1}{2}\left(1+ae^2\right)\left(1+2w|w|\right) + w|w(1+ae^2)|\right] + b\right) \\
&\quad - x_n\sigma\left(e^2\left[\frac{-1}{2}\left(1+ae^2\right)\left(1+2w|w|\right) + w|w(1+ae^2)|\right] + b\right) \\
&\quad + \sigma\left(e^2\left[\frac{-1}{2}\left(1+ae^2\right)\left(1+2w|w|\right) + w|w(1+ae^2)|\right] + b\right) && (89)
\end{aligned}$$

Since both $f_{\boldsymbol{\gamma}_\star}(x_1^n)$ and $\mathbb{P}\left(x_{n+1} = 1 \mid x_n\right)$ are linear functions of x_n, equating them for all vallues of $x_n \in \{0,1\}$ implies that the respective coeffecients in these functions in Eq. (86) and Eq. (89) are also equal, i.e.

$$\begin{aligned}
1-p-q &= \sigma\left(e^2\left[\frac{1}{2}\left(1+ae^2\right)\left(1+2w|w|\right) + w|w(1+ae^2)|\right] + b\right) \\
&\quad - \sigma\left(e^2\left[\frac{-1}{2}\left(1+ae^2\right)\left(1+2w|w|\right) + w|w(1+ae^2)|\right] + b\right) \\
p &= \sigma\left(e^2\left[\frac{-1}{2}\left(1+ae^2\right)\left(1+2w|w|\right) + w|w(1+ae^2)|\right] + b\right)
\end{aligned}$$

and hence

$$\begin{aligned}
\sigma\left(e^2\left[\frac{1}{2}\left(1+ae^2\right)\left(1+2w|w|\right) + w|w(1+ae^2)|\right] + b\right) &= 1-q \\
\sigma\left(e^2\left[\frac{-1}{2}\left(1+ae^2\right)\left(1+2w|w|\right) + w|w(1+ae^2)|\right] + b\right) &= p
\end{aligned} \tag{90}$$

Since $\sigma(z) = y$ for $y \in (0,1)$ implies $z = \log\frac{y}{1-y}$, Eq. (90) can be rewritten as

$$e^2\left[\frac{1}{2}\left(1+ae^2\right)\left(1+2w|w|\right) + w|w(1+ae^2)|\right] + b = \log\frac{1-q}{q}$$

$$e^2\left[\frac{-1}{2}\left(1+ae^2\right)(1+2w|w|)+w|w(1+ae^2)|\right]+b=\log\frac{p}{1-p}$$

Adding and subtracting the above two equations, we obtain:

$$\begin{aligned} e^2w|w(1+ae^2)|+b&=\frac{1}{2}\log\frac{p(1-q)}{q(1-p)}\\ e^2\left(1+ae^2\right)(1+2w|w|)&=\log\frac{(1-q)(1-p)}{pq}\end{aligned} \tag{91}$$

Thus $\boldsymbol{\gamma}_\star\in\mathbb{R}^3$ is a global minimum for $L(\cdot)$ if and only if it satisfies Eq. (91) (note that it's already a critical point, as established in Thm. 6). Thus, the set of all global minimum $\boldsymbol{\Gamma}_\star(p,q)$ is given by

$$\begin{aligned}\boldsymbol{\Gamma}_\star(p,q)\triangleq\{\boldsymbol{\gamma}_\star=(e,w,b,a)\in\mathbb{R}^4:\ &e^2w|w(1+ae^2)|+b=\frac{1}{2}\log\frac{p(1-q)}{q(1-p)},\\ &e^2\left(1+ae^2\right)(1+2w|w|)=\log\frac{(1-q)(1-p)}{pq}\}\end{aligned}$$

Since the prediction $f_{\boldsymbol{\gamma}_\star}(\cdot)$ equals the Markov kernel for any $\boldsymbol{\gamma}_\star\in\boldsymbol{\Gamma}_\star$, it follows from Thm. 4 (or (**?**)Lemma 1]makkuva2024attention) that $L(\boldsymbol{\gamma}_\star)=H(x_{n+1}\mid x_n)$, the entropy rate of the Markov chain.

M.3. Stationary points

Reproducing the derivates from the previous equations (and setting them to zero)

$$\begin{aligned}\frac{L(\boldsymbol{\theta})}{\partial b}&=\mathbb{E}_X\left[f_1X+f_2\right]=0\\ \frac{L(\boldsymbol{\theta})}{\partial e}&=\mathbb{E}_X\left[(f_1X+f_2)\left(X-\frac{1}{2}\right)\right]2e\left(1+ae^2\right)(1+2w|w|)\\ &\quad+\mathbb{E}_X\left[(f_1X+f_2)\left(X-\frac{1}{2}\right)\right]2e^3a\left(1+2w|w|\right)=0\\ \frac{L(\boldsymbol{\theta})}{\partial w}&=\mathbb{E}_X\left[(f_1X+f_2)\left(X-\frac{1}{2}\right)\right]2e^2\left(1+ae^2\right)(|w|+\operatorname{sign}(w)\,w)=0\\ \frac{L(\boldsymbol{\theta})}{\partial a}&=\mathbb{E}_X\left[(f_1X+f_2)\left(X-\frac{1}{2}\right)\right]e^4\left(1+2w|w|\right)=0\end{aligned} \tag{92}$$

From the above equations, there are multiple regions of stationarity. Here are the following regions of stationarity:

1. $\mathbb{E}_X\left[f_1X+f_2\right]=0, e=0$
2. $\mathbb{E}_X\left[f_1X+f_2\right]=0, e\neq 0, 1+ae^2=0, 1+2w|w|=0$

