THE CROSS-ENTROPY GAP: SCALING LAWS FOR SIGN DESCENT AND GRADIENT DESCENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Adam is currently one of the most used optimizers in deep learning, yet our understanding of it remains incomplete. We investigate a simple setting, but one that is rich enough to explain the key advantage of sign-like algorithms such as Adam. We analyze next-token prediction under a linear bigram model with Zipfian token frequencies $\pi_k \propto 1/k$ for the k-th token, $k \in \mathbb{N}^*$ and study gradient descent and sign descent, a common proxy for Adam, under the cross-entropy loss. Although widely used in practice, the cross-entropy loss remains hard to analyze in theory and exact asymptotic results are rare outside restrictive settings. To address this, we introduce a technique of independent interest: a two-stage reduction that matches discrete gradient descent to gradient flow despite a diverging step-size and replaces the coupled softmax flow by a decoupled exponential flow with closed-form trajectories. For sign descent, we use a fixed step-size that balances errors due to progress and oscillations and model the oscillations as independent uniform variables, which matches experimental results. With this method, we are able to show that gradient descent takes on the order of d^2 iterations to reach an error ε, while sign descent needs only $\sqrt{\log(d)/\varepsilon}$, with only logarithmic dependence on the vocabulary size d. This gives a strong advantage to sign descent on language tasks and confirms empirically observed behavior. Our predictions are verified on synthetic data and on OpenWebText.

1 INTRODUCTION

Adam (Kingma and Ba, 2015) is the default optimizer for training large language models as it outperforms stochastic gradient descent in these settings (Zhang et al., 2020; Kunstner et al., 2023). Recent work suggests that heavy-tailed imbalance in the distribution of the tokens results in imbalanced gradient magnitudes across parameters associated with different tokens (Kunstner et al., 2024). The coordinate-wise sign-like normalization of Adam has been proposed as the key mechanism for its benefit (Balles and Hennig, 2018; Bernstein et al., 2018; Kunstner et al., 2023).

The benefit of the sign-like normalization with heavy-tailed tokens has been provably established on a linear bigram model with Zipfian token frequencies $\pi_k \propto 1/k$ under the quadratic loss as the vocabulary size d grows (Kunstner and Bach, 2025). Their results show that gradient descent (GD) requires on the order of $t \sim d$ iterations to reach a small relative error, while sign descent (SD) needs only $t \sim \sqrt{d}$. However, this analysis only holds for the square loss, whereas classification problems are typically trained with the cross-entropy loss. Under cross-entropy, Kunstner and Bach (2025) observed a substantially larger gap between SD and GD. Figure 1 shows that the gap between sign descent and gradient descent under mean squared error (MSE) is already significant, but is even larger under cross-entropy (CE) loss; gradient descent is slower on the cross-entropy loss than on the quadratic loss, while sign descent is faster. However, describing the dynamics of gradient descent on the cross-entropy loss was left open, because the softmax function couples the weights together, making it impossible to derive a closed-form expression for the iterates. Analyses of the dynamics of gradient descent on the cross-entropy loss therefore typically describe the asymptotic behavior of the margins (e.g., Soudry et al., 2018) or rely on simplified models (e.g., Cabannes et al., 2024).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

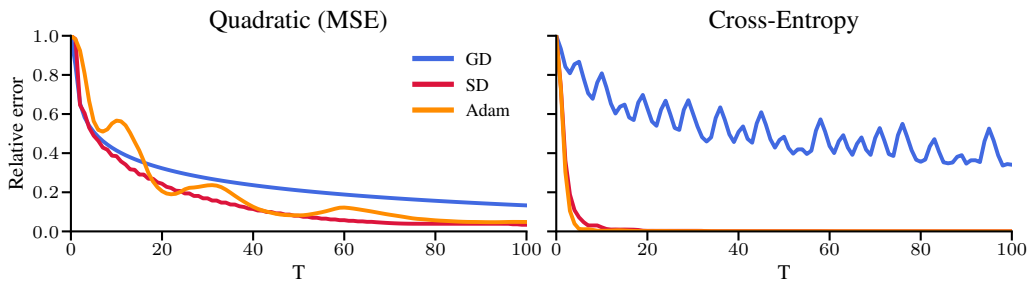

Figure 1: **The gap between GD and SD is much larger under cross-entropy loss** than under quadratic loss. Linear bigram model, $d = 500$ words, step-size tuned by grid-search for a budget T.

1.1 Contributions

The goal of this work is to characterize the performance gap between GD and SD under the cross-entropy (CE) loss. We derive scaling laws for GD and SD that show that the number of iterations required to reach a relative error of ε scales as d^2 for GD, while for SD it scales with $\sqrt{\log(d)/\varepsilon}$.

Table 1: **Gradient and sign descent have a much larger gap on cross-entropy loss than on quadratic loss.** Number of iterations T needed to reach a relative error ε as $d \to \infty, \varepsilon \to 0$.

	GD	**SD**	
Quadratic	d	$\sqrt{d/\varepsilon}$	(Kunstner and Bach, 2025)
Cross-entropy	d^2	$\sqrt{\log(d)/\varepsilon}$	This paper

The intuition for the worse performance of gradient descent on the cross-entropy loss compared to the square error is that the difficulty in the case of the square error comes from the imbalance in the inputs, while the cross-entropy loss depends on both the inputs *and* outputs. In the quadratic case, the spectrum of the Hessian depends on the distribution of the inputs, with eigenvalues corresponding to the frequencies of the input words in the vocabulary. For Zipf-distributed data, this leads to a condition number of d. In the case of the cross-entropy loss, the spectrum of the Hessian at the minimum is given by the distribution of the *joint* probability of the input and output words, leading to a condition number of d^2. The main form of our result is the following asymptotic equivalence.

Informal theorem. *As the vocabulary size $d \to \infty$, the excess loss $\mathcal{L}(T) - \mathcal{L}^*$, of the linear bigram model described later, where $\mathcal{L}^*$ is the optimal loss, after T steps of gradient descent (GD) or sign descent (SD) behaves asymptotically as follows.*

$$\text{For GD, after } T = d^\tau \text{ steps,} \qquad \mathcal{L}(T) - \mathcal{L}^* \sim \frac{(2-\tau)^3}{3}\left(\mathcal{L}(0) - \mathcal{L}^*\right), \text{ for } \tau \in [1, 2].$$

$$\text{For SD, after } T \gg \log(d) \text{ steps,} \qquad \mathcal{L}(T) - \mathcal{L}^* \sim \frac{\log(d)}{12T^2}\left(\mathcal{L}(0) - \mathcal{L}^*\right).$$

The results depend on the selection of a constant step-size, and a random residual model for SD. The cases $\tau \in [0, 1]$ for GD and T independent of d for SD are covered in Theorems 3.5 and 4.5, resp.

The main technical tool we introduce to derive this result is a reduction of the dynamics of gradient descent on the coupled cross-entropy loss to the gradient flow on a decoupled system where the softmax is replaced by a simple exponential, leading to closed-form trajectories. This result is based on the large-d limit and the heavy-tailedness of the distribution over the classes, which may be of interest to other applications of the cross-entropy loss. For sign descent, we derive conditional limits under an empirically supported random model of the oscillations. Our results (Corollary 4.6) are consistent with empirical results on synthetic data and real data from OpenWebText (Gokaslan et al., 2019).

1.2 Related Work

Adam and SGD. The performance gap between Adam and SGD has been widely observed, with multiple proposed explanations. The heavy-tailed noise hypothesis of Zhang et al. (2020) suggests gradient clipping and adaptive methods help in mitigating large outliers, but the performance gap persists under deterministic training (Kunstner et al., 2023) and disappears with a small batch size (Srećković et al., 2025; Marek et al., 2025), motivating explanations based on deterministic optimization. One such explanation is ill-conditioning, seen in the heterogeneous structure of the Hessian (Zhang et al., 2024), especially in the last classification layer with many imbalanced classes (Dong et al., 2025; Zhang et al., 2025; Zhao et al., 2025). Low-frequency classes are learned slowly, leading to the poor performance of GD (Francazi et al., 2023; Kunstner et al., 2024). We complement these works by proving a stronger separation of gradient and sign descent on the cross-entropy loss.

Dynamics under cross-entropy loss. Multi-step trajectories are difficult to obtain because the softmax couples the weights. Existing analyses instead study asymptotic properties such as the implicit bias toward max-margin solutions (Soudry et al., 2018; Gunasekar et al., 2018; Fan et al., 2025), use one-step analyses as a proxy (Cabannes et al., 2024; Kim et al., 2026b), study invariants of the dynamics (Varre et al., 2026), design special initializations that decouple the dynamics (Garrod et al., 2026), or reduce them to systems of coupled ODEs that can be integrated (Collins-Woodfin et al., 2024; Collins-Woodfin and Seroussi, 2025). Most closely related to our analysis of gradient descent, Tsiolis et al. (2026a) derive a similar asymptotic equivalence for gradient descent on a Gaussian mixture with many imbalanced components. Their analysis relies on the uniform confusion among incorrect classes to reduce the d-dimensional coupled dynamics to a solvable two-dimensional system. We instead show that a closed-form trajectory is available in the large-d limit under Zipfian data and uniform initialization. Our model allows nonuniform output-class probabilities and yields a GD timescale of $t \sim d^2$ instead of the $t \sim d$ scaling under the mixture model of Tsiolis et al. (2026a).

Scaling laws and heaviness of tails. Power-law source and capacity assumptions are widely used to describe asymptotic learning behavior (Caponnetto and De Vito, 2007). Related work derives scaling laws from power-law decay in the spectrum and target coefficients (e.g., Berthier et al., 2020; Cui et al., 2021; Lyu et al., 2025; Paquette et al., 2024). These analyses typically operate in a power law regime with a large exponent $\alpha > 1$. Our Zipfian setting has heavier tails, with $\alpha = 1$, which produces dimension-dependent iteration complexity and might be a more relevant description of the behavior observed when changing the vocabulary size (Gowda and May, 2020; Tao et al., 2024).

Sign-based optimization. Closest to our optimizer comparison, Yadav et al. (2025) use ℓ_∞ smoothness to derive improved bounds for SD over normalized GD when fitting a Zipfian unigram distribution under KL divergence. This smoothness argument does not capture our bigram setting, where the relevant ℓ_∞ smoothness constant scales with d. We instead derive asymptotically exact expressions for both gradient and sign descent, revealing a larger separation in the bigram setting.

2 Problem Set-up

We start by defining the linear bigram problem and describing our strategy for the analysis.

Definition 2.1 (Linear bigram model)**.** *We consider a linear bigram model where $\boldsymbol{x}_i, \boldsymbol{y}_i \in \{0,1\}^d$ for $i \in \{1, \dots, n\}$ are the one-hot encodings of the current token and its corresponding next token. We aim to minimize the empirical cross-entropy loss over the dataset to learn the conditional distribution of the next token given the current one, parameterized by the weight matrix $\boldsymbol{W} \in \mathbb{R}^{d\times d}$:*

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{W}; \boldsymbol{x}_i, \boldsymbol{y}_i), \quad \textit{where} \quad \ell(\boldsymbol{W}; \boldsymbol{x}, \boldsymbol{y}) = \log\left(\sum_{j=1}^{d} \exp((\boldsymbol{W}\boldsymbol{x})_j)\right) - \boldsymbol{y}^\top \boldsymbol{W}\boldsymbol{x}.$$

In what follows, we assume the infinite data limit ($n \to \infty$). Thus, the empirical risk converges to the expected loss over the true distribution. We assume this distribution to be the Zipfian distribution to reflect the frequency of words in natural language (Piantadosi, 2014). This approximation is discussed and verified empirically by Kunstner and Bach (2025) on the dataset OpenWebText.

Assumption 2.2 (Zipf assumption)**.** *We assume the frequency of words over the vocabulary $\pi_1, ..., \pi_d$ and the conditional distribution of the next word j given the current word k, $\pi_{j\,|\,k}$, follow Zipf's law,*

$$\pi_k = \frac{1}{kH_d} \text{ for all } k, \quad \text{and} \quad \pi_{k|j} = \frac{1}{kH_d} \text{ for all } j, k, \quad \text{where} \quad H_d = \sum_{k=1}^{d} \frac{1}{k} \sim \log(d).$$

We assume that the conditional distributions are sorted and identical for all input words to simplify notation, but this assumption is made without loss of generality (we could allow each conditional distribution to follow Zipf's law after a permutation of the indices, such that each input word has a different most likely output word). This permutation does not change the dynamics of the loss under gradient or sign descent. Under Assumption 2.2 the objective becomes the following expected loss.

Proposition 2.3 (Closed-form Expected Loss and Gradient)**.** *The loss can be written as:*

$$\mathcal{L}(\boldsymbol{W}) = -\sum_{j=1}^{d}\sum_{k=1}^{d} \pi_j \pi_{k|j} \log\left(\operatorname{softmax}(\boldsymbol{W}_{:,j})_k\right), \text{ where } \operatorname{softmax}(\boldsymbol{W}_{:,j})_k = \frac{e^{\boldsymbol{W}_{k,j}}}{\sum_{l=1}^{d} e^{\boldsymbol{W}_{l,j}}}, \quad (1)$$

where $\boldsymbol{W}_{:,j}$ denotes the j-th column of $\boldsymbol{W}$. The gradient of $\mathcal{L}(\boldsymbol{W})$ with respect to $\boldsymbol{W}_{k,j}$ is

$$\frac{\partial \mathcal{L}(\boldsymbol{W})}{\partial \boldsymbol{W}_{k,j}} = \pi_j \left(\operatorname{softmax}(\boldsymbol{W}_{:,j})_k - \pi_{k|j}\right).$$

The solution is such that $\boldsymbol{W}^*$ satisfies $\operatorname{softmax}(\boldsymbol{W}^*_{:,j}) = \pi_{\cdot|j}$ for all j.

As the loss at initialization diverges as $d \to \infty$, we use the relative error to measure progress.

Definition 2.4 (Relative Error)**.** *As $\mathcal{L}(\boldsymbol{W}(0)) - \mathcal{L}^* \sim \frac{1}{2}\log(d)$, we work with the relative error*

$$r(\boldsymbol{W}(t)) := \frac{\mathcal{L}(\boldsymbol{W}(t)) - \mathcal{L}^*}{\mathcal{L}(\boldsymbol{W}(0)) - \mathcal{L}^*} = \frac{1}{\mathcal{L}(\boldsymbol{W}(0)) - \mathcal{L}^*} \sum_{j=1}^{d}\sum_{k=1}^{d} \pi_j \pi_{k|j} \log\left(\frac{\pi_{k|j}}{\operatorname{softmax}(\boldsymbol{W}_{:,j}(t))_k}\right),$$

which starts at $r(\boldsymbol{W}(0)) = 1$ and converges to $r(\boldsymbol{W}^*) = 0$.

Plan for the analysis. We outline the main ideas used in the analysis of gradient and sign descent.

For gradient descent, deriving a closed-form trajectory under CE loss is impossible in general. Our approach is to get an asymptotic result as $d \to \infty$ for Zipf-distributed data. We show that gradient descent is asymptotically equivalent to gradient flow, and that gradient flow on the CE loss is asymptotically equivalent to a simplified flow where the softmax is replaced by an exponential. This substitution removes the coupling between the weights, and allows for a closed-form solution.

For sign descent, the weights are constrained to a discrete grid of mesh size η, so sign descent cannot converge with a fixed step-size. We derive a constant step-size for a given budget T that balances errors due to oscillations and under-optimization, and simplify the analysis by modeling the oscillatory phase as a random process. We confirm that our predictions match empirical observations.

3 Dynamics of Gradient Descent

We can now study the GD dynamics, given by the following update for each weight:

$$\boldsymbol{W}_{k,j}(t+1) = \boldsymbol{W}_{k,j}(t) - \eta\pi_j \left(\operatorname{softmax}(\boldsymbol{W}_{:,j}(t))_k - \pi_{k|j}\right). \quad (2)$$

The equation decouples with respect to the columns of $\boldsymbol{W}$, meaning that the dynamics of each $\boldsymbol{W}_{:,j}$ can be studied independently.

Setting the Step-Size. We restrict the analysis to constant step-sizes for which gradient descent is globally convergent. Typical analyses of gradient descent use step-size $\eta \leq 1/L$ where L is the smoothness constant of the loss. Instead, we use $\eta \leq 2/\lambda^*_{\max}$ where $\lambda^*_{\max}$ is the largest eigenvalue of the Hessian at the *optimum*, which is required for local stability of gradient descent at the optimum. In separable logistic regression, arbitrarily large step-sizes can still converge and may improve optimization through non-monotone dynamics (Wu et al., 2024). By contrast, our problem has a

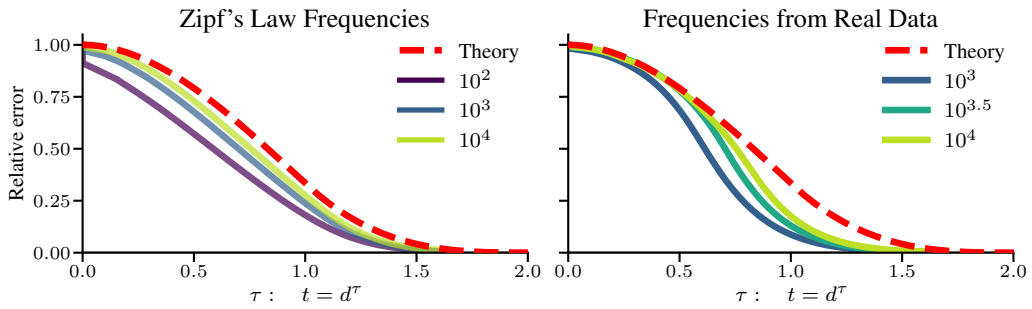

Figure 2: The empirical error converges to our theoretical limit as d increases. Global relative error as a function of the rescaled time $\tau = \frac{\log(t)}{\log(d)}$ across different dimensions, comparing GD dynamics to our theory on synthetic data (**Left**) and real data from the OpenWebText dataset (**Right**).

finite optimum, where step-sizes above the local stability threshold lead to instability (Meng et al., 2024). We state our results for $\eta = 1/\lambda^*_{\max}$, which is within a factor of two of the largest stable step-size.[1] We verify numerically that this value is within a constant factor of the best step-size by grid search (see Figure 7, Appendix C).

Proposition 3.1. *The largest eigenvalue of the Hessian at the optimum is* $\lambda^*_{\max} \sim 1/H_d^2$.

The proof is given in Appendix C. Note that $\eta \sim \log(d)^2$ diverges as $d \to \infty$ and is much larger than the maximum step-size allowed by global smoothness arguments, which is $\eta \leq 8$.

3.1 Continuous Approximation via Gradient Flow

To simplify the analysis, we consider the following gradient flow initialized at $\boldsymbol{W}^{\mathrm{F}}(0)$ with update

$$\dot{\boldsymbol{W}}^{\mathrm{F}}_{k,j}(t) = -\eta\pi_j \left(\frac{e^{\boldsymbol{W}^{\mathrm{F}}_{k,j}(t)}}{Z_j(t)} - \pi_{k|j} \right), \qquad \text{where} \quad Z_j(t) = \sum_{k=1}^{d} e^{\boldsymbol{W}^{\mathrm{F}}_{k,j}(t)}. \tag{3}$$

Note that we keep the step-size in the definition of the flow to have $\boldsymbol{W}(t) \approx \boldsymbol{W}^{\mathrm{F}}(t)$ for $t \in \mathbb{N}$. We can show that the discrete and continuous dynamics are asymptotically equivalent.

Lemma 3.2 (Equivalence of Discrete and Continuous Dynamics)**.** *The difference in relative error between the discrete* $\boldsymbol{W}(t)$ *and continuous* $\boldsymbol{W}^{\mathrm{F}}(t)$ *started at a uniform initialization vanishes as* $d \to \infty$*; the following bound holds uniformly for all* $t > 0$*:*

$$|r(\boldsymbol{W}(t)) - r(\boldsymbol{W}^{\mathrm{F}}(t))| = O\left(\frac{\log(\log(d))}{\log(d)} \right). \tag{4}$$

Proof sketch. The full proof is deferred to Appendix D. This result might seem counter-intuitive as the step-size $\eta \sim \log^2(d)$ diverges. The result holds because the effective step-size for column j is $\eta_j = \eta\pi_j \sim \log(d)/j$, which is small for the vast majority of j, leading to a negligible discretization error. For the frequent words with a large effective step-size, we use the fact that Zipf's law is heavy-tailed and that the majority of weight of the distribution is concentrated on its tail. The error associated with the first $\lfloor \log^2(d) \rfloor$ words has a vanishing impact on the total relative error as $d \to \infty$. □

With this approximation result, we can derive an implicit formula for gradient flow, proven in Appendix D.1, which will also apply to the discrete dynamics.

Proposition 3.3 (Exact Trajectory of the Gradient Flow)**.** *Under the gradient flow equation (3) with uniform initialization* $W_{k,j}(0) = -\log(d)$*, the exact trajectory is given by:*

$$\exp(\boldsymbol{W}_{k,j}(t)) = \frac{\exp\left(t\eta\pi_j\pi_{k|j}\right)}{d + \eta\pi_j \int_0^t \frac{\exp\left(s\eta\pi_j\pi_{k|j}\right)}{Z_j(s)}\,\mathrm{d}s}, \quad \textit{where} \quad Z_j(s) = \sum_{l=1}^{d} \exp(\boldsymbol{W}_{l,j}(s)). \tag{5}$$

[1] A larger step-size of $\eta = c/\lambda^*_{\max}$ for $c \in (0, 2)$ would only change the rate by a constant factor; our analysis reduces to gradient flow, for which multiplying the step-size by c is equivalent to rescaling time by c.

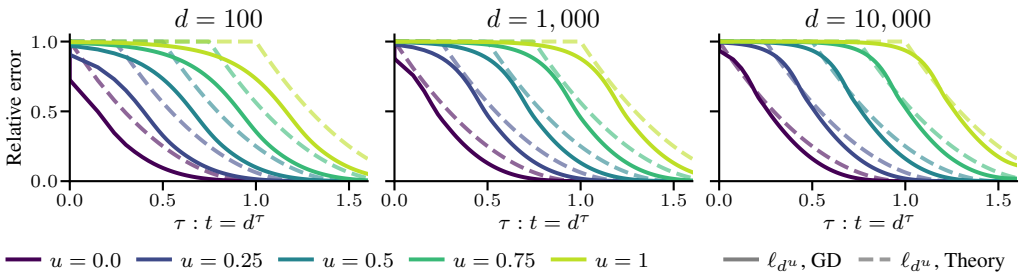

Figure 3: **Sequential learning of words under gradient descent.** Words are learned in waves, the more frequent a word is, the sooner it is learned. Gradient descent dynamics of the per-word relative error for different ratios u_j where $u_j = \log(j)/\log(d)$.

While this result is implicit, as it depends on $Z_j(s)$, we can show that $Z_j(s)$ is bounded between 1 and H_d (Lemma D.2). We can further show that taking the same trajectory but replacing $Z_j(s)$ with 1 produces a trajectory which effectively corresponds to the dynamics under a decoupled exponential loss, which is asymptotically equivalent to the dynamics of gradient descent with cross-entropy.

Lemma 3.4 (Equivalence of softmax and Exponential Dynamics)**.** *Let $\boldsymbol{W}^{\exp}(t)$ denote the weight trajectory governed by $\dot{\boldsymbol{W}}^{\exp}_{k,j}(t) = -\eta\pi_j(\exp(\boldsymbol{W}^{\exp}_{k,j}(t)) - \pi_{k|j})$. Then, the difference in relative error between the exact softmax continuous flow $\boldsymbol{W}^{\mathrm{F}}(t)$ and the exponential flow $\boldsymbol{W}^{\exp}(t)$ vanishes:*

$$|r(\boldsymbol{W}^{\mathrm{F}}(t)) - r(\boldsymbol{W}^{\exp}(t))| = O\left(\frac{\log(\log(d))}{\log(d)}\right). \tag{6}$$

The proof is given in Appendix D.6. What remains is to analyze the asymptotic limit of the relative error under the exponential flow, which gives the following result.

Theorem 3.5 (Asymptotic Limit of the Relative Error)**.** *Under the time scaling $t = d^\tau$ with $\tau \in [0, 2]$, the relative error converges as $d \to \infty$ to:*

$$\lim_{d\to\infty} r(\boldsymbol{W}(d^\tau)) = 2\iint_{\tau<u+v} (1-v)\ \mathrm{d}u\ \mathrm{d}v = \begin{cases} 1 - \tau^2 + \frac{\tau^3}{3} & \text{if } \tau \in [0,1], \\ \frac{(2-\tau)^3}{3} & \text{if } \tau \in [1,2]. \end{cases}$$

Proof sketch. The full proof is deferred to Appendix D.6. Replacing $\boldsymbol{W}^{\exp}$ by its limit in the equation for the loss (1), using $\mathcal{L}(\boldsymbol{W}(0)) - \mathcal{L}^* \sim \log(d)/2$ and that $1 \le Z \le H_d$ (Lemma D.2), we have:

$$r(\boldsymbol{W}^{\exp}(t)) \sim \frac{2}{\log(d)} \sum_{j=1}^{d}\sum_{k=1}^{d} \frac{1}{jkH_d^2}\left[\log(\pi_k) - t\eta\pi_j\pi_{k|j} + \log\left(d + \frac{1}{\pi_{k|j}}\left(e^{t\eta\pi_j\pi_{k|j}} - 1\right)\right)\right],$$

$$= \frac{2}{\log(d)} \sum_{j=1}^{d}\sum_{k=1}^{d} \frac{1}{jkH_d^2} \log\left(1 - e^{-\frac{t}{jk}} + e^{-\frac{t}{jk}}\frac{d}{kH_d}\right).$$

Then, we approximate the sum with a continuous integral (see Proposition D.17), yielding

$$r(\boldsymbol{W}^{\exp}(t)) \sim \frac{2}{\log(d)} \int_1^d \int_1^d \frac{1}{xyH_d^2} \log\left(1 - e^{-\frac{t}{xy}} + e^{-\frac{t}{xy}}\frac{d}{yH_d}\right) \mathrm{d}x\ \mathrm{d}y.$$

We use the change of variables $x = d^u$ and $y = d^v$, and the rescaling of time $t = d^\tau$. Because $\mathrm{d}x = \log(d)d^u\ \mathrm{d}u$ and $\mathrm{d}y = \log(d)d^v\ \mathrm{d}v$, the resulting $\log(d)^2 d^{u+v}$ terms cancel with xyH_d^2:

$$r(\boldsymbol{W}^{\exp}(d^\tau)) \sim \frac{2}{\log(d)} \int_0^1 \int_0^1 \log\left(1 - e^{-d^{\tau-u-v}} + e^{-d^{\tau-u-v}}\frac{d^{1-v}}{H_d}\right) \mathrm{d}u\ \mathrm{d}v.$$

Then, two distinct behaviors emerge depending on whether $\tau < u+v$ or $\tau > u+v$. When $\tau > u+v$, the integrand goes to 0, meaning that the weights (j,k) have been learned since they produce zero error. Conversely, when $\tau < u+v$, the logarithm scales as $(1-v)\log(d)$. When multiplied by the global $2/\log(d)$ prefactor, this region contributes $2(1-v)$ to the integral. Thus:

$$\lim_{d\to\infty} r(\boldsymbol{W}^{\exp}(d^\tau)) = 2\iint_{\tau<u+v} (1-v)\ \mathrm{d}u\ \mathrm{d}v.$$

The result applies to discrete gradient descent on the cross-entropy loss by Lemmas 3.2 and 3.4 □

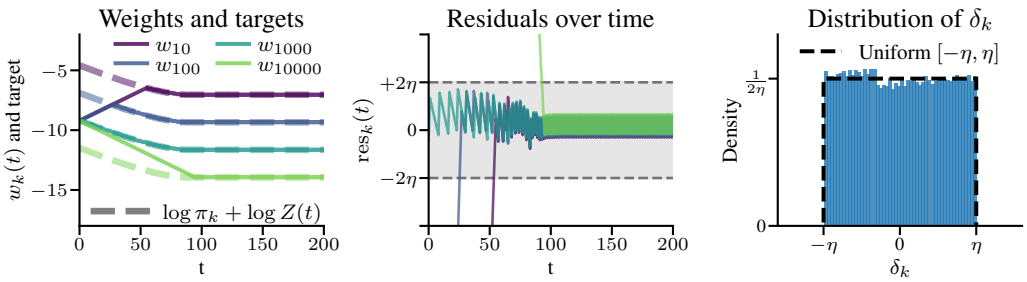

Figure 4: **Dynamics of sign descent: The weights move linearly until they catch their targets. (Left)** The weights track their associated targets $\log(\pi_k) + \log(Z(t))$ over time (shown as dashed lines). **(Center)** The residuals stay contained within the bound $[-2\eta, 2\eta]$. **(Right)** Empirical distribution of δ_k vs. the uniform assumption.

3.2 Column-by-column Analysis and Sequential Learning

The following decomposition of the problem into a per-word loss will be useful for the analysis as the dynamics separate along the columns of $\boldsymbol{W}$, leading to independent dynamics for each word.