Slightly changing the variable order, for any $\boldsymbol{\gamma}=(b,e,w,a)\in\mathbb{R}^4$, we define

$$\boldsymbol{H}(\boldsymbol{\gamma})\triangleq\nabla^2L(\boldsymbol{\gamma})=\begin{bmatrix}\frac{\partial^2L}{\partial b^2}&\frac{\partial^2L}{\partial b\partial e}&\frac{\partial^2L}{\partial b\partial w}&\frac{\partial^2L}{\partial b\partial a}\\ \frac{\partial^2L}{\partial e\partial b}&\frac{\partial^2L}{\partial e^2}&\frac{\partial^2L}{\partial e\partial w}&\frac{\partial^2L}{\partial e\partial a}\\ \frac{\partial^2L}{\partial w\partial b}&\frac{\partial^2L}{\partial w\partial e}&\frac{\partial^2L}{\partial w^2}&\frac{\partial^2L}{\partial w\partial a}\\ \frac{\partial^2L}{\partial a\partial b}&\frac{\partial^2L}{\partial a\partial e}&\frac{\partial^2L}{\partial a\partial w}&\frac{\partial^2L}{\partial a^2}\end{bmatrix}\in\mathbb{R}^{4\times4}. \tag{93}$$

Note that:

$$f_1 = \sigma\left(\underbrace{e^2\left[\frac{1}{2}\left(1+ae^2\right)\left(1+2w|w|\right)+w|w(1+ae^2)|\right]+b}_{z_1}\right)$$
$$-\sigma\left(\underbrace{e^2\left[\frac{-1}{2}\left(1+ae^2\right)\left(1+2w|w|\right)+w|w(1+ae^2)|\right]+b}_{z_2}\right)$$
$$+p+q-1$$
$$f_2 = \sigma\left(\underbrace{e^2\left[\frac{-1}{2}\left(1+ae^2\right)\left(1+2w|w|\right)+w|w(1+ae^2)|\right]+b}_{z_2}\right)-p$$

From Eq. (92), we see that the second derivaties of L depend on the first-derivatives of f_1 and f_2, which we now compute. Recall that the derivative of the sigmoid function obeys $\sigma'(z) = \sigma(z)(1-\sigma(z)) = \sigma(z)\sigma(-z)$ for any $z \in \mathbb{R}$. Now the gradients of f_1 and f_2 with respect to b, e, w and a are

$$
\begin{aligned}
\frac{\partial f_1}{\partial b} &= \sigma(z_1)\sigma(-z_1) - \sigma(z_2)\sigma(-z_2) \\
\frac{\partial f_2}{\partial b} &= \sigma(z_2)\sigma(-z_2) \\
\frac{\partial f_1}{\partial e} &= \sigma(z_1)\sigma(-z_1)\left\{2e\left[\frac{1}{2}\left(1+ae^2\right)(1+2w|w|) + w|w(1+ae^2)|\right]\right\} \\
&\quad + \sigma(z_1)\sigma(-z_1)\left\{2ae^3\left[\frac{1}{2}(1+2w|w|) + w|w|\text{sign}(1+2w|w|)\right]\right\} \\
&\quad - \sigma(z_2)\sigma(-z_2)\left\{2e\left[-\frac{1}{2}\left(1+ae^2\right)(1+2w|w|) + w|w(1+ae^2)|\right]\right\} \\
&\quad - \sigma(z_2)\sigma(-z_2)\left\{2ae^3\left[-\frac{1}{2}(1+2w|w|) + w|w|\text{sign}(1+2w|w|)\right]\right\} \\
\frac{\partial f_2}{\partial e} &= \sigma(z_2)\sigma(-z_2)\left\{2e\left[-\frac{1}{2}\left(1+ae^2\right)(1+2w|w|) + w|w(1+ae^2)|\right]\right\} \\
&\quad + \sigma(z_2)\sigma(-z_2)\left\{2ae^3\left[-\frac{1}{2}(1+2w|w|) + w|w|\text{sign}(1+2w|w|)\right]\right\} \\
\frac{\partial f_1}{\partial w} &= \sigma(z_1)\sigma(-z_1)\left\{2e^2\left[\frac{1}{2}\left(1+ae^2\right)|w| + |w||1+ae^2|\right]\right\} \\
&\quad - \sigma(z_2)\sigma(-z_2)\left\{2e^2\left[-\frac{1}{2}\left(1+ae^2\right)|w| + |w||1+ae^2|\right]\right\} \\
\frac{\partial f_2}{\partial w} &= \left\{2e^2\left[-\frac{1}{2}\left(1+ae^2\right)|w| + |w||1+ae^2|\right]\right\} \\
\frac{\partial f_1}{\partial a} &= \sigma(z_1)\sigma(-z_1)\left\{e^4\left[\frac{1}{2}(1+2w|w|) + w|w|\text{sign}\left(1+ae^2\right)\right]\right\} \\
&\quad - \sigma(z_2)\sigma(-z_2)\left\{e^4\left[-\frac{1}{2}(1+2w|w|) + w|w|\text{sign}\left(1+ae^2\right)\right)\right]\right\} \\
\frac{\partial f_2}{\partial a} &= \sigma(z_2)\sigma(-z_2)\left\{e^4\left[-\frac{1}{2}(1+2w|w|) + w|w|\text{sign}\left(1+ae^2\right)\right]\right\}
\end{aligned}
\tag{94}
$$

M.3.1. Stationary Points Where $\mathbb{E}_X[f_1X + f_2] = 0, e = 0$

When $e = 0 \implies z_1 = b, z_2 = b$. Hence:

$$
\begin{aligned}
f_1 &= \sigma(b) + p + q - 1 - \sigma(b) = p + q - 1 \\
f_2 &= \sigma(b) - p
\end{aligned}
$$