Definition 3.6 (Per-word Error)**.** *For each j, we define the per-word loss and its optimum:*

$$\ell_j(\boldsymbol{W}) := -\sum_{k=1}^{d} \pi_{k|j} \log\left(\mathrm{softmax}(\boldsymbol{W}_{:,j})_k\right), \quad \textit{and} \quad r_j(\boldsymbol{W}(t)) := \frac{\ell_j(\boldsymbol{W}(t)) - \ell_j^*}{\mathcal{L}(\boldsymbol{W}(0)) - \mathcal{L}^*},$$

where $\ell_j^* := -\sum_{k=1}^{d} \pi_{k|j} \log \pi_{k|j}$ is the optimal per-word loss. This gives the per-input word decomposition $\mathcal{L}(\boldsymbol{W}) = \sum_{j=1}^{d} \pi_j \ell_j(\boldsymbol{W})$ and $r(\boldsymbol{W}) = \sum_{j=1}^{d} \pi_j r_j(\boldsymbol{W})$.

Since the algorithm decouples the columns of the weight matrix, it is natural to analyze the relative error of a single context column j. To study a specific column in the high-dimensional limit, we parameterize it by its log-rank, that is we consider a column $j = d^u$ for $u \in [0, 1]$.

Proposition 3.7 (Limit of the Per-Context Relative Error)**.** *Under the time scaling $t = d^\tau$, the relative error associated with input word $j = d^u$ for $u \in (0, 1)$ converges to:*

$$\lim_{d\to\infty} r_{d^u}(\boldsymbol{W}(d^\tau)) = (1 - \max(0, \tau - u))^2 \mathbb{1}_{\{\tau - u \le 1\}}.$$

The proof is in Appendix D.7. Figure 3 confirms our theory. Words are learned sequentially: word j is not learned ($r_j(\tau) = 1$) until time reaches $\tau = u_j$ ($t = j$) and is learned at $\tau = 1 + u_j$ ($t = d \cdot j$). Very recently, (Tsiolis et al., 2026b) analyzed scaling laws for logistic regression and also identified this phenomenon of sequential learning.

4 Dynamics of Sign Descent

We now extend our analysis to sign descent, with the following update rule.

Definition 4.1 (Sign Descent Update)**.** *The update rule is given by:*

$$\begin{aligned} \boldsymbol{W}_{k,j}(t+1) &= \boldsymbol{W}_{k,j}(t) - \eta \,\mathrm{sign}\big(\pi_j\big(\mathrm{softmax}(\boldsymbol{W}_{:,j}(t))_k - \pi_{k|j}\big)\big), \\ &= \boldsymbol{W}_{k,j}(t) - \eta \,\mathrm{sign}\big(\boldsymbol{W}_{k,j}(t) - \log(\pi_{k|j}) - \log Z_j(t)\big). \end{aligned} \tag{7}$$

The sign update does not depend on the context j as the sign ignores the magnitude of π_j, and under our assumption about the conditional distributions, we have $\pi_{k|j} = \pi_{k\,|\,j'}$ (Assumption 2.2). We thus omit the index j in this section. The dynamics of sign descent are first linear, as the step is always η, then oscillate around the target, as illustrated in Figure 4. We can formalize this behavior in the following proposition, proved in Appendix E.1.

Proposition 4.2 (Weight Dynamics)**.** *Let $\log(\pi_k) + \log Z(t)$ be the (moving) target of weight $W_k(t)$. Under sign descent, W_k always increases (or always decreases) until it reaches its target, then oscillates around it. That is, if $W_k(t) - \log(\pi_k) - \log Z(t)$ first changes sign at T_k,*

$$W_k(t) = W_0 \pm \eta t \textit{ for } t \le T_k, \qquad \textit{and} \qquad |W_k(t) - \log(\pi_k) - \log Z(t)| \le 2\eta \textit{ for } t > T_k.$$

The system reaches the oscillatory regime when all weights oscillate, at $T_{\mathrm{osc}} := \max\{T_1, \ldots, T_d\}$.

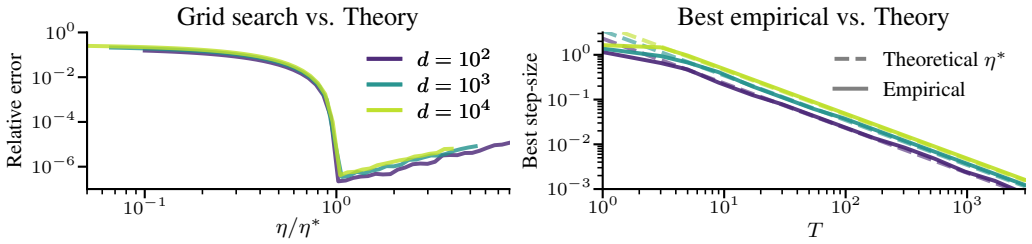

Figure 5: **The theoretical step-size $\eta^* = \log(d)/2T$ is close to optimal. Left:** Excess loss vs. step-size η for varying dimensions d at $T = 1000$. **Right:** Optimal η from grid-search vs. budget T.

Setting the Step-Size. Due to the oscillations, the error cannot vanish unless the step-size η goes to 0 as $T \to \infty$. To minimize the error of sign descent after a fixed budget of T steps, the step-size η must be chosen to trade-off between two sources of errors. If η is too large, the weights will oscillate around their targets with an amplitude proportional to η, leading to a large error. If η is too small, the error will be large because the weights are still far from their targets after T steps.

Minimizing the error exactly for a finite dimension d and budget T is difficult because the dynamics of sign descent on the logistic regression are not linear and are coupled across weights, leading to a much more complex analysis than for the gradient descent case (Kunstner and Bach, 2025). As a heuristic, we compute a lower bound on the step-size η needed to guarantee that the weights reach their targets when optimized using a gradient flow approximation of the dynamics.

Proposition 4.3 (Minimum step-size for sign flow)**.** *For sign flow with step-size η,*

$$\dot{\boldsymbol{w}}(t) = -\eta \operatorname{sign}(\boldsymbol{w}(t) - \log \pi - \log Z(t)),$$

the step-size needs to be at least $\eta \geq \log(d)/2T$ to hit the solution at time T.

The proof is given in Appendix E.2. While this is only a lower bound for the continuous-time version of the algorithm, it appears to be a good approximation for the discrete version of the dynamics. We verify empirically that η^* is close to the empirical minimizer of the relative error via grid search in Figure 5, and conjecture that it is optimal for the large d, large T regime.

4.1 Relative Error Analysis

Once the system has entered the oscillatory regime, the weights oscillate around their targets. To simplify, we model the oscillatory phase as a random process, where the residuals are i.i.d. uniform random variables of magnitude at most η. This is the only source of randomness in the result below.

Assumption 4.4 (Uniform Residuals)**.** *We assume that, with the step-size $\eta = \log(d)/(2T)$, all weights have reached the oscillatory regime at time T and satisfy*

$$W_k(T) = \log \pi_k + c_t + \delta_k, \qquad k = 1, \dots, d,$$

where c_t is a constant and $\delta_1, \dots, \delta_d$ are i.i.d. with distribution $\mathcal{U}([-\eta, \eta])$.

The main purpose of this assumption is to avoid having to track the exact trajectory of the weights, which depends heavily on the initial conditions. Empirically, the assumption appears to be a good approximation, as the distribution of the residuals is close to uniform, see Figure 4 (Right), and this assumption is sufficient to obtain a closed-form expression for the limit of the relative error that is a good approximation of the empirical results.

Theorem 4.5 (Asymptotic Relative Error of Sign Descent)**.** *Under Assumption 4.4, sign descent with step-size $\eta = \log(d)/(2T)$, as $d \to \infty$, the relative error satisfies*

$$r(\boldsymbol{W}(T)) \;=\; \frac{2}{\log(d)} \log\left(\frac{\sinh \eta}{\eta}\right) \big(1 + o_{\mathbb{P}}(1)\big).$$

Proof sketch. The full proof is deferred to Appendix E.3. We can express the excess loss as $\mathcal{L}(T) - \mathcal{L}^* = -\sum_k \pi_k \delta_k + \log(\sum_j \pi_j \exp(\delta_j))$. Concentration inequalities give that the first term converges to 0 in probability while the second concentrates around $\log(\sinh(\eta)/\eta)$. Normalizing by the initial loss $\mathcal{L}(\boldsymbol{W}(0)) - \mathcal{L}^* \sim \log(d)/2$ gives the result. □

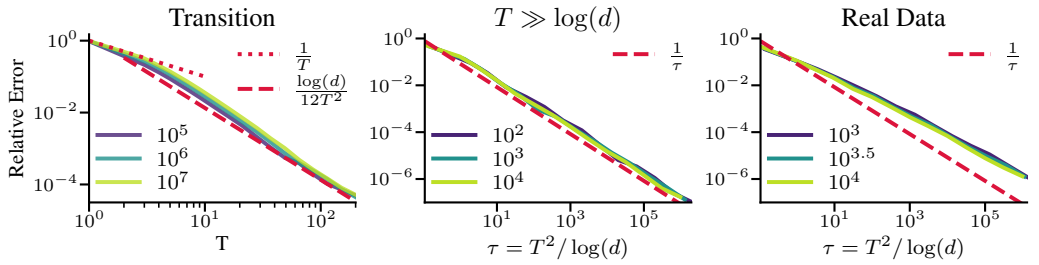

Figure 6: Empirical results match our theory: the relative error in the oscillatory regime is $\frac{\log(d)}{12T^2}$ for T sufficiently large and $1/T$ when T is small. Final relative error plotted against the final time for large d using $\eta = \log(d)/2T$ (**Left**), and using a grid-search for the step-size on synthetic data (**Center**) and real data (**Right**).

Corollary 4.6 (Different regimes)**.** *Depending on the relationship between T and d, we have the simplified asymptotic equivalences:*

$$\text{If } T \ll \log(d), \qquad r(\boldsymbol{W}(T)) \sim \frac{1}{T} \qquad \text{in probability.}$$

$$\text{If } T \gg \log(d), \qquad r(\boldsymbol{W}(T)) \sim \frac{\log(d)}{12\,T^2} \qquad \text{in probability.}$$

Proof. **If** $T \ll \log(d)$, then $\eta = \log(d)/(2T) \to \infty$ and $\log(\sinh \eta/\eta) \sim \eta$, giving that $(2/\log(d))\,\log(\sinh \eta/\eta) \sim 2\eta/\log(d) = 1/T$.

If $T \gg \log(d)$, then $\eta = \log(d)/(2T) \to 0$, and by Taylor expansion $\log(\sinh \eta/\eta) = \eta^2/6 + O(\eta^4)$, giving that $r(\boldsymbol{W}(T)) = (2/\log(d)) \cdot \eta^2/6 \cdot (1 + o_{\mathbb{P}}(1)) \sim \log(d)/(12T^2)$ in probability. □

The first regime corresponds to scaling the vocabulary size d to infinity while keeping the budget T fixed or growing more slowly than $\log(d)$. However, it is not realistic in practice because $\log(d)$ grows slowly; even with $d = 10^{10}$, it would only describe a budget of $T \ll 23$ iterations. This switch is confirmed experimentally in Figure 6; the $1/T$ behavior is visible only for small T at $d = 10^8$ before the transition to the $\log(d)/(12T^2)$ regime that governs the rest of the dynamics, explaining the $1/T^2$ scaling observed empirically by Kunstner and Bach (2025).

5 Conclusion

We presented scaling laws for GD and SD as the vocabulary size grows under the cross-entropy loss. To do so, we showed that under a power-law with exponent 1, the dynamics of GD are asymptotically equivalent to a continuous-time flow with a simpler exponential loss that decouples everything, leading to the same behavior asymptotically. The analysis of the linear bigram model shows that the sign normalization mitigates the difficulty encountered by gradient descent. In particular, sign descent depends only weakly on the dimension in the asymptotic limit, and not at all in the $T \ll \log(d)$ regime, which is surprising given the behavior of gradient descent. In the more realistic regime where $\eta = \log(d)/2T$ is small, the number of iterations scales as $\sqrt{\log(d)/\varepsilon}$, which remains almost independent of the dimension as $\log(d)$ grows slowly with d. Our analysis shows that the gap between gradient and sign descent is much larger for the CE loss than for the MSE loss.

However, our approach has limitations. We only treat gradient and sign descent in the deterministic setting for simplicity. An interesting direction is to combine our gradient-flow analysis, which handles the coupling induced by the cross-entropy loss, with tools for analyzing stochastic gradient descent and gradient flow (Kim et al., 2026a; Xiao et al., 2025; Bian et al., 2026). Studying the impact of momentum would also be interesting; we conjecture that momentum could improve the dimension dependence of gradient descent from d^2 to d. Finally, the linear bigram model is deliberately simple. It remains unclear how imbalances in the inputs and outputs interact with depth in more complex models, and whether limited capacity—such as network width that grows more slowly than the vocabulary size, as is common in practice— changes the scaling behavior of optimizers.

AI USE STATEMENT

In this work, we used generative AI tools to: provide a fewingredients for proving mathematical claims, assist in the writing of proofs, assist with translation, and implement methods in code.

We have not used generative AI tools to: generate synthetic datasets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, clean and reformat datasets, support qualitative and thematic data analysis or interpret results.

Additionally, we used generative AI tools to: create or modify scientific figures, create or edit software code, refine and polish the text for the final paper and review the final paper to check for mistakes.

We have reviewed all AI-assisted work. Both the code and the mathematical proofs were reviewed by their direct author and checked by a co-author. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Balles, Lukas and Philipp Hennig (2018). "Dissecting Adam: The Sign, Magnitude and Variance of Stochastic Gradients". In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research, pp. 404–413.

Bernstein, Jeremy, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar (2018). "signSGD: Compressed Optimisation for Non-Convex Problems". In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80, pp. 560–569.

Berthier, Raphaël, Francis Bach, and Pierre Gaillard (2020). "Tight Nonparametric Convergence Rates for Stochastic Gradient Descent under the Noiseless Linear Model". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33.

Bian, Zeyu, Dmitriy Drusvyatskiy, and Tianhao Wang (2026). "Stochastic Gradient Descent on the Linear Bigram Model: Bias-Variance Scaling and Critical Batch Size". In: *ICML Workshop on High-dimensional Learning Dynamics (HILD)*.

Cabannes, Vivien, Berfin Simsek, and Alberto Bietti (2024). "Learning Associative Memories with Gradient Descent". In: *International Conference on Machine Learning (ICML)*.

Caponnetto, A. and E. De Vito (2007). "Optimal Rates for the Regularized Least-Squares Algorithm". In: *Foundations of Computational Mathematics* 7.3, pp. 331–368.

Collins-Woodfin, Elizabeth, Courtney Paquette, Elliot Paquette, and Inbar Seroussi (2024). "Hitting the High-dimensional notes: an ODE for SGD learning dynamics on GLMs and multi-index models". In: *Information and Inference: A Journal of the IMA* 13.4, iaae028.

Collins-Woodfin, Elizabeth and Inbar Seroussi (2025). *Exact Dynamics of Multi-class Stochastic Gradient Descent*. Preprint. arXiv/2510.14074.