On computing the expectation:

$$
\begin{aligned}
\mathbb{E}_X[f_1X + f_2] &= (p+q-1)\,\mathbb{E}_X[X] + \sigma(b) - p = 0 \\
&= (p+q-1)\,\pi_1 + \sigma(b) - p = 0
\end{aligned}
$$

Rearranging and simplifying:

$$
\sigma(b) = \frac{p}{p+q}
$$

$$\implies b = \log \frac{p}{q}$$

Using the fact that $\sigma\left(\log \frac{p}{q}\right) = \frac{p}{p+q} = \pi_1$ and $\sigma\left(-\log \frac{p}{q}\right) = \frac{q}{p+q} = \pi_0$, the above gradients evaluated for any $\boldsymbol{\gamma} = (b = \log \frac{p}{q}, e = 0, w, a)$ further reduce to

$$\begin{aligned}
\left.\frac{\partial f_1}{\partial b}\right|_{\boldsymbol{\gamma}} = 0, \quad \left.\frac{\partial f_2}{\partial b}\right|_{\boldsymbol{\gamma}} = \pi_0\pi_1, \\
\left.\frac{\partial f_1}{\partial e}\right|_{\boldsymbol{\gamma}} = 0, \quad \left.\frac{\partial f_2}{\partial e}\right|_{\boldsymbol{\gamma}} = 0, \\
\left.\frac{\partial f_1}{\partial w}\right|_{\boldsymbol{\gamma}} = 0, \quad \left.\frac{\partial f_2}{\partial w}\right|_{\boldsymbol{\gamma}} = 0. \\
\left.\frac{\partial f_1}{\partial a}\right|_{\boldsymbol{\gamma}} = 0, \quad \left.\frac{\partial f_2}{\partial a}\right|_{\boldsymbol{\gamma}} = 0.
\end{aligned} \tag{95}$$

Now substituting Eq. (95) when computing the second-derivatives of L in Eq. (92), we obtain

$$\begin{aligned}
\left.\frac{\partial^2 L}{\partial b^2}\right|_{\boldsymbol{\gamma}} &= \mathbb{E}_X\left[\left.\frac{\partial f_1}{\partial b}\right|_{\boldsymbol{\gamma}} X + \left.\frac{\partial f_2}{\partial b}\right|_{\boldsymbol{\gamma}}\right] = \pi_0\pi_1, \\
\left.\frac{\partial^2 L}{\partial b \partial e}\right|_{\boldsymbol{\gamma}} &= \mathbb{E}_X\left[\left.\frac{\partial f_1}{\partial e}\right|_{\boldsymbol{\gamma}} X + \left.\frac{\partial f_2}{\partial e}\right|_{\boldsymbol{\gamma}}\right] = 0, \\
\left.\frac{\partial^2 L}{\partial b \partial w}\right|_{\boldsymbol{\gamma}} &= \mathbb{E}_X\left[\left.\frac{\partial f_1}{\partial w}\right|_{\boldsymbol{\gamma}} X + \left.\frac{\partial f_2}{\partial w}\right|_{\boldsymbol{\gamma}}\right] = 0, \\
\left.\frac{\partial^2 L}{\partial b \partial a}\right|_{\boldsymbol{\gamma}} &= \mathbb{E}_X\left[\left.\frac{\partial f_1}{\partial a}\right|_{\boldsymbol{\gamma}} X + \left.\frac{\partial f_2}{\partial a}\right|_{\boldsymbol{\gamma}}\right] \\
\left.\frac{\partial^2 L}{\partial e^2}\right|_{\boldsymbol{\gamma}} &= \mathbb{E}_X\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] 2\left(1 + ae^2\right)(1 + 2w|w|)\Big|_{\boldsymbol{\gamma}} \\
&\quad + \underbrace{\mathbb{E}_X\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] 4e^2 a\,(1 + 2w|w|)\Big|_{\boldsymbol{\gamma}}}_{0} \\
&\quad + \underbrace{\mathbb{E}_X\left[\left(\left.\frac{\partial f_1}{\partial e}\right|_{\boldsymbol{\gamma}} X + \left.\frac{\partial f_2}{\partial e}\right|_{\boldsymbol{\gamma}}\right)\left(X - \frac{1}{2}\right)\right] 2\left(1 + ae^2\right)(1 + 2w|w|)}_{0} \\
&= \mathbb{E}_X\left[(2f_1(1 + 2w|w|) - f_1 + 2f_2(1 + 2w|w|)X - f_2\right]\Big|_{\boldsymbol{\gamma}} \\
&= \mathbb{E}_X\left[(f_1(1 + 4w|w|) + f_2(2 + 4w|w|)X - f_2\right]\Big|_{\boldsymbol{\gamma}} \\
&= (f_1(1 + 4w|w|) + f_2(2 + 4w|w|))\pi_1 - f_2\Big|_{\boldsymbol{\gamma}} \\
&\overset{(a)}{=} ((p+q-1)(1 + 4w|w|) - \pi_1(p+q-1)(2 + 4w|w|))\pi_1 + \pi_1(p+q-1) \\
&= \pi_1(p+q-1)\left(1 + 4w|w| - \pi_1(2 + 4w|w|) + 1\right) \\
&\overset{(b)}{=} 2\pi_1\pi_0(p+q-1)(1 + 2w|w|),
\end{aligned} \tag{96}$$

where (a) follows from the fact that $f_1|_{\boldsymbol{\gamma}} = p+q-1$, $f_2|_{\boldsymbol{\gamma}} = \sigma(b) - p = \frac{p}{p+q} - p = \frac{-p}{p+q}(p+q-1) = -\pi_1(p+q-1)$ and (b) from $1-\pi_1 = \pi_0$. Returning to the remaining second derivatives,

$$
\begin{aligned}
\left.\frac{\partial^2 L}{\partial e \partial w}\right|_{\boldsymbol{\gamma}} &= \frac{\partial}{\partial e}\mathbb{E}_X\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] 4e^2\left(1+ae^2\right)|w|\Big|_{\boldsymbol{\gamma}} \\
&= \mathbb{E}_X\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right]\left(8e\left(1+ae^2\right)|w| + 8e^3 a|w|\right)\Big|_{\boldsymbol{\gamma}} \\
&\quad + \mathbb{E}_X\left[\left(\left.\frac{\partial f_1}{\partial e}\right|_{\boldsymbol{\gamma}} X + \left.\frac{\partial f_2}{\partial e}\right|_{\boldsymbol{\gamma}}\right)\left(X - \frac{1}{2}\right)\right] 4e^2\left(1+ae^2\right)(|w|) \\
&= 0 \\
\left.\frac{\partial^2 L}{\partial e \partial a}\right|_{\boldsymbol{\gamma}} &= \frac{\partial}{\partial e}\mathbb{E}_X\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] e^4\left(1+2w|w|\right)\Big|_{\boldsymbol{\gamma}} \\
&= \mathbb{E}_X\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] 4e^3\left(1+2w|w|\right)\Big|_{\boldsymbol{\gamma}} \\
&\quad + \mathbb{E}_X\left[\left(\left.\frac{\partial f_1}{\partial e}\right|_{\boldsymbol{\gamma}} X + \left.\frac{\partial f_2}{\partial e}\right|_{\boldsymbol{\gamma}}\right)\left(X - \frac{1}{2}\right)\right] e^4\left(1+2w|w|\right) \\
&= 0 \\
\left.\frac{\partial^2 L}{\partial w^2}\right|_{\boldsymbol{\gamma}} &= \frac{\partial}{\partial w}\mathbb{E}_X\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] 4e^2\left(1+ae^2\right)|w|\Big|_{\boldsymbol{\gamma}} \\
&= \mathbb{E}_X\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] 4e^2\left(1+ae^2\right)\operatorname{sign}(w) \\
&\quad + \mathbb{E}_X\left[\left(\left.\frac{\partial f_1}{\partial w}\right|_{\boldsymbol{\gamma}} X + \left.\frac{\partial f_2}{\partial w}\right|_{\boldsymbol{\gamma}}\right)\left(X - \frac{1}{2}\right)\right] 4e^2\left(1+ae^2\right)|w| \\
&= 0 \\
\left.\frac{\partial^2 L}{\partial w \partial a}\right|_{\boldsymbol{\gamma}} &= \frac{\partial}{\partial w}\mathbb{E}_X\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] e^4\left(1+2w|w|\right)\Big|_{\boldsymbol{\gamma}} \\
&= \mathbb{E}_X\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] 2e^4|w|\Big|_{\boldsymbol{\gamma}} \\
&\quad + \mathbb{E}_X\left[\left(\left.\frac{\partial f_1}{\partial w}\right|_{\boldsymbol{\gamma}} X + \left.\frac{\partial f_2}{\partial w}\right|_{\boldsymbol{\gamma}}\right)\right]_{\boldsymbol{\gamma}} e^4\left(1+2w|w|\right) \\
&= 0 \\
\left.\frac{\partial^2 L}{\partial a^2}\right|_{\boldsymbol{\gamma}} &= \frac{\partial}{\partial a}\mathbb{E}_X\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] e^4\left(1+2w|w|\right)\Big|_{\boldsymbol{\gamma}} \\
&= \mathbb{E}_X\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] 2e^4|w|\Big|_{\boldsymbol{\gamma}} \\
&\quad + \mathbb{E}_X\left[\left(\left.\frac{\partial f_1}{\partial w}\right|_{\boldsymbol{\gamma}} X + \left.\frac{\partial f_2}{\partial w}\right|_{\boldsymbol{\gamma}}\right)\right]_{\boldsymbol{\gamma}} e^4\left(1+2w|w|\right) \\
&= 0
\end{aligned}
\tag{97}
$$

Congregating all the second derivatives from Eq. (96) and Eq. (97) into the Hessian $\boldsymbol{H}(\boldsymbol{\gamma})$ in Eq. (93), we finally obtain

$$\boldsymbol{H}(\boldsymbol{\gamma}) = \pi_0\pi_1 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2(p+q-1)(1+2w|w|) & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

It is trivial to see that (borrowing the result proved in E)

Similar to the notation defined before $\boldsymbol{\Gamma}_{\min}(p,q) \subseteq \mathbb{R}^4$ and $\boldsymbol{\Gamma}_{\text{sad}} \subseteq \mathbb{R}^4$ as follows:

$$\boldsymbol{\Gamma}_{\min}(p,q) \triangleq \left\{ \boldsymbol{\gamma}_{\min} = (e,w,b,a) \in \mathbb{R}^4 : e = 0, (p+q-1)(1+2w|w|) > 0, b = \log\frac{p}{q} \right\},$$
$$\boldsymbol{\Gamma}_{\text{sad}}(p,q) \triangleq \left\{ \boldsymbol{\gamma}_{\text{sad}} = (e,w,b,a) \in \mathbb{R}^4 : e = 0, (p+q-1)(1+2w|w|) \leq 0, b = \log\frac{p}{q} \right\}.$$

M.3.2. Stationary Points Where $\mathbb{E}_X[f_1X + f_2] = 0, e \neq 0, 1+ae^2 = 0, 1+2w|w| = 0$