Cui, Hugo, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová (2021). "Generalization Error Rates in Kernel Regression: The Crossover from the Noiseless to Noisy Regime". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 34, pp. 10131–10143.

Dong, Zhaorui, Yushun Zhang, Jianfeng Yao, and Ruoyu Sun (2025). "Towards Quantifying the Hessian Structure of Neural Networks". In: *NeurIPS Workshop on Optimization for Machine Learning (OPT)*.

Fan, Chen, Mark Schmidt, and Christos Thrampoulidis (2025). "Implicit Bias of Spectral Descent and Muon on Multiclass Separable Data". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Francazi, Emanuele, Marco Baity-Jesi, and Aurelien Lucchi (2023). "A Theoretical Analysis of the Learning Dynamics under Class Imbalance". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202, pp. 10285–10322.

Garrod, Connall, Jonathan P. Keating, and Christos Thrampoulidis (2026). "Diagonalizing the Softmax: Hadamard Initialization for Tractable Cross-Entropy Dynamics". In: *ICML Workshop on Weight-Space Symmetries: from Foundations to Practical Applications*.

Gokaslan, Aaron, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex (2019). *OpenWebText Corpus*. `http://Skylion007.github.io/OpenWebTextCorpus`.

Gowda, Thamme and Jonathan May (2020). "Finding the Optimal Vocabulary Size for Neural Machine Translation". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3955–3964.

Gunasekar, Suriya, Jason Lee, Daniel Soudry, and Nathan Srebro (2018). "Characterizing Implicit Bias in Terms of Optimization Geometry". In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80, pp. 1832–1841.

Kim, Jihwan, Dogyoon Song, and Chulhee Yun (2026a). "Scaling Laws of SignSGD in Linear Regression: When Does It Outperform SGD?" In: *International Conference on Learning Representations (ICLR)*.

Kim, Juno, Eshaan Nichani, Denny Wu, Alberto Bietti, and Jason D. Lee (2026b). *Sharp Capacity Scaling of Spectral Optimizers in Learning Associative Memory*. Preprint. arXiv/2603.26554.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations (ICLR)*.

Kunstner, Frederik and Francis Bach (2025). "Scaling Laws for Gradient Descent and Sign Descent for Linear Bigram Models under Zipf's Law". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Kunstner, Frederik, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt (2023). "Noise Is Not the Main Factor Behind the Gap Between Sgd and Adam on Transformers, But Sign Descent Might Be". In: *International Conference on Learning Representations (ICLR)*.

Kunstner, Frederik, Robin Yadav, Alan Milligan, Mark Schmidt, and Alberto Bietti (2024). "Heavy-Tailed Class Imbalance and Why Adam Outperforms Gradient Descent on Language Models". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Lyu, Bochen, Di Wang, and Zhanxing Zhu (2025). "A Solvable Attention for Neural Scaling Laws". In: *International Conference on Learning Representations (ICLR)*.

Marek, Martin, Sanae Lotfi, Aditya Somasundaram, Andrew Gordon Wilson, and Micah Goldblum (2025). "Small Batch Size Training for Language Models: When Vanilla SGD Works, and Why Gradient Accumulation is Wasteful". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Meng, Si Yi, Antonio Orvieto, Daniel Yiming Cao, and Christopher De Sa (2024). *Gradient Descent on Logistic Regression with Non-Separable Data and Large Step Sizes*. Preprint. arXiv/2406.05033.

Paquette, Elliot, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington (2024). "4+3 Phases of Compute-Optimal Neural Scaling Laws". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Piantadosi, Steven T. (2014). "Zipf's word frequency law in natural language: A critical review and future directions". In: *Psychonomic Bulletin & Review* 21.5, pp. 1112–1130.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). "Neural Machine Translation of Rare Words with Subword Units". In: *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Soudry, Daniel, Elad Hoffer, Mor Spigel Nacson, and Nathan Srebro (2018). "The Implicit Bias of Gradient Descent on Separable Data". In: *International Conference on Learning Representations (ICLR)*.

Srećković, Teodora, Jonas Geiping, and Antonio Orvieto (2025). "Is your batch size the problem? Revisiting the Adam-SGD gap in language modeling". In: *ICML Workshop on High-dimensional Learning Dynamics (HILD)*.

Tao, Chaofan, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong (2024). "Scaling Laws with Vocabulary: Larger Models Deserve Larger Vocabularies". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Tsiolis, Konstantinos Christopher, Denny Wu, Christos Thrampoulidis, and Murat A. Erdogdu (2026a). *SGD in Multiclass Logistic Regression: Sequential Learning and Scaling Laws*.

Tsiolis, Konstantinos Christopher, Denny Wu, Christos Thrampoulidis, and Murat A. Erdogdu (2026b). "SGD in Multiclass Logistic Regression: Sequential Learning and Scaling Laws". In: *arXiv preprint arXiv:2609.07868*.

Varre, Aditya, Mark Rofin, and Nicolas Flammarion (2026). *Gradient Flow Polarizes Softmax Outputs towards Low-Entropy Solutions*. Preprint. arXiv/2603.06248.

Wu, Jingfeng, Peter L. Bartlett, Matus Telgarsky, and Bin Yu (2024). "Large Stepsize Gradient Descent for Logistic Loss: Non-Monotonicity of the Loss Improves Optimization Efficiency". In: *Proceedings of Thirty Seventh Conference on Learning Theory*. Vol. 247, pp. 5019–5073.

Xiao, Ke Liang, Noah Marshall, Atish Agarwala, and Elliot Paquette (2025). "Exact risk curves of signSGD in High-Dimensions: quantifying preconditioning and noise-compression effects". In: *Forty-second International Conference on Machine Learning*.

Yadav, Robin, Shuo Xie, Tianhao Wang, and Zhiyuan Li (2025). "Provable Benefit of Sign Descent: A Minimal Model Under Heavy-Tailed Class Imbalance". In: *NeurIPS Workshop on Optimization for Machine Learning (OPT)*.

Zhang, Jingzhao, Tianxing He, Suvrit Sra, and Ali Jadbabaie (2020). "Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity". In: *International Conference on Learning Representations (ICLR)*.

Zhang, Yushun, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo (2024). "Why Transformers Need Adam: A Hessian Perspective". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhang, Yushun, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik P. Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun (2025). "Adam-mini: Use Fewer Learning Rates To Gain More". In: *International Conference on Learning Representations (ICLR)*.

Zhao, Rosie, Depen Morwani, David Brandfonbrener, Nikhil Vyas, and Sham M. Kakade (2025). "Deconstructing What Makes a Good Optimizer for Autoregressive Language Models". In: *International Conference on Learning Representations (ICLR)*.

Appendix

Table of Contents

A Experimental details **13**

B Proofs of the set-up and definitions **14**

C Setting the Step-Size for Gradient Descent **14**

D Gradient flow is equivalent to gradient descent **15**
- D.1 Preliminary results . . . 16
- D.2 Restricting to small step-sizes . . . 17
- D.3 Bounding the difference between GF and GD . . . 19
- D.4 Behavior of the gradient flow dynamics . . . 20
- D.5 Exact bounds on the trajectory and loss . . . 22
- D.6 Proof of the Limit of GD . . . 24
- D.7 Sequential learning . . . 29

E Dynamics of Sign Descent **31**
- E.1 The oscillatory phase . . . 31
- E.2 Picking the step-size . . . 31
- E.3 Limit theorems . . . 31

A Experimental details

Synthetic data experiments run gradient descent on the expected loss, using the distribution $\pi_k \propto 1/k$ and $\pi_{k|j} \propto 1/(k)$ described in Assumption 2.2. The weights $\boldsymbol{W}$ are initialized at $\log(1/d)$, but any constant initialization would yield the same results.

Experiments using real data follow the protocol of Kunstner and Bach (2025) on the OpenWebText (Gokaslan et al., 2019) dataset, tokenized with BPE (Sennrich et al., 2016) and a vocabulary of size d. The token and bigram frequencies are then computed over the dataset, and used to define the empirical distribution π_k and $\pi_{k|j}$.

Optimizers and hyperparameters. We use plain gradient and sign descent. We run Adam in Figure 1 without momentum, $\beta_1 = 0$, and other parameters set to default values ($\beta_2 = 0.999$, $\epsilon = 10^{-8}$, no weight decay). The step-size used is either set by grid-search with a logarithmically spaced grid, ensuring we get within 20% of the optimum step-size. The grid-search is run separately for every dimension d for gradient descent, and for budget T and dimension d for sign descent.

- Figures 1, 5 and 7 use a grid search. Figures 2 and 3 use the theoretically-motivated step-size.
- Figure 6 uses the theoretical step-size for the left panel, and a grid search for the other panels.
- Figure 4 uses an arbitrary step-size of $\eta = 0.05$

Horizon vs. runtime. For gradient descent, the step-size is independent of the horizon and the loss goes down over time. Therefore there is no difference between looking at the loss of a single run at time T and looking at the best loss achieved over a run of T steps. For sign descent however, the step-size depends on the horizon T, and to show the scaling law we need to show the loss as a function of the horizon T. The results of Figures 1 and 6 therefore do not show the result of a single run over time, but the lower envelope of what is achieved by sign descent.

Oscillations. Even with an exact grid search, the performance of sign descent can depend a lot on the exact time at which it is stopped. We observe period-2 oscillations where the loss at even

steps is higher than the loss at odd steps. This effect is more pronounced for small T, with period-2 oscillations that are empirically of order $1/T$. To smooth this effect in Figure 1, we show the average of the loss over two steps for all algorithms.

B Proofs of the set-up and definitions

Proposition 2.3 (Closed-form Expected Loss and Gradient)**.** *The loss can be written as:*

$$\mathcal{L}(\boldsymbol{W}) = -\sum_{j=1}^{d}\sum_{k=1}^{d} \pi_j \pi_{k|j} \log\left(\mathrm{softmax}(\boldsymbol{W}_{:,j})_k\right), \text{ where } \mathrm{softmax}(\boldsymbol{W}_{:,j})_k = \frac{e^{\boldsymbol{W}_{k,j}}}{\sum_{l=1}^{d} e^{\boldsymbol{W}_{l,j}}}, \quad (1)$$

where $\boldsymbol{W}_{:,j}$ denotes the j-th column of $\boldsymbol{W}$. The gradient of $\mathcal{L}(\boldsymbol{W})$ with respect to $\boldsymbol{W}_{k,j}$ is

$$\frac{\partial \mathcal{L}(\boldsymbol{W})}{\partial \boldsymbol{W}_{k,j}} = \pi_j \left(\mathrm{softmax}(\boldsymbol{W}_{:,j})_k - \pi_{k|j}\right).$$

The solution is such that $\boldsymbol{W}^*$ satisfies $\mathrm{softmax}(\boldsymbol{W}^*_{:,j}) = \pi_{\cdot|j}$ for all j.

Proof of Proposition 2.3. If $\boldsymbol{x}$ is the one-hot encoding of k and $\boldsymbol{y}$ the one-hot encoding of j, then $\boldsymbol{y}^\top \boldsymbol{W} \boldsymbol{x} = W_{k,j}$, so the pointwise loss reduces to $\ell(\boldsymbol{W}; j, k) = -\log\left(\mathrm{softmax}(\boldsymbol{W}_{:,j})_k\right)$. Expanding the expectation over the joint distribution then gives the formula for the loss.

Differentiating the expression with respect to $W_{k,j}$ gives:

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{W})}{\partial W_{k,j}} &= -\sum_{k'=1}^{d} \pi_j \pi_{k'|j} \frac{\partial}{\partial W_{k,j}} \log \mathrm{softmax}(\boldsymbol{W}_{:,j})_{k'}, \\
&= -\sum_{k'=1}^{d} \pi_j \pi_{k'|j} \left(\mathbf{1}_{k'=k} - \mathrm{softmax}(\boldsymbol{W}_{:,j})_k\right), \\
&= \pi_j \left(\mathrm{softmax}(\boldsymbol{W}_{:,j})_k - \pi_{k|j}\right).
\end{aligned}$$

□

Claim B.1. *The initial error scales as $d \to \infty$: $\mathcal{L}(0) - \mathcal{L}^* \sim \frac{1}{2}\log(d)$.*

Proof. At initialization, the softmax is uniform, i.e. $\mathrm{softmax}(\boldsymbol{W}_{:,j})_k = 1/d$. Therefore,

$$\begin{aligned}
\mathcal{L}(0) - \mathcal{L}^* &= \sum_{j=1}^{d}\sum_{k=1}^{d} \pi_j \pi_{k|j} \log\left(\frac{\pi_{k|j}}{1/d}\right) = \log(d) + \sum_j \pi_j \sum_k \pi_{k|j} \log(\pi_{k|j}), \\
&= \log(d) - \log(H_d) \sum_{k=1}^{d} \frac{1}{k H_d} - \frac{1}{H_d} \sum_{k=1}^{d} \frac{\log(k)}{k}.
\end{aligned}$$

Using $H_d \sim \log(d)$ and $\sum_{k=1}^{d} \frac{\log(k)}{k} \sim \frac{1}{2}\log^2(d)$, we obtain $\mathcal{L}(0) - \mathcal{L}^* \sim \frac{1}{2}\log(d)$. □

C Setting the Step-Size for Gradient Descent

In this section, we show that for gradient descent to converge we need $\eta \leq 2\log(d)^2$. However, this does not prove that $\eta = \log(d)^2$ is optimal. For that, we use some empirical evidence that $\eta = H_d^2$ is optimal up to a constant (Figure 7). As explained in the main text, to prove the equivalence for the largest eigenvalue (Proposition 3.1), we examine the Hessian at the optimum.

Lemma C.1. *We analyze the per-word loss ℓ_j for a given context j. At the optimum $\boldsymbol{W}^*$, the Hessian of the loss is block diagonal with the j-th block being:*

$$B_j = \pi_j \left[\mathrm{Diag}(P_j) - P_j P_j^\top\right], \quad \text{where} \quad P_j = (\pi_{1|j}, \ldots, \pi_{d|j})^\top.$$

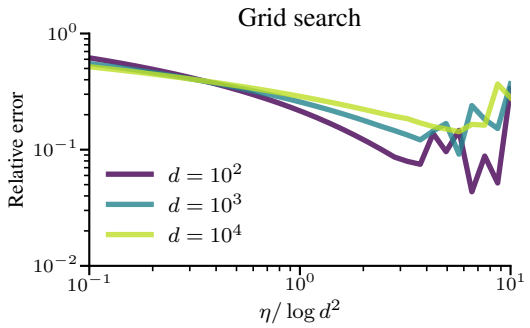

Figure 7: The best step-size is close, up to a constant, to the local-stability limit $\eta = \log^2(d)$. Final relative error $r(T)$ as a function of the scaled step-size $\eta/\log^2(d)$, evaluated at $T = d$.