For this set of points, the Hessian remains undefined because $\frac{\partial f_1}{\partial e}, \frac{\partial f_2}{\partial e}, \frac{\partial f_1}{\partial a}, \frac{\partial f_2}{\partial a}$ do not exist. This non-existence arises since sign $\left(1+ae^2\right)$ lacks definition when $1+ae^2 = 0$. However, even in this scenario,when $e \neq 0, 1+ae^2 = 0, 1+2w|w| = 0 \implies z_1 = b, z_2 = b$. Hence:

$$f_1 = \sigma(b) + p + q - 1 - \sigma(b) = p + q - 1$$
$$f_2 = \sigma(b) - p$$

Hence, on computing the Expectation:

$$\begin{aligned} \mathbb{E}_X[f_1X + f_2] &= (p+q-1)\,\mathbb{E}_X[X] + \sigma(b) - p = 0 \\ &= (p+q-1)\,\pi_1 + \sigma(b) - p = 0 \end{aligned}$$

On simplifying:

$$\sigma(b) = \frac{p}{p+q}$$
$$\implies b = \log\frac{p}{q}$$

We could attempt to understand the characterization of the points on this manifold through local perturbation analysis. However, in this work, we classify them as stationary points and leave the comprehensive characterization for future research. Therefore,

$\boldsymbol{\Gamma}_{\text{station}}(p,q) \subseteq \mathbb{R}^4$ as follows:

$$\boldsymbol{\Gamma}_{\text{station}}(p,q) \triangleq \left\{ \boldsymbol{\gamma}_{\min} = (e,w,b,a) \in \mathbb{R}^4 : e \neq 0, 1+ae^2 = 0, 1+2w|w| = 0, b = \log\frac{p}{q} \right\},$$

M.3.3. Computing the Optimal Bias

Redefining quantities:

$$f_1 = \sigma\left(\underbrace{e^2\left[\frac{1}{2}\left(1+ae^2\right)(1+2w|w|)+w|w(1+ae^2)|\right]+b}_{z_1}\right) - \sigma\left(\underbrace{e^2\left[\frac{-1}{2}\left(1+ae^2\right)(1+2w|w|)+w|w(1+ae^2)|\right]+b}_{z_2}\right) + p + q - 1$$

$$f_2 = \sigma\left(\underbrace{e^2\left[\frac{-1}{2}\left(1+ae^2\right)(1+2w|w|)+w|w(1+ae^2)|\right]+b}_{z_2}\right) - p$$

Finding the manifold of stationary points for the bias:

$$\begin{aligned}\frac{L(\boldsymbol{\theta})}{\partial b} &= \mathbb{E}_X\left[f_1 X + f_2\right] = 0\\ &= \left(\sigma(z_1)-\sigma(z_2)+p+q-1\right)\mathbb{E}_X\left[X\right]+\sigma(z_2)-p=0\\ &= \left(\sigma(z_1)-\sigma(z_2)+p+q-1\right)\pi_1+\sigma(z_2)-p=0\end{aligned}$$

Hence on grouping terms together and simplifying:

$$\begin{aligned}&\left(\sigma(z_1)-\sigma(z_2)+p+q-1\right)\pi_1 = p-\sigma(z_2)\\ \implies &\left(\sigma(z_1)-1\right)\pi_1-\sigma(z_2)\pi_1+p = p-\sigma(z_2)\\ \implies &\left(\sigma(z_1)-1\right)\pi_1 = \sigma(z_2)(\pi_1-1)\\ \implies &\frac{\sigma(z_2)}{1-\sigma(z_1)} = \frac{\pi_1}{1-\pi_1} = \frac{p}{q}\end{aligned}$$

On using the definition of the sigmoid function and rearranging:

$$\begin{aligned}&\frac{1+\exp(z_1)}{1+\exp(-z_2)} = \frac{p}{q}\\ \implies &\exp(z_1)+1 = \frac{p}{q}\left(1+\exp(-z_2)\right)\\ \implies &\exp(2z_1)+\exp(z_1) = \frac{p}{q}\exp(z_1)+\frac{p}{q}\exp(z_1-z_2)\\ \implies &\exp(z_1)^2+\exp(z_1)(1-\frac{p}{q})-\frac{p}{q}\cdot\exp(z_1-z_2)=0\end{aligned}$$

Basis the definitions of z_1 and z_2, we have $z_1 - z_2 = e^2(1+ae^2)(1+2w|w|)$. We define $A = \exp(e^2(1+ae^2)(1+2w|w|))$. Hence we obtain a quadratic equations for for $\exp(z_1)$:

$$\exp(z_1)^2+\exp(z_1)(1-\frac{p}{q})-\frac{p}{q}\cdot A = 0$$

On solving the quadratic equation for $\exp(z_1)$:

$$\exp(z_1) = \frac{1}{2}\left[\frac{p}{q} - 1 + \sqrt{\left(\frac{p}{q}-1\right)^2 + 4\cdot\frac{p}{q}\cdot A}\right] \tag{98}$$

$$\implies z_1 = \log\left(\frac{1}{2}\left[\frac{p}{q} - 1 + \sqrt{\left(\frac{p}{q}-1\right)^2 + 4\cdot\frac{p}{q}\cdot A}\right]\right) \tag{99}$$

$$\implies b_{\text{station}} = \log\left(\frac{1}{2}\left[\frac{p}{q} - 1 + \sqrt{\left(\frac{p}{q}-1\right)^2 + 4\cdot\frac{p}{q}\cdot A}\right]\right) \tag{100}$$

$$- e^2\left[\frac{1}{2}\left(1+ae^2\right)\left(1+2w|w|\right) + w|w(1+ae^2)|\right], \tag{101}$$

where b_{station} denotes points where the bias remains a stationary quantity.

M.4. Proof of 15

Proof. Recall from Lemma 13 that for $\boldsymbol{\theta} = (e, w, a) \in \mathbb{R}^3$, we have

$$\begin{aligned}
\frac{\partial L}{\partial e} &= -\mathbb{E}\left[(f_1X + f_2)\left(X - \frac{1}{2}\right)\right]\cdot 2e\left(1+ae^2\right)\left(1+2w|w|\right) \\
&\quad - \mathbb{E}\left[(f_1X + f_2)\left(X - \frac{1}{2}\right)\right]\cdot 2e^3a\left(1+2w|w|\right), \\
\frac{\partial L}{\partial w} &= -\mathbb{E}\left[(f_1X + f_2)\left(X - \frac{1}{2}\right)\right]\cdot 2e^2\left(1+ae^2\right)\left(|w| + \text{sign}\left(w\right)w\right), \\
\frac{\partial L}{\partial a} &= -\mathbb{E}\left[(f_1X + f_2)\left(X - \frac{1}{2}\right)\right]\cdot e^4\left(1+2w|w|\right),
\end{aligned}$$

where $X \in \{0,1\}$ is a Bernoulli random variable with $X \sim \text{Bern}(p/(p+q))$, and

$$\begin{aligned}
&f_1 \triangleq 1 - p - q - \phi_1 + \phi_0, \quad f_2 \triangleq p - \phi_0, \\
&\phi_1 \triangleq \sigma\left(e^2\left(\frac{1}{2}\left(1+ae^2\right)\left(1+2w|w|\right) + w|w(1+ae^2)|\right) + b_\star\right), \\
&\phi_0 \triangleq \sigma\left(e^2\left(\frac{-1}{2}\left(1+ae^2\right)\left(1+2w|w|\right) + w|w(1+ae^2)|\right) + b_\star\right),
\end{aligned}$$

where the optimal bias $b_\star$ is obtained by solving $\pi_1 f_1 + f_2 = 0$. Using the definition of the gradient flow that $\dot{\boldsymbol{\theta}} = -\nabla L(\boldsymbol{\theta})$ for $\boldsymbol{\theta} = \boldsymbol{\theta}_t$, we have

$$\begin{aligned}
\dot{w} = -\frac{L(\boldsymbol{\theta})}{\partial w} &= \mathbb{E}\left[(f_1X + f_2)\left(X - \frac{1}{2}\right)\right]\cdot 2e^2\left(1+ae^2\right)\left(|w| + \text{sign}\left(w\right)w\right) \\
\implies \frac{\dot{w}}{e^2\left(|w| + \text{sign}\left(w\right)w\right)} &= \mathbb{E}\left[(f_1X + f_2)\left(X - \frac{1}{2}\right)\right] 2\left(1+ae^2\right).
\end{aligned} \tag{102}$$

Similarly for a,

$$\begin{aligned}
\dot{a} = -\frac{L(\boldsymbol{\theta})}{\partial a} &= \mathbb{E}\left[(f_1X + f_2)\left(X - \frac{1}{2}\right)\right] e^4\left(1+2w|w|\right) \\
\implies \frac{\dot{a}}{e^4} &= \mathbb{E}\left[(f_1X + f_2)\left(X - \frac{1}{2}\right)\right]\left(1+2w|w|\right)
\end{aligned} \tag{103}$$

Likewise, for e:

$$\begin{aligned}\dot{e} = -\frac{L(\boldsymbol{\theta})}{\partial e} &= \mathbb{E}\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] 2\left(1 + ae^2\right) e\left(1 + 2w|w|\right) \\ &+ \mathbb{E}\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] \left(1 + 2w|w|\right) 2e^3 a\end{aligned} \tag{104}$$

By substituting the expressions of 102 and 103 into 104:

$$\begin{aligned}\dot{e} = &\underbrace{\mathbb{E}\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] 2\left(1 + ae^2\right) e\left(1 + 2w|w|\right)}_{102} \\ &+ \underbrace{\mathbb{E}\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] \left(1 + 2w|w|\right) 2e^3 a}_{103}\end{aligned} \tag{105}$$

We obtain:

$$\dot{e} = \frac{\dot{w}}{2e^2\left(|w| + \operatorname{sign}(w)\, w\right)} e\left(1 + 2w|w|\right) + \frac{\dot{a}}{e^4} 2e^3 a \tag{106}$$

On rearranging and simplifying:

$$\begin{aligned} e\dot{e} &= \frac{\dot{w}}{\left(|w| + \operatorname{sign}(w)\, w\right)} \left(1 + 2w|w|\right) + \dot{a} 2a \\ \implies e\dot{e} &= \frac{\dot{w}}{2\left(|w|\right)} \left(1 + 2w|w|\right) + 2a\dot{a} \end{aligned} \tag{107}$$

Integrating the above equation on both sides:

$$\begin{aligned} \int e\dot{e} &= \int \frac{\dot{w}}{4\left(|w|\right)} \left(1 + 2w|w|\right) + \int 2a\dot{a} \\ \implies \frac{e^2(t)}{2} &= \frac{\operatorname{sign}(w(t)) \cdot \log|w(t)| + w(t)^2}{2} + a(t)^2 + \frac{c}{2} \\ \implies e^2(t) &= \operatorname{sign}(w(t)) \cdot \log|w(t)| + w(t)^2 + 2a(t)^2 + c \end{aligned} \tag{108}$$