Proof. Since the objective completely decouples across the columns of $\boldsymbol{W}$, the global Hessian is block diagonal. For a given context j, the per-word loss $\ell_j(\boldsymbol{W}_{:,j})$ is the standard cross-entropy loss with respect to the logits $\boldsymbol{W}_{:,j}$. Computing the Hessian of this loss directly yields:

$$\nabla^2 \ell_j(\boldsymbol{W}_{:,j}) = \operatorname{diag}(S_j) - S_j S_j^\top, \quad \text{where} \quad S_j = \operatorname{softmax}(\boldsymbol{W}_{:,j}).$$

At the global optimum $\boldsymbol{W}^*$, the predicted distribution perfectly matches the target conditional distribution, meaning $S_j = P_j$. Since the global loss is $\mathcal{L}(\boldsymbol{W}) = \sum_{j=1}^{d} \pi_j \ell_j(\boldsymbol{W}_{:,j})$. Multiplying the local Hessian by the context frequency π_j gives the j-th block of the global Hessian:

$$B_j = \pi_j \left[\operatorname{diag}(P_j) - P_j P_j^\top\right]. \qquad \square$$

With this result, we can prove the asymptotic equivalent of the largest eigenvalue.

Proposition 3.1. *The largest eigenvalue of the Hessian at the optimum is* $\lambda^*_{\max} \sim 1/H_d^2$.

Proof of Proposition 3.1. Since $P_j P_j^\top$ is a positive semi-definite matrix, we have the matrix inequality $B_j \preceq \pi_j \operatorname{diag}(P_j)$. Thus :

$$\lambda_{\max}(B_j) \leq \pi_j \max_k(\pi_{k|j}) = \pi_j \pi_{1|j}.$$

Taking the maximum over all blocks j, the largest eigenvalue $\lambda^*_{\max}$ satisfies:

$$\lambda^*_{\max} \leq \pi_1 \pi_{1|1} = \pi_1^2 = \frac{1}{H_d^2}.$$

We then use the fact that $\lambda_{\max}(B_1) \geq \boldsymbol{u}^\top B_1 \boldsymbol{u}$ for any unit vector $\boldsymbol{u}$. Taking $\boldsymbol{u} = \mathbf{e}_1$, we get:

$$\lambda^*_{\max} \geq \lambda_{\max}(B_1) \geq \mathbf{e}_1^\top B_1 \mathbf{e}_1 = \pi_1(\pi_{1|1} - \pi_{1|1}^2) = \pi_1^2(1 - \pi_1) = \frac{1}{H_d^2}\left(1 - \frac{1}{H_d}\right).$$

Combining the upper and lower bounds yields:

$$\frac{1}{H_d^2}\left(1 - \frac{1}{H_d}\right) \leq \lambda^*_{\max} \leq \frac{1}{H_d^2}.$$

This gives the result as both side scales as $\frac{1}{\log(d)^2}$. $\square$

D GRADIENT FLOW IS EQUIVALENT TO GRADIENT DESCENT

In this section, we prove Lemma 3.2.

Lemma 3.2 (Equivalence of Discrete and Continuous Dynamics)**.** *The difference in relative error between the discrete $\boldsymbol{W}(t)$ and continuous $\boldsymbol{W}^{\mathrm{F}}(t)$ started at a uniform initialization vanishes as $d \to \infty$; the following bound holds uniformly for all $t > 0$:*

$$|r(\boldsymbol{W}(t)) - r(\boldsymbol{W}^{\mathrm{F}}(t))| = O\left(\frac{\log(\log(d))}{\log(d)}\right). \tag{4}$$

This proof proceeds in four main steps. First, we establish preliminary bounds on the partition function (Appendix D.1). Second, we show that the most frequent context words do not affect the asymptotic relative error (Appendix D.2). Third, we bound the discrepancy between the two dynamics on the remaining words (Appendix D.3). Finally, we analyze the continuous dynamics to bound this total variation and conclude on the global convergence of the relative error (Appendices D.4 and D.5).

D.1 Preliminary results

Proposition 3.3 (Exact Trajectory of the Gradient Flow)**.** *Under the gradient flow equation (3) with uniform initialization $W_{k,j}(0) = -\log(d)$, the exact trajectory is given by:*

$$\exp(\boldsymbol{W}_{k,j}(t)) = \frac{\exp\big(t\eta\pi_j\pi_{k|j}\big)}{d + \eta\pi_j \int_0^t \frac{\exp\big(s\eta\pi_j\pi_{k|j}\big)}{Z_j(s)}\,\mathrm{d}s}, \quad \textit{where} \quad Z_j(s) = \sum_{l=1}^{d} \exp(\boldsymbol{W}_{l,j}(s)). \tag{5}$$

Proof of Proposition 3.3. We fix a column j and drop the index j for readability. We introduce $y_k(t) := \exp(-W_k(t))$. Differentiating and substituting the gradient flow equation 3 yields:

$$\dot{y}_k(t) = -y_k(t)\dot{W}_k(t) = \eta\, y_k(t)\left(\frac{e^{W_k(t)}}{Z(t)} - \pi_k\right) = \eta\left(\frac{1}{Z(t)} - \pi_k y_k(t)\right),$$

where we used $y_k(t)e^{W_k(t)} = 1$. Rearranging gives the linear first-order ODE:

$$\dot{y}_k(t) + \eta\pi_k y_k(t) = \frac{\eta}{Z(t)}.$$

We can then solve it as a classic ODE. Recalling $y_k(t) = \exp(-W_k(t))$, we obtain the result. □

Lemma D.1 (Conservation of the Sum of Weights)**.** *For both the continuous gradient flow and the discrete gradient descent, the sum of the weights in any column j is strictly conserved over time. Given the uniform initialization $\boldsymbol{W}_{k,j}(0) = -\log(d)$, we have:*

$$\sum_{k=1}^{d} \boldsymbol{W}_{k,j}(t) = -d\log(d) \qquad \forall t \geq 0.$$

Proof. We drop the column index j for readability.

For the continuous gradient flow, we sum the time derivatives of the weights over all k:

$$\frac{d}{dt}\sum_{k=1}^{d} \boldsymbol{W}_k(t) = \sum_{k=1}^{d} \dot{\boldsymbol{W}}_k(t) = -\eta\pi_j \sum_{k=1}^{d}\left(\frac{e^{\boldsymbol{W}_k(t)}}{Z(t)} - \pi_{k|j}\right) = -\eta\pi_j(1-1) = 0.$$

Thus, the sum is constant over time. For the discrete gradient descent, we sum the updates over all k:

$$\sum_{k=1}^{d} \boldsymbol{W}_k(t+1) - \sum_{k=1}^{d} \boldsymbol{W}_k(t) = -\eta\pi_j \sum_{k=1}^{d}\big(\mathrm{softmax}(\boldsymbol{W}(t))_k - \pi_{k|j}\big) = -\eta\pi_j(1-1) = 0.$$

Thus, the sum is strictly conserved at each iteration. Since both dynamics start from the uniform initialization $\boldsymbol{W}_k(0) = -\log(d)$, the sum at $t = 0$ is exactly $d \times (-\log(d)) = -d\log(d)$. This value is therefore conserved for all $t \geq 0$. □

Lemma D.2. *For both the continuous gradient flow and the discrete gradient descent, the partition function for any column j satisfies: $1 \leq Z_j(t) \leq H_d, \forall t \geq 0$.*

Proof. We drop the column index j for readability, denoting $\eta_{\text{eff}} = \eta_{j,\text{eff}} = \eta\pi_j$ and $\pi_k = \pi_{k|j}$.

Lower bound (Both cases): By Jensen's inequality and Lemma D.1:

$$Z(t) = \sum_{k=1}^{d} \exp(w_k(t)) = d \sum_{k=1}^{d} \frac{1}{d} \exp(w_k(t)) \geq d \exp\left(\frac{1}{d}\sum_{k=1}^{d} w_k(t)\right) = d\exp(-\log(d)) = 1.$$

Upper bound for the continuous case: Consider any time interval $[0, T]$. Since $Z(t)$ is continuous, it reaches its maximum at some time t^*. Let $Z_{\max} = Z(t^*)$. By Proposition 3.3 and bounding $Z(s) \leq Z_{\max}$ inside the integral, we obtain:

$$Z_{\max} \leq \sum_{k=1}^{d} \frac{e^{\eta_{\text{eff}}\pi_k t^*}}{d + \frac{\eta_{\text{eff}}}{Z_{\max}} \int_0^{t^*} e^{\eta_{\text{eff}}\pi_k s}\, \mathrm{d}s} = \sum_{k=1}^{d} \frac{e^{\eta_{\text{eff}}\pi_k t^*}}{d + \frac{1}{\pi_k Z_{\max}} \left(e^{\eta_{\text{eff}}\pi_k t^*} - 1\right)}.$$

Dividing both sides by $Z_{\max}$ and multiplying the numerator and denominator by π_k, we obtain:

$$1 \leq \sum_{k=1}^{d} \frac{\pi_k}{1 + e^{-\eta_{\text{eff}}\pi_k t^*}(dZ_{\max}\pi_k - 1)}. \tag{8}$$

Assume that $Z_{\max} > H_d$. The term in the denominators satisfies:

$$dZ_{\max}\pi_k - 1 > dH_d \left(\frac{1}{kH_d}\right) - 1 = \frac{d}{k} - 1 \geq 0 \quad (\text{since } k \leq d).$$

Because $e^{-\eta_{\text{eff}}\pi_k t^*}$ is strictly positive, the denominator of every fraction in the sum is greater than 1. This leads to: $1 \leq \sum_{k=1}^{d} \frac{\pi_k}{>1} < \sum_{k=1}^{d} \pi_k = 1$. We have the contradiction $1 < 1$. Thus, $Z(t) \leq H_d$ for all $t \geq 0$ in both dynamics.

Upper bound for the discrete case: Let $y_k(t) = \exp(-w_k^d(t))$. The discrete update rule $w_k(t+1) = w_k(t) - \eta_{\text{eff}}\big(\text{softmax}(\boldsymbol{w}^d(t))_k - \pi_k\big)$ combined with the inequality $e^x \geq 1 + x$ yields:

$$y_k(t+1) = y_k(t) \exp\Big(\eta_{\text{eff}}\big(\text{softmax}(\boldsymbol{w}^d(t))_k - \pi_k\big)\Big) \geq y_k(t)\Big(1 + \eta_{\text{eff}}\big(\text{softmax}(\boldsymbol{w}^d(t))_k - \pi_k\big)\Big).$$

Using $S_k(t) = \frac{1}{y_k(t)Z(t)}$, this becomes $y_k(t+1) \geq y_k(t)(1 - \eta_{\text{eff}}\pi_k) + \frac{\eta_{\text{eff}}}{Z(t)}$. Unrolling this inequality from 0 to $t^* = \arg\max_{t\in[0,T]} Z(t)$ ($1 - \eta_{\text{eff}}\pi_k = 1 - 1/(kj) \geq 0$), and bounding $Z(s) \leq Z_{\max}$ inside the sum gives:

$$y_k(t^*) \geq d(1 - \eta_{\text{eff}}\pi_k)^{t^*} + \frac{\eta_{\text{eff}}}{Z_{\max}} \sum_{s=0}^{t^*-1} (1 - \eta_{\text{eff}}\pi_k)^s = d(1 - \eta_{\text{eff}}\pi_k)^{t^*} + \frac{1 - (1 - \eta_{\text{eff}}\pi_k)^{t^*}}{Z_{\max}\pi_k}.$$

Injecting this lower bound back into the definition $Z_{\max} = \sum_{k=1}^{d} \frac{1}{y_k(t^*)}$ and rearranging the terms exactly as we did for the continuous case, we get:

$$1 \leq \sum_{k=1}^{d} \frac{\pi_k}{1 + (1 - \eta_{\text{eff}}\pi_k)^{t^*}(dZ_{\max}\pi_k - 1)}. \tag{9}$$

We conclude with the same technique as the continuous case. □

D.2 Restricting to small step-sizes

A challenge in showing that gradient flow and gradient descent match is that the step-size $\eta = H_d^2 \sim \log^2(d)$ diverges as $d \to \infty$. However, the dynamics of each column j are governed by an effective step-size $\eta_j = H_d/j$, which remains small for all but a fraction of contexts. Our next step is therefore to isolate the first, most frequent words, where η_j is large and show that their contribution to the relative error is negligible. We formalize this in the following theorem:

Theorem D.3. *The contribution of the first $K = \lfloor \log^2(d) \rfloor$ context words to the relative error vanishes as $d \to \infty$:*

$$\frac{1}{\mathcal{L}(W(0)) - \mathcal{L}(W^*)} \sum_{j=1}^{K} \pi_j \left| \ell_j(\boldsymbol{w}_j^c(t)) - \ell_j(\boldsymbol{w}_j^d(t)) \right| \xrightarrow[d\to\infty]{} 0.$$

Proof. We first show that we can bound the per-column loss ℓ_j, for all j:

$$\ell_j(\boldsymbol{w}_j^c(t)) \le \log(d), \qquad \ell_j(\boldsymbol{w}_j^d(t)) = O(\log(d)) \qquad \forall t \ge 0.$$

For the continuous case. By the chain rule:

$$\frac{d}{dt}\ell_j(\boldsymbol{w}_j^c(t)) = \langle \nabla \ell_j(\boldsymbol{w}_j^c), \dot{\boldsymbol{w}}_j^c \rangle = -\eta_j \|\nabla \ell_j(\boldsymbol{w}_j^c)\|_2^2 \le 0.$$

Since the loss is non-increasing and $\ell_j(\boldsymbol{w}_j^c(0)) = \log(d)$, hence the result.

For the discrete case. To bound the loss $\ell_j(\boldsymbol{w}_j^d(t)) = -\sum_k \pi_{k|j} w_k^d(t) + \log Z_j^d(t)$, we first establish a lower bound on the weights $w_k^d(t)$. We drop the column index j for readability, setting $\eta_{\text{eff}} = \eta\pi_j$ and $\pi_k = \pi_{k|j}$.