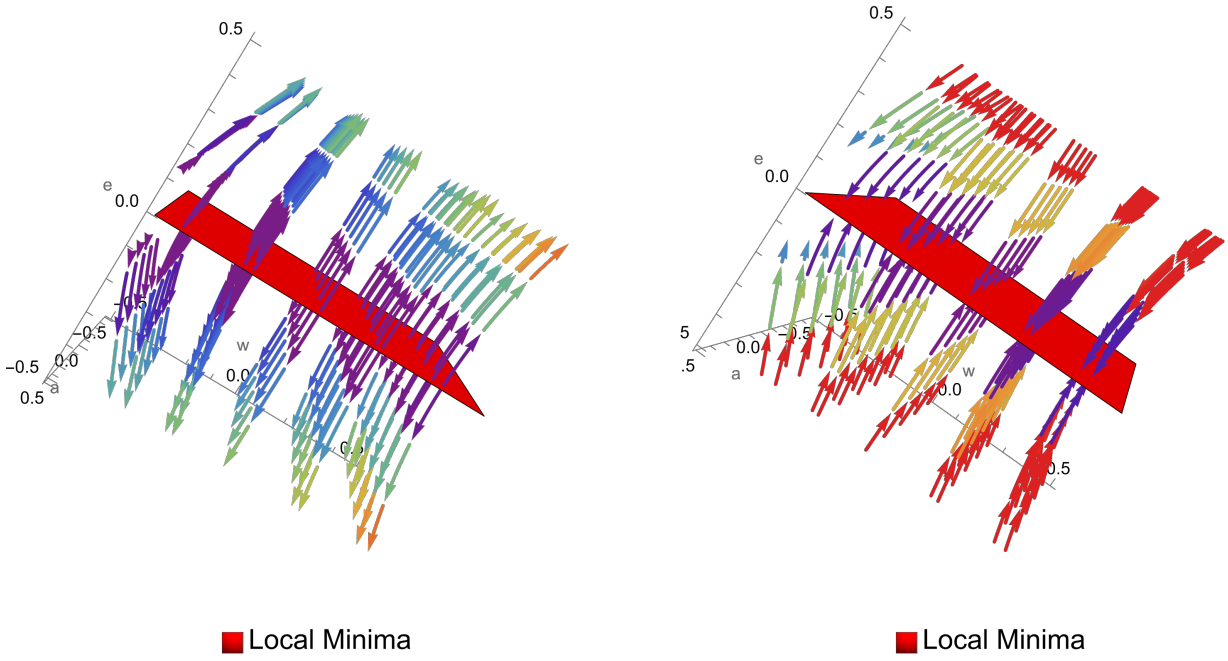
Note that here $c \in R$, is a quantity that depends on the initial conditions. Thus the energy $\mathcal{E}(\boldsymbol{\theta}_t) = \mathcal{E}(\boldsymbol{\theta}_0)$ for $w_0 \neq 0$.

□

M.5. Role of Standard Initialization

Proof. [Informal] Reproducing the derivatives for $\boldsymbol{\theta} = (e) \in \mathbb{R}^1$, assuming optimal b:

$$\begin{aligned}\frac{L(\boldsymbol{\theta})}{\partial e} &= \mathbb{E}_X\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] 2e\left(1 + ae^2\right)\left(1 + 2w|w|\right) \\ &+ \mathbb{E}_X\left[(f_1 X + f_2)\left(X - \frac{1}{2}\right)\right] 2e^3 a\left(1 + 2w|w|\right) = 0\end{aligned} \tag{109}$$



(a) Gradient flow around origin ($p + q < 1$)

(b) Gradient flow around origin ($p + q > 1$)

Figure 5: Gradient flow dynamics in $\mathbb{R}^3$, near the origin, for the transformer parameters with attention scalar a (Sec. **??**). The local minima are repellors for $p + q < 1$, while attracting for $p + q > 1$.

Where:

$$f_1 = \sigma\left(\underbrace{e^2\left[\frac{1}{2}\left(1+ae^2\right)\left(1+2w|w|\right)+w|w(1+ae^2)|\right]+b}_{z_1}\right) - \sigma\left(\underbrace{e^2\left[\frac{-1}{2}\left(1+ae^2\right)\left(1+2w|w|\right)+w|w(1+ae^2)|\right]+b}_{z_2}\right) + p + q - 1$$

$$f_2 = \sigma\left(\underbrace{e^2\left[\frac{-1}{2}\left(1+ae^2\right)\left(1+2w|w|\right)+w|w(1+ae^2)|\right]+b}_{z_2}\right) - p$$

We assume the initialization is very small, making any product of quantities in $\boldsymbol{\theta} = (e, w, a, b)$ much smaller than the individual quantities. Therefore, we can consider these products to be approximately zero. That is, $\forall x, y \in \boldsymbol{\theta}, x \geq xy \;\; \& \;\; y \geq xy \;\; \& \;\; xy \approx 0$. Hence:

$$z_1 = \sigma(b)$$
$$z_2 = \sigma(b)$$

$$f_1 = p + q - 1$$
$$f_2 = \sigma(b) - p$$

Similarly:

$$\frac{L(\boldsymbol{\theta})}{\partial e} = 2\mathbb{E}_X \left[(f_1 X + f_2)(X) \right] e \tag{110}$$

Analysing $\mathbb{E}_X \left[(f_1 X + f_2) \right]$:

$$\begin{aligned} \mathbb{E}_X \left[(f_1 X + f_2) X \right] &= (f_1 + f_2)\pi_1 \\ &= f_1 \pi_1 - f_1 \pi^2 \\ &= (p + q - 1)(\pi_1 - \pi_1^2) \end{aligned}$$