The discrete update rule is $w_k(t+1) = w_k(t) - \eta_{\text{eff}}\big(\text{softmax}(\boldsymbol{w}^d(t))_k - \pi_k\big)$ so the weight w_k decreases if and only if $\text{softmax}(\boldsymbol{w}^d(t))_k > \pi_k$. Suppose w_k strictly decreases at step t^*. Its value after the update is:

$$w_k(t^*+1) = \log\big(\text{softmax}(\boldsymbol{w}^d(t^*))_k Z(t^*)\big) - \eta_{\text{eff}}\big(\text{softmax}(\boldsymbol{w}^d(t^*))_k - \pi_k\big)$$

Consider the function $f(x) = \log\big(x \cdot Z(t^*)\big) - \eta_{\text{eff}}(x - \pi_k)$ defined on the interval $x \in [\pi_k, 1]$. Its second derivative is $f''(x) = -1/x^2 < 0$. Because f is strictly concave, its minimum on this interval necessarily occurs at one of the boundaries. And because $Z \ge 1$ (Lemma D.2):

$$\begin{aligned} \text{At } x = \pi_k: &\quad f(\pi_k) = \log\big(\pi_k Z(t^*)\big) \ge \log \pi_k, \\ \text{At } x = 1: &\quad f(1) = \log Z(t^*) - \eta_{\text{eff}}(1 - \pi_k) \ge -\eta_{\text{eff}}. \end{aligned}$$

Therefore, whenever a weight decreases, its new value is bounded below by $\min(\log \pi_k, -\eta_{\text{eff}})$. This guarantees that any value reached via a decrease is bounded by:

$$w_k(t^*+1) \ge \log \pi_k - \eta_{\text{eff}} \ge -\log(d) - \log H_d - \eta_{\text{eff}}.$$

Since the initial value is $w_k(0) = -\log(d)$, it satisfies this lower bound. Thus, at any time t and for all $k \in \{1, \dots, d\}$, we have:

$$w_k(t) \ge -\log(d) - \log H_d - \eta_{\text{eff}}.$$

Injecting this bound into the loss formulation and remembering that $Z \le H_d$ (Lemma D.2), yields:

$$\begin{aligned} \ell_j(\boldsymbol{w}_j^d(t)) &= -\sum_k \pi_k w_k(t) + \log Z(t), \\ &\le -\sum_k \pi_k(-\log(d) - \log H_d - \eta_{\text{eff}}) + \log Z(t), \\ &= \log(d) + \log H_d + \eta_{\text{eff}} + \log Z(t) = O(\log(d)). \end{aligned}$$

Using the bounds established, we can now bound the absolute difference for any context j:

$$|\ell_j(\boldsymbol{w}_j^c(t)) - \ell_j(\boldsymbol{w}_j^d(t))| \le \ell_j(\boldsymbol{w}_j^c(t)) + \ell_j(\boldsymbol{w}_j^d(t)) \le \log(d) + O(\log(d)) = O(\log(d)).$$

Next, we sum this bound over the first $K = \lfloor \log(d)^2 \rfloor$ context words:

$$\sum_{j=1}^{K} \pi_j \left| \ell_j(\boldsymbol{w}_j^c(t)) - \ell_j(\boldsymbol{w}_j^d(t)) \right| \le O(\log(d)) \sum_{j=1}^{K} \pi_j.$$

The cumulative probability mass of the head is given by:

$$\sum_{j=1}^{K} \pi_j = \frac{H_K}{H_d} \sim \frac{\log(\log^2(d))}{\log(d)} = \frac{2\log(\log(d))}{\log(d)}.$$

Therefore, this sum is bounded by:

$$\sum_{j=1}^{K} \pi_j \left|\ell_j(\boldsymbol{w}_j^c(t)) - \ell_j(\boldsymbol{w}_j^d(t))\right| \leq O(\log(d)) \times \left(\frac{2\log(\log(d))}{H_d}\right) = O(\log(\log(d))).$$

Normalizing by the initial loss $\mathcal{L}(W(0)) - \mathcal{L}(W^*) \sim \frac{1}{2}\log(d)$, we obtain:

$$\frac{1}{\mathcal{L}(W(0)) - \mathcal{L}(W^*)} \sum_{j=1}^{K} \pi_j \left|\ell_j(\boldsymbol{w}_j^c(t)) - \ell_j(\boldsymbol{w}_j^d(t))\right| = O\left(\frac{\log(\log(d))}{\log(d)}\right) \xrightarrow[d\to\infty]{} 0. \qquad \square$$

D.3 Bounding the difference between GF and GD

We can now focus on j where $j > K$, meaning $\eta_{\text{eff},j} = \eta\pi_j = H_d/j \leq 1/H_d \leq 1$. We omit the index j for readability.

Definition D.4. *We define the following operators on the vector $\boldsymbol{w}$:*

- ***Discrete operator:*** *$D(\boldsymbol{w}) := \boldsymbol{w} - \eta\nabla\ell(\boldsymbol{w}) = \boldsymbol{w} - \eta(S(\boldsymbol{w}) - \pi)$, so that $\boldsymbol{w}^d(t+1) = D(\boldsymbol{w}^d(t))$.*
- ***Continuous operator:*** *$C(\boldsymbol{w}) := \boldsymbol{w}(1)$, where $\boldsymbol{w}(\cdot)$ solves $\dot{\boldsymbol{w}}(s) = -\eta\nabla\ell(\boldsymbol{w}(s))$, $\boldsymbol{w}(0) = \boldsymbol{w}$; so that $\boldsymbol{w}^c(t+1) = C(\boldsymbol{w}^c(t))$ for integer t.*

To bound the difference between the two paths, we first show a property of the D operator.

Lemma D.5 (L_1 Non-expansiveness of D)**.** *The operator D is non-expansive in the L_1 norm :*

$$\|D(\boldsymbol{w}_1) - D(\boldsymbol{w}_2)\|_1 \leq \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_1 \qquad \forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d.$$

Proof. By the fundamental theorem of calculus applied to $u \mapsto D(\boldsymbol{w}_2 + u(\boldsymbol{w}_1 - \boldsymbol{w}_2))$,

$$D(\boldsymbol{w}_1) - D(\boldsymbol{w}_2) = \left(\int_0^1 J_D(\boldsymbol{w}_u)\ \mathrm{d}u\right)(\boldsymbol{w}_1 - \boldsymbol{w}_2),$$

where $J_D(\boldsymbol{w}) = I - \eta\nabla^2\ell(\boldsymbol{w})$ and $\nabla^2\ell(\boldsymbol{w}) = \mathrm{diag}(S(\boldsymbol{w})) - S(\boldsymbol{w})S(\boldsymbol{w})^\top$. We have $J_{ij} = \eta S_i S_j$ for $i \neq j$ and $J_{ii} = 1 - \eta S_i(1 - S_i)$ on the diagonal. The maximum possible value of $S_i(1 - S_i)$ is $1/4$. Since $\eta \leq 1$ (because η is the effective step-size and $j > K$), the diagonal elements are non-negative: $1 - \eta S_i(1 - S_i) \geq 0$.

The off-diagonal elements are positive. The L_1 operator norm of a matrix is its maximum absolute column sum ($\|A\|_1 = \max_j \sum_i |A_{ij}|$). Thus, for a fixed column j:

$$\sum_{i=1}^{d} |J_{ij}| = J_{jj} + \sum_{i\neq j} J_{ij} = 1 - \eta S_j(1 - S_j) + \eta S_j \sum_{i\neq j} S_i = 1,$$

because probabilities sum to 1. Therefore, $\|J_D(\boldsymbol{w}_u)\|_1 \leq 1$ for all u and the operator is non-expansive. $\square$

We now use this result to bound the difference between the two paths. More precisely:

Theorem D.6. *Let $\boldsymbol{S}^c(s) = \mathrm{softmax}(\boldsymbol{w}^c(s))$ denote the softmax probabilities along the continuous trajectory. For every integer $t \geq 1$, we have:*

$$\|\boldsymbol{w}^c(t) - \boldsymbol{w}^d(t)\|_1 \leq \eta_{\text{eff}} \int_0^t \|\dot{\boldsymbol{S}}^c(s)\|_1\ \mathrm{d}s.$$

Proof. Let $\boldsymbol{e}_k := \boldsymbol{w}^c(k) - \boldsymbol{w}^d(k)$ with $\boldsymbol{e}_0 = \mathbf{0}$. We can decompose the error at step $k+1$ as follows:

$$\boldsymbol{e}_{k+1} = C(\boldsymbol{w}^c(k)) - D(\boldsymbol{w}^d(k)) = \big[C(\boldsymbol{w}^c(k)) - D(\boldsymbol{w}^c(k))\big] + \big[D(\boldsymbol{w}^c(k)) - D(\boldsymbol{w}^d(k))\big].$$

Taking L_1 norms and using the non-expansiveness of D (Lemma D.5) on the second term gives:

$$\|\boldsymbol{e}_{k+1}\|_1 \le \|C(\boldsymbol{w}^c(k)) - D(\boldsymbol{w}^c(k))\|_1 + \|\boldsymbol{e}_k\|_1.$$

We apply Taylor's theorem to the continuous flow $f(u) := \boldsymbol{w}^c(k+u)$, which gives:

$$C(\boldsymbol{w}^c(k)) - D(\boldsymbol{w}^c(k)) = \eta^2 \int_0^1 (1-u)\, \nabla^2 \ell(\boldsymbol{w}^c(k+u))\, \nabla \ell(\boldsymbol{w}^c(k+u))\, \mathrm{d}u.$$

The time derivative of the softmax is $\dot{S}^c(s) = \nabla^2 \ell(\boldsymbol{w}^c(s))\, \dot{\boldsymbol{w}}^c(s) = -\eta\, \nabla^2 \ell(\boldsymbol{w}^c(s))\, \nabla \ell(\boldsymbol{w}^c(s))$. Substituting this into our Taylor expansion simplifies the expression to:

$$C(\boldsymbol{w}^c(k)) - D(\boldsymbol{w}^c(k)) = -\eta \int_0^1 (1-u)\, \dot{S}^c(k+u)\, \mathrm{d}u.$$

Since $1 - u \le 1$ on $[0,1]$, taking the L_1 norm bounds the local error by:

$$\|C(\boldsymbol{w}^c(k)) - D(\boldsymbol{w}^c(k))\|_1 \le \eta \int_0^1 \|\dot{S}^c(k+u)\|_1\, \mathrm{d}u.$$

Telescoping the recurrence $\|\boldsymbol{e}_{k+1}\|_1 \le \eta \int_0^1 \|\dot{S}^c(k+u)\|_1\, \mathrm{d}u + \|\boldsymbol{e}_k\|_1$ from $k = 0$ to $t-1$, substituting $s = k+u$, and concatenating the intervals into a single integral over $[0,t]$ concludes the proof. □

D.4 Behavior of the gradient flow dynamics

Bounding the discrepancy between the discrete and continuous paths requires controlling the variation of the softmax probabilities in the continuous setting. We therefore study the continuous dynamics in more detail by introducing the following quantities:

Definition D.7. *We define the normalized probabilities $x_i(t)$ as the ratio between the continuous softmax prediction $\boldsymbol{S}_i^c(t)$ and the target probability π_i: $x_i(t) := \frac{\boldsymbol{S}_i^c(t)}{\pi_i}$. Based on this ratio, we partition the indices into two sets: the head $\mathcal{H}(t)$ and the tail $\mathcal{T}(t)$, defined as:*

$$\mathcal{H}(t) = \{i \mid x_i(t) \le 1\} \quad \textit{and} \quad \mathcal{T}(t) = \{i \mid x_i(t) > 1\}.$$

This dynamic partition is crucial because it governs the entire trajectory of the optimization. Specifically, we will demonstrate that any word initialized below its target probability (i.e., starting in the head) permanently remains below it, making the head an absorbing state. Conversely, probabilities initialized above their targets (in the tail) strictly decrease. Some eventually cross their target and enter the head, where they remain trapped. In terms of the weights, this induces a simple behavior: weights in the head monotonically increase whereas those in the tail strictly decrease until they reach a minimum after which they monotonically increase.

We first show that the predicted probabilities stay strictly ordered at all times.

Lemma D.8. *For the continuous gradient flow, $S_1(t) > S_2(t) > \cdots > S_d(t)$ for all $t > 0$.*

Proof. Fix $i < k$ and set $\delta(t) := w_i(t) - w_k(t)$, so $\dot{\delta}(t) = \eta(\pi_i - \pi_k) - \eta(S_i(t) - S_k(t))$. At $t = 0$, $\delta(0) = 0$ and $\dot{\delta}(0) = \eta(\pi_i - \pi_k) > 0$, so $\delta(t) > 0$ for small $t > 0$. If δ returned to 0 for the first time at some $t_0 > 0$, then $w_i(t_0) = w_k(t_0)$, hence $S_i(t_0) = S_k(t_0)$ and $\dot{\delta}(t_0) = \eta(\pi_i - \pi_k) > 0$ — but this contradicts δ decreasing toward 0 from above. Hence $\delta(t) > 0$ for all $t > 0$, i.e., $S_1(t) > \cdots > S_d(t)$. □

Next, we show that two tail elements can never cross each other. Since $x_i(0) = \frac{1/d}{\pi_i}$ is strictly increasing in i, this means the elements that are present from the beginning stay ordered by increasing index for as long as they remain in the tail.

Lemma D.9. *For any $i < k$ with $1 < x_i(0) < x_k(0)$, the order $x_i(t) < x_k(t)$ holds for as long as both $x_i(t), x_k(t) > 1$. In particular, a higher-index tail element cannot reach the boundary $x = 1$ before a lower-index one.*

Proof. Set $\Delta(t) := \log x_k(t) - \log x_i(t)$. Since $\frac{d}{dt} \log x_i = \eta\pi_i(1 - x_i) - \dot{Z}^c/Z^c$, the partition-function term cancels in the difference:

$$\dot{\Delta}(t) = \eta\pi_k(1 - x_k(t)) - \eta\pi_i(1 - x_i(t)).$$

If Δ first vanished at some t_0 with $x_i(t_0) = x_k(t_0) = x > 1$, then $\dot{\Delta}(t_0) = \eta(1 - x)(\pi_k - \pi_i) > 0$, contradicting Δ decreasing to 0. So $\Delta(t) > 0$, i.e., $x_i(t) < x_k(t)$, as long as both exceed 1. □

With these two lemmas, we can show that Z^c grows monotonically.

Lemma D.10 (Monotonicity of the Partition Function)**.** *For gradient flow, $\dot{Z}^c(t) \geq 0$ for all $t \geq 0$.*

Proof. Write $\dot{Z}^c(t) = \eta Z^c(t) A(t)$ with $A(t) := \sum_{i=1}^d S_i(t)(\pi_i - S_i(t))$, and let

$$T^* := \sup\{T \geq 0 \mid \dot{Z}^c(t) \geq 0 \; \forall t \in [0, T]\}.$$

At $t = 0$, uniform initialization gives $A(0) = \sum_{i=1}^d \frac{1}{d}(\pi_i - \frac{1}{d}) = 0$, so $\dot{Z}^c(0) = 0$. Since under continuous flow $\dot{S}_i(0) = \frac{\eta}{d}(\pi_i - \frac{1}{d})$,

$$\dot{A}(0) = \sum_{i=1}^d \dot{S}_i(0)(\pi_i - 2S_i(0)) = \frac{\eta}{d} \sum_{i=1}^d \Big(\pi_i - \frac{1}{d}\Big)\Big(\pi_i - \frac{2}{d}\Big).$$

Using $\sum_i (\pi_i - \frac{1}{d}) = 0$ to expand the product, the cross term vanishes and this reduces to

$$\dot{A}(0) = \frac{\eta}{d} \sum_{i=1}^d \Big(\pi_i - \frac{1}{d}\Big)^2 > 0.$$

Since A is C^1, $A(0) = 0$ and $\dot{A}(0) > 0$ imply $A(t) > 0$ just after $t = 0$, so $T^* > 0$.

Suppose for contradiction that $T^* < \infty$. We first show that, on $[0, T^*]$, the head and tail are two contiguous blocks of indices, i.e., $H(t) = \{1, \ldots, m(t)\}$ for some $m(t)$. By Lemma D.9, tail elements leave the tail in index order, so the head only ever gains its newest index at the bottom of the tail. It remains to check that the head never loses an index.

On $[0, T^*]$, continuity gives $\dot{Z}^c(t) \geq 0$, so

$$\dot{x}_i(t) = x_i(t)\Big[\eta\pi_i(1 - x_i(t)) - \frac{\dot{Z}^c(t)}{Z^c(t)}\Big] \leq \eta\pi_i x_i(t)(1 - x_i(t)).$$

With $y_i(t) := x_i(t) - 1$, this reads $\dot{y}_i(t) + \eta\pi_i x_i(t) y_i(t) \leq 0$. Multiplying by the integrating factor $\mu(t) := \exp\big(\int_0^t \eta\pi_i x_i(s)\, \mathrm{d}s\big) > 0$ gives

$$\frac{d}{dt}\big[y_i(t)\mu(t)\big] \leq 0.$$

So if $y_i(t_0) \leq 0$ for some t_0 (index i has entered the head), then $y_i(t)\mu(t) \leq y_i(t_0)\mu(t_0) \leq 0$, hence $x_i(t) \leq 1$, for all $t \in [t_0, T^*]$. The head is thus absorbing and the head and tail are contiguous blocks,

$$A(T^*) = \sum_{i=1}^d \big(S_i(T^*) - c\big)\big(\pi_i - S_i(T^*)\big).$$

We pick $c \in \big(S_{m(T^*)+1}(T^*), S_{m(T^*)}(T^*)\big)$. Then:

- Head terms ($i \leq m(T^*)$): $S_i(T^*) \geq c$ and $\pi_i \geq S_i(T^*)$ (since $i \in H$), so the product is ≥ 0.
- Tail terms ($i > m(T^*)$): $S_i(T^*) < c$ and $\pi_i < S_i(T^*)$ (since $i \in T$), so the product is again ≥ 0.

Every term is non-negative. Since $S(T^*) \neq \pi$, at least one $S_i(T^*) \neq \pi_i$ and hence one term is strictly positive. Thus $A(T^*) > 0$ and $\dot{Z}^c(T^*) > 0$. By continuity of $\dot{Z}^c$, there is $\varepsilon > 0$ with $\dot{Z}^c(t) \geq 0$ on $[0, T^* + \varepsilon]$, contradicting the definition of T^*. So $T^* = \infty$, i.e., $\dot{Z}^c(t) \geq 0$ for all $t \geq 0$. □

Now that the global monotonicity of $Z^c(t)$ is established, we can fully characterize the trajectory of the normalized ratios $x_i(t)$. We begin by formalizing a crucial consequence: the absorbing property of the head, which was derived locally in the previous proof, holds globally.

Lemma D.11 (Positive invariance of the Head)**.** *If an element is in the head at some time t_0 (i.e., $x_i(t_0) \leq 1$), it remains in the head for all $t \geq t_0$ along the continuous gradient flow.*

Proof. Since $S_i = e^{w_i}/Z^c$, the time derivative is $\dot{S}_i = S_i\big(\dot{w}_i - \dot{Z}^c/Z^c\big)$. Dividing by π_i and using the flow definition $\dot{w}_i = -\eta\pi_i(x_i - 1)$, we obtain the exact dynamics of the ratio:

$$\dot{x}_i(t) = x_i(t)\left[\eta\pi_i\big(1 - x_i(t)\big) - \frac{\dot{Z}^c(t)}{Z^c(t)}\right]. \tag{10}$$

By Lemma D.10, we now know that $\dot{Z}^c(t)/Z^c(t) \geq 0$ for all $t \geq 0$. Dropping this non-negative contribution yields the differential inequality:

$$\dot{x}_i(t) \leq \eta\pi_i\, x_i(t)\big(1 - x_i(t)\big).$$

As previously demonstrated in Step 2 of the proof of Lemma D.10, applying the strictly positive integrating factor $\mu(t) = \exp(\int_{t_0}^{t} \eta\pi_i x_i(s)\, \mathrm{d}s)$ to the shifted variable $y(t) = x_i(t) - 1$ guarantees that $y(t_0) \leq 0 \implies y(t) \leq 0$ for all $t \geq t_0$. Thus, $x_i(t) \leq 1$ globally. □

Conversely, while an element remains in the tail ($x_i > 1$), the monotonicity of $Z^c(t)$ forces its ratio to strictly decrease toward the head.

Lemma D.12 (Strict decrease outside the invariant region)**.** *If $x_i(t) > 1$, then $S_i(t)$ (and equivalently $x_i(t)$) strictly decreases at time t along the continuous gradient flow, meaning $\dot{x}_i(t) < 0$.*

Proof. Recall the exact dynamics from Equation equation 10. When $x_i(t) > 1$, the first bracketed term, $\eta\pi_i\big(1 - x_i(t)\big)$, is strictly negative. Since $\dot{Z}^c(t)/Z^c(t) \geq 0$ (Lemma D.10) and the ratio $x_i(t)$ is strictly positive, the entire expression in the brackets is strictly negative. Consequently, $\dot{x}_i(t) < 0$. □

Now we have all the keys to understand the dynamics. At initialization, the indices are separated into those above their target and those below it. Those below always remain below, whereas those above decrease. Some decrease enough to cross the target and remain below for the rest of the trajectory, while others decrease but stay above their target. Since $\dot{w}_i(t) = \eta\pi_i(1 - x_i(t))$, this dynamic directly implies that $w_i(t)$ increases when $x_i(t) \leq 1$ and decreases when $x_i(t) > 1$. Therefore, weights in the head only increase, while weights starting in the tail decrease to a minimum before increasing if they enter the head.

D.5 Exact bounds on the trajectory and loss

Now that we understand the dynamics, we establish the following bounds. First:

Theorem D.13. *For all $t \geq 0$ and all coordinates i, the continuous gradient flow satisfies:*

$$\mathrm{softmax}(\boldsymbol{w}^c(t))_i \leq \max\left(\frac{1}{d}, \pi_i\right).$$

Proof. Fix i. **Case** $\pi_i \geq 1/d$**.** Then $x_i(0) = \frac{S_i(0)}{\pi_i} = \frac{1/d}{\pi_i} \leq 1$. By Lemma D.11, $x_i(t) \leq 1$ for all $t \geq 0$, i.e. $S_i(t) \leq \pi_i = \max(1/d, \pi_i)$.

Case $\pi_i < 1/d$**.** Then $x_i(0) > 1$ and $\max(1/d, \pi_i) = 1/d$. Two sub-cases:

- If $x_i(t) > 1$ for all $t \geq 0$: by Lemma D.12, S_i is strictly decreasing at every t, so $S_i(t) \leq S_i(0) = 1/d$ for all t.

- Otherwise, let $t_1 := \inf\{t \geq 0 : x_i(t) \leq 1\}$ (finite, by continuity). For $t < t_1$, $x_i(t) > 1$, so by Lemma D.12, S_i is decreasing on $[0, t_1]$, giving $S_i(t) \leq S_i(0) = 1/d$ for $t \leq t_1$. For $t \geq t_1$, $x_i(t_1) \leq 1$, so by Lemma D.11, $x_i(t) \leq 1$ for all $t \geq t_1$, i.e. $S_i(t) \leq \pi_i < 1/d$.

In both sub-cases, $S_i(t) \leq 1/d = \max(1/d, \pi_i)$ for all $t \geq 0$, which concludes the proof. □

Theorem D.14 (Total Variation of the softmax)**.** *For the continuous gradient flow, the total variation is bounded as:*

$$\int_0^\infty \|\dot{S}^c(t)\|_1 \, \mathrm{d}t = O(\log(\log(d)))$$

Proof. From $\dot{S}_i = S_i(\dot{w}_i - \dot{Z}^c/Z^c)$ and $\dot{Z}^c \geq 0$, the triangle inequality gives $|\dot{S}_i| \leq S_i|\dot{w}_i| + S_i \frac{\dot{Z}^c}{Z^c}$. Summing over i and using $\sum_i S_i = 1$ yields:

$$\|\dot{S}^c(t)\|_1 \leq \sum_{i=1}^d S_i(t)|\dot{w}_i(t)| + \frac{\dot{Z}^c(t)}{Z^c(t)}$$

Integrating over $[0, \infty)$ and using $Z^c(0) = 1$:

$$\int_0^\infty \|\dot{S}^c(t)\|_1 \, \mathrm{d}t \leq \sum_{i=1}^d \int_0^\infty S_i(t)|\dot{w}_i(t)| \, \mathrm{d}t + \log Z^c(\infty)$$

We bound the integral sum by splitting it between the initial head $\mathcal{H}(0)$ and tail $\mathcal{T}(0)$.

Initial Head classes ($i \in \mathcal{H}(0)$): Since $x_i(t) \leq 1$, $w_i(t)$ is monotonically increasing, so $|\dot{w}_i(t)| = \dot{w}_i(t)$. Substituting $\dot{w}_i = \frac{\dot{S}_i}{S_i} + \frac{\dot{Z}^c}{Z^c}$:

$$\sum_{i \in \mathcal{H}(0)} \int_0^\infty S_i(t)\dot{w}_i(t) \, \mathrm{d}t = \sum_{i \in \mathcal{H}(0)} \big(S_i(\infty) - S_i(0)\big) + \int_0^\infty \Big(\sum_{i \in \mathcal{H}(0)} S_i(t)\Big) \frac{\dot{Z}^c(t)}{Z^c(t)} \, \mathrm{d}t$$

Since $\sum_{i \in \mathcal{H}(0)} S_i(t) \leq 1$ at all times, this contribution is bounded by $1 + \log Z^c(\infty) = O(\log(\log(d)))$.

Initial Tail classes ($i \in \mathcal{T}(0)$): By Theorem D.13, $S_i(t) \leq 1/d$. Therefore:

$$\sum_{i \in \mathcal{T}(0)} \int_0^\infty S_i(t)|\dot{w}_i(t)| \, \mathrm{d}t \leq \frac{1}{d} \sum_{i \in \mathcal{T}(0)} TV(w_i)$$

where $TV(w_i) = \int_0^\infty |\dot{w}_i(t)| \, \mathrm{d}t$. For these classes, w_i decreases to a minimum where $\dot{w}_i = 0$ (i.e., $S_i = \pi_i$), then increases. Since $Z^c(t) \geq 1$, this minimum satisfies $w_i^{\min} = \log(S_i Z^c) \geq \log \pi_i$. The total variation is $TV(w_i) \leq w_i(0) - w_i^{\min} + w_i(\infty) - w_i^{\min} \leq w_i(0) + w_i(\infty) - 2\log \pi_i$. Using $w_i(0) = -\log(d)$ and $w_i(\infty) \leq \log(\pi_i Z^c(\infty))$, we obtain $TV(w_i) \leq \log Z^c(\infty) - \log(d\pi_i)$. Summing over $\mathcal{T}(0)$:

$$\frac{1}{d} \sum_{i \in \mathcal{T}(0)} TV(w_i) \leq \frac{|\mathcal{T}(0)|}{d} \log Z^c(\infty) + \frac{1}{d} \sum_{i \in \mathcal{T}(0)} \log\left(\frac{1}{d\pi_i}\right)$$

By Zipf's law, $\frac{1}{d\pi_i} = \frac{iH_d}{d} \leq H_d$. The logarithmic sum is bounded by $\log H_d$. Since $|\mathcal{T}(0)| < d$, this contribution is bounded by $\log Z^c(\infty) + \log H_d = O(\log(\log(d)))$. Finally:

$$\int_0^\infty \|\dot{S}^c(t)\|_1 \, \mathrm{d}t \leq O(\log(\log(d))).$$

□

Theorem D.15 (Absolute Loss Bound)**.** *The absolute difference in cross-entropy loss between discrete gradient descent and continuous gradient flow is bounded by:*

$$|\ell(\boldsymbol{w}^d(t)) - \ell(\boldsymbol{w}^c(t))| \leq \frac{O(\log(\log(d)))}{\log(d)}$$

Proof. Recall that Theorem D.6 yields:

$$\|\boldsymbol{w}^d(t) - \boldsymbol{w}^c(t)\|_1 \leq \eta \int_0^\infty \|\dot{S}^c(s)\|_1 \, \mathrm{d}s$$

To bound the difference in loss, we apply the Mean Value Theorem using the dual pairing of L_1 and L_∞ norms. For some intermediate point $\boldsymbol{w}_u$:

$$|\ell(\boldsymbol{w}^d(t)) - \ell(\boldsymbol{w}^c(t))| \leq \|\nabla \ell(\boldsymbol{w}_u)\|_\infty \|\boldsymbol{w}^d(t) - \boldsymbol{w}^c(t)\|_1$$

The gradient of the cross-entropy loss is $\nabla_i \ell(\boldsymbol{w}) = S_i(\boldsymbol{w}) - \pi_i$. Since both S_i and π_i are valid probabilities in $[0, 1]$, its L_∞ norm is bounded in $\mathbb{R}^d$:

$$\|\nabla \ell(\boldsymbol{w}_u)\|_\infty = \max_i |S_i(\boldsymbol{w}_u) - \pi_i| \leq 1$$

Injecting the L_1 parameter bound and using $\eta \leq \frac{1}{\log(d)}$ gives (recall that η is here the effective step-size which is small since $j > K$):

$$|\ell(\boldsymbol{w}^d(t)) - \ell(\boldsymbol{w}^c(t))| \leq 1 \times \left(\frac{1}{\log(d)} \int_0^\infty \|\dot{S}^c(s)\|_1 \, \mathrm{d}s \right)$$

By Theorem D.14, the total variation is bounded by $O(\log(\log(d)))$. Thus, the absolute difference in loss is bounded by $O\left(\log(\log(d))/\log(d)\right)$. □

Corollary D.16 (Convergence of the Relative Error)**.** *For any time t, the difference in relative error for the full model vanishes as the vocabulary size d grows.*

Proof. We combine the contributions of $j \leq K$ and $j > K$ to the relative error. As demonstrated earlier, the normalized error for $j \leq K$ is bounded by:

$$\frac{\sum_{j=1}^K \pi_j |\ell_j(\boldsymbol{w}_j^c(t)) - \ell_j(\boldsymbol{w}_j^d(t))|}{\mathcal{L}(W(0)) - \mathcal{L}(W^*)} = O\left(\frac{\log(\log(d))}{\log(d)} \right).$$

For $j > K$, Theorem D.15 establishes that the absolute loss error satisfies:

$$|\ell_j(\boldsymbol{w}_j^d(t)) - \ell_j(\boldsymbol{w}_j^c(t))| \leq O\left(\frac{\log(\log(d))}{\log(d)} \right).$$

We compute the total absolute error for this part by summing these individual column errors weighted by their context frequencies π_j. Since $\sum_{j>K} \pi_j \leq 1$, the absolute error is bounded by the same term:

$$\sum_{j>K} \pi_j |\ell_j(\boldsymbol{w}_j^d) - \ell_j(\boldsymbol{w}_j^c)| \leq O\left(\frac{\log(\log(d))}{\log(d)} \right).$$

Normalizing this absolute error by the initial loss $\mathcal{L}(W(0)) - \mathcal{L}(W^*) \sim \frac{1}{2}\log(d)$, we get

$$\frac{\sum_{j>K} \pi_j |\ell_j(\boldsymbol{w}_j^d) - \ell_j(\boldsymbol{w}_j^c)|}{\mathcal{L}(W(0)) - \mathcal{L}(W^*)} \leq O\left(\frac{\log(\log(d))}{\log^2(d)} \right).$$

Combining these results, we obtain:

$$|r_d(t) - r_c(t)| \leq O\left(\frac{\log(\log(d))}{\log(d)} \right) + O\left(\frac{\log(\log(d))}{\log^2(d)} \right) = O\left(\frac{\log(\log(d))}{\log(d)} \right) \xrightarrow[d\to\infty]{} 0.$$

□

D.6 Proof of the Limit of GD

In this section, we give the full proof of Theorem 3.5, following the sketch given in the main text.