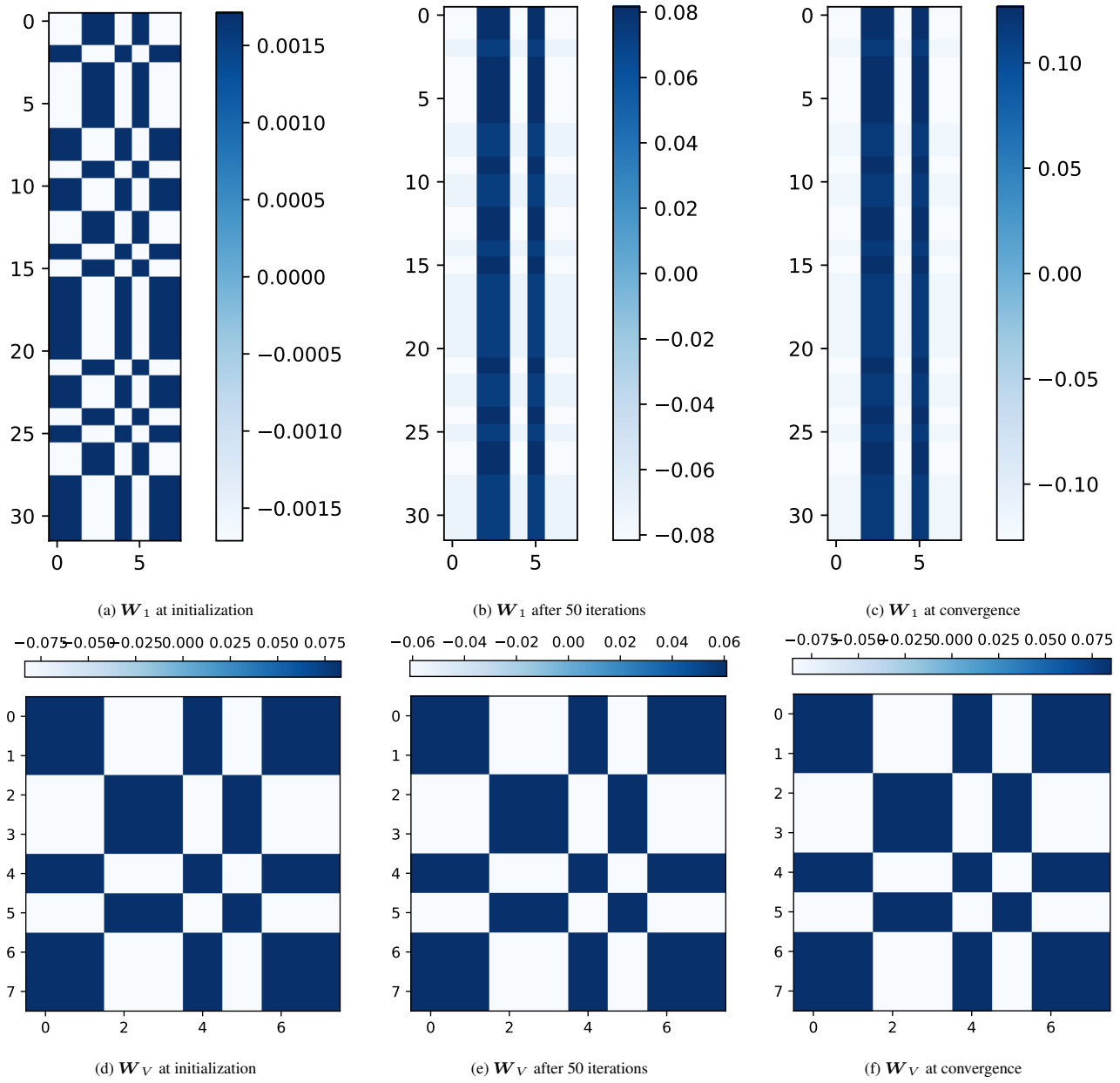
We used the fact that b is optimal in the above equations, specifically where $f_1 \pi_1 + f_2 = 0$. On computing the gradient flow:

$$\dot{e} = -\frac{L(\boldsymbol{\theta})}{\partial e} = -(p + q - 1)(\pi_1 - \pi_1^2) e \implies e = e_0 \exp(-(p + q - 1)(\pi_1 - \pi_1^2) t)$$

Since $(p + q - 1)(\pi_1 - \pi_1^2) > 0$, $e \to 0$, which denotes it converges to the local minima.

□

N. Additional empirical results



(a) $\boldsymbol{W}_1$ at initialization (b) $\boldsymbol{W}_1$ after 50 iterations (c) $\boldsymbol{W}_1$ at convergence

(d) $\boldsymbol{W}_V$ at initialization (e) $\boldsymbol{W}_V$ after 50 iterations (f) $\boldsymbol{W}_V$ at convergence

Figure 6: Evolution of parameters $\boldsymbol{W}_1$ and $\boldsymbol{W}_V$ across iterations, starting from a rank-one initialization. The parameters maintain a rank-one structure across the entire training.

O. Model architecture and hyper-parameters

Table 1: Parameters in the transformer architecture with their shape.

Parameter	Matrix shape
transformer.wte	$2 \times d$
transformer.wpe	$N \times d$
transformer.h.ln_1	$d \times 1$
transformer.h.attn.c_attn	$3d \times d$
transformer.h.attn.c_proj	$d \times d$
transformer.h.ln_2	$d \times 1$
transformer.h.mlp.c_fc	$4d \times d$
transformer.h.mlp.c_proj	$d \times 4d$
transformer.ln_f	$d \times 1$

Table 2: Settings and parameters for the transformer model used in the experiments.

Dataset	k-th order binary Markov source
Architecture	Based on the GPT-2 architecture as implemented in (Pagliardini, 2023)
Batch size	Grid-searched in $\{16, 50\}$
Accumulation steps	1
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)
Learning rate	0.001
Scheduler	Cosine
# Iterations	8000
Weight decay	1×10^{-3}
Dropout	0
Sequence length	Grid-searched in $\{512, 1024, 2048\}$
Embedding dimension	Grid-searched in $\{4, 8, 16, 32, 64\}$
Transformer layers	1
Attention heads	1
Repetitions	3 or 5