Lemma 3.4 (Equivalence of softmax and Exponential Dynamics)**.** *Let $\boldsymbol{W}^{\exp}(t)$ denote the weight trajectory governed by $\dot{\boldsymbol{W}}_{k,j}^{\exp}(t) = -\eta\pi_j(\exp(\boldsymbol{W}_{k,j}^{\exp}(t)) - \pi_{k|j})$. Then, the difference in relative error between the exact softmax continuous flow $\boldsymbol{W}^{\mathrm{F}}(t)$ and the exponential flow $\boldsymbol{W}^{\exp}(t)$ vanishes:*

$$|r(\boldsymbol{W}^{\mathrm{F}}(t)) - r(\boldsymbol{W}^{\exp}(t))| = O\left(\frac{\log(\log(d))}{\log(d)}\right). \tag{6}$$

Proof of Lemma 3.4. Let $N(t) = \mathcal{L}(\boldsymbol{W}(t)) - \mathcal{L}^*$ be the exact numerator of the relative error and $\hat{N}(t)$ be the approximated numerator where we set $Z_j(s) = 1$. We define the absolute difference $\Delta(t) = |N(t) - \hat{N}(t)|$. Substituting the weight trajectory into the loss gap, we have:

$$N(t) = \sum_{j,k} \pi_j \pi_{k|j} \left[\log \pi_{k|j} - t\eta\pi_j\pi_{k|j} + \log \underbrace{\left(d + \eta\pi_j \int_0^t \frac{e^{s\eta\pi_j\pi_{k|j}}}{Z_j(s)}\, \mathrm{d}s \right)}_{:=A_{j,k}} + \log Z_j(t) \right].$$

For the approximated numerator $\hat{N}(t)$, substituting $Z_j(s) = 1$ allows us to compute the integral:

$$\eta\pi_j \int_0^t e^{s\eta\pi_j\pi_{k|j}}\, \mathrm{d}s = \frac{1}{\pi_{k|j}}(e^{t\eta\pi_j\pi_{k|j}} - 1),$$

$$\text{giving } \hat{N}(t) = \sum_{j,k} \pi_j \pi_{k|j} \left[\log \pi_{k|j} - t\eta\pi_j\pi_{k|j} + \log \underbrace{\left(d + \frac{1}{\pi_{k|j}} \left(e^{t\eta\pi_j\pi_{k|j}} - 1 \right) \right)}_{:=B_{j,k}} \right].$$

The absolute difference is then given by:

$$\Delta(t) = \left| \sum_{j=1}^{d} \sum_{k=1}^{d} \pi_j \pi_{k|j} \left[\log\left(\frac{A_{j,k}}{B_{j,k}} \right) + \log Z_j(t) \right] \right|.$$

To bound this, we use : $1 \le Z_j(s) \le H_d$:

$$\frac{1}{H_d} \int_0^t e^{s\eta\pi_j\pi_{k|j}}\, \mathrm{d}s \le \int_0^t \frac{e^{s\eta\pi_j\pi_{k|j}}}{Z_j(s)}\, \mathrm{d}s \le \int_0^t e^{s\eta\pi_j\pi_{k|j}}\, \mathrm{d}s.$$

Translating this to $A_{j,k}$ and $B_{j,k}$, we obtain $d + \frac{B_{j,k}-d}{H_d} \le A_{j,k} \le B_{j,k}$. Since $d \ge \frac{d}{H_d}$, it follows that $\frac{1}{H_d} B_{j,k} \le A_{j,k} \le B_{j,k}$. Taking the logarithm gives:

$$-\log H_d \le \log\left(\frac{A_{j,k}}{B_{j,k}} \right) \le 0.$$

On the other hand, the bounds on $Z_j(t)$ imply that $0 \le \log Z_j(t) \le \log H_d$. This means that:

$$-\log H_d \le \log(A_{j,k}/B_{j,k}) + \log Z_j(t) \le \log H_d \implies |\log(A_{j,k}/B_{j,k}) + \log Z_j(t)| \le \log H_d.$$

We can now bound the total absolute difference $\Delta(t)$:

$$\Delta(t) \le \sum_{j=1}^{d} \sum_{k=1}^{d} \pi_j \pi_{k|j} |\log(A_{j,k}/B_{j,k}) + \log Z_j(t)| \le \log H_d \sum_{j=1}^{d} \sum_{k=1}^{d} \pi_j \pi_{k|j} = \log H_d.$$

We have established that $\Delta(t) = O(\log(\log(d)))$. Finally:

$$|r(\boldsymbol{W}(t)) - r(\boldsymbol{W}^{\exp})(t)| = \frac{\Delta(t)}{L_0 - \mathcal{L}^*} = O\left(\frac{\log(\log(d))}{\log(d)}\right).$$

As $d \to \infty$, this term vanishes, proving that $r(t) \sim \hat{r}(t)$. □

Proposition D.17 (Continuous approximation via Euler–Maclaurin). *Let $t = d^\tau$ with $\tau \in [0, 2]$. Then, uniformly in τ,*

$$\hat{r}(d^\tau) \;=\; I_d \;+\; O\!\left(\frac{1}{\log(d)}\right), \qquad I_d := \frac{2}{\log(d)} \int_1^d \!\!\int_1^d F_d(x, y)\; \mathrm{d}x\; \mathrm{d}y.$$

where the continuous summand function $F_d(x, y)$ is defined as:

$$F_d(x, y) = \frac{1}{xyH_d^2} \log \underbrace{\left(1 - e^{-\frac{d^\tau}{xy}} + e^{-\frac{d^\tau}{xy}} \frac{d}{yH_d}\right)}_{:=A(x,y)}.$$

Proof. **Step 1: Euler-Maclaurin theorem** Let $G(x) = \sum_{k=1}^d F_d(x, k)$. Applying the first-order 1D Euler-Maclaurin formula to this inner sum yields:

$$G(x) = \frac{F_d(x, 1) + F_d(x, d)}{2} + \int_1^d F_d(x, y)\; \mathrm{d}y + R_x,$$

where the remainder is bounded by the integral of the derivative: $|R_x| \le \int_1^d \left|\frac{\partial F_d}{\partial y}(x, y)\right| \mathrm{d}y$. Summing $G(j)$ over $j \in \{1, \dots, d\}$ and applying Euler-Maclaurin again to the outer sum gives:

$$\sum_{j=1}^d G(j) = \frac{G(1) + G(d)}{2} + \int_1^d G(x)\; \mathrm{d}x + R,$$

where the final remainder R is bounded by $|R| \le \int_1^d \sum_{k=1}^d \left|\frac{\partial F_d}{\partial x}(x, k)\right| \mathrm{d}x$. By substituting $G(x)$ back into this expression, we obtain the exact 2D decomposition:

$$\sum_{j=1}^d \sum_{k=1}^d F_d(j, k) = \int_1^d \!\!\int_1^d F_d(x, y)\, \mathrm{d}x\, \mathrm{d}y + \mathcal{B}_{0D} + \mathcal{B}_{1D} + \frac{R_1 + R_d}{2} + \int_1^d R_x\; \mathrm{d}x + R,$$

$$\text{where } \mathcal{B}_{0D} = \frac{1}{4}[F_d(1, 1) + F_d(1, d) + F_d(d, 1) + F_d(d, d)],$$

$$\text{and } \mathcal{B}_{1D} = \frac{1}{2}\left[\int F_d(1, y)\, \mathrm{d}y + \int F_d(d, y)\, \mathrm{d}y + \int F_d(x, 1)\, \mathrm{d}x + \int F_d(x, d)\, \mathrm{d}x\right].$$

Step 2: Bounding the boundary terms First, we establish a global bound on $|\log A(x, y)|$. We can rewrite $A(x, y)$ as a convex combination:

$$A(x, y) = (1 - e^{-z}) \cdot 1 + e^{-z} \left(\frac{d}{yH_d}\right).$$

Since the domain restricts $y \le d$, we know $d/y \ge 1$ and $d/(yH_d) \ge 1/H_d$. Therefore, $A(x, y)$ is a convex combination of 1 and a value $\ge 1/H_d$. This guarantees its minimum is at least $1/H_d$:

$$A(x, y) \ge \min\left(1, \frac{1}{H_d}\right) = \frac{1}{H_d}.$$

As a result, the logarithm is bounded from below by $\log A(x, y) \ge -\log H_d$. Since we already established the upper bound $A(x, y) \le 1 + \frac{d}{yH_d} \le 1 + d$, we have:

$$-\log H_d \le \log A(x, y) \le \log(1 + d).$$

Because $H_d \le 1 + d$, this gives the absolute bound $|\log A(x, y)| \le \log(1 + d)$. Thus, the absolute value of the function is globally bounded by:

$$|F_d(x, y)| \le \frac{|\log A(x, y)|}{xyH_d^2} \le \frac{\log(1 + d)}{xyH_d^2} \sim \frac{\log(d)}{xyH_d^2}.$$

This implies that $|F_d(1, 1)| \le \log(1 + d)/H_d^2 \sim 1/\log(d) = O(1)$. We can apply the same reasoning to the other corners, yielding $\mathcal{B}_{0D} = O(1)$. Similarly, $\int_1^d |F_d(1, y)|\, \mathrm{d}y \le \log(d + 1)/H_d^2 \int_1^d \frac{1}{y}\, \mathrm{d}y$ which is $O(1)$. We can do the same for the other terms of $\mathcal{B}_{1D}$. Thus, $\mathcal{B}_{1D} = O(1)$.

Step 3: Computing the Partial Derivative. To bound the remainders, we must differentiate $F_d(x, y)$:

$$\frac{\partial F_d}{\partial y}(x,y) = \frac{\partial}{\partial y}\left[\frac{1}{xyH_d^2}\log A(x,y)\right], = -\frac{1}{xy^2H_d^2}\log A(x,y) + \frac{1}{xyH_d^2}\frac{1}{A(x,y)}\frac{\partial A(x,y)}{\partial y}.$$

$$\begin{aligned}\frac{\partial A(x,y)}{\partial y} &= \frac{\partial}{\partial y}\left(1 - e^{-\frac{d^\tau}{xy}} + e^{-\frac{d^\tau}{xy}}\frac{d}{yH_d}\right),\\ &= -e^{-\frac{d^\tau}{xy}}\left(\frac{d^\tau}{xy^2}\right) + e^{-\frac{d^\tau}{xy}}\left(\frac{d^\tau}{xy^2}\right)\frac{d}{yH_d} - e^{-\frac{d^\tau}{xy}}\frac{d}{y^2H_d},\\ &= \frac{1}{y}\left[e^{-\frac{d^\tau}{xy}}\left(\frac{d^\tau}{xy}\right)\left(\frac{d}{yH_d} - 1\right) - e^{-\frac{d^\tau}{xy}}\frac{d}{yH_d}\right].\end{aligned}$$

To analyze this expression, we introduce $z = \frac{d^\tau}{xy}$ and $c(y) = \frac{d}{yH_d} - 1$. Notice that $A(x,y) = 1 + e^{-z}c(y)$. Substituting these into the logarithmic derivative yields:

$$\frac{1}{A(x,y)}\frac{\partial A(x,y)}{\partial y} = \frac{1}{y}\Bigg[\underbrace{\frac{ze^{-z}c(y)}{1+e^{-z}c(y)}}_{:=h(z)} - \underbrace{\frac{e^{-z}\frac{d}{yH_d}}{1+e^{-z}c(y)}}_{:=g(z,y)}\Bigg].$$

Step 4: Bounding the Derivative Terms We bound the two distinct terms inside the bracket:

- **Bounding $g(z,y)$:** The denominator expands to $1 - e^{-z} + e^{-z}d/(yH_d)$. Since $z > 0$, we have that $1 - e^{-z} > 0$. The denominator is greater than the numerator $e^{-z}d/(yH_d)$, and $g(z,y) \le 1$.
- **Bounding $h(z)$:** Assume $c(y) \ge 0$. We use the standard bound $\frac{X}{X+1} \le \min(1, X)$ with $X = e^{-z}c(y)$, which gives:
$$h(z) \le \min\big(z,\ zc(y)e^{-z}\big).$$
We analyze this over the domain of z:
  - **If $z \le \log c(y)$:** We automatically have $h(z) \le z \le \log c(y)$.
  - **If $z > \log c(y)$:** We separate the cases depending on the magnitude of $c(y)$:
    - If $c(y) > e$: The function $u \mapsto ue^{-u}$ is decreasing for $u > 1$. Since $z > \log c(y) > 1$, we have $ze^{-z} < (\log c(y))e^{-\log c(y)} = \frac{\log c(y)}{c(y)}$. Thus, $h(z) \le c(y)\frac{\log c(y)}{c(y)} = \log c(y)$.
    - If $0 < c(y) \le e$: The global maximum of $u \mapsto ue^{-u}$ is $1/e$, so $h(z) \le c(y)\frac{1}{e} \le 1$.

  If $-1 < c(y) < 0$, the absolute value satisfies:
$$|h(z)| = \frac{ze^{-z}|c(y)|}{1 - e^{-z}|c(y)|} \le \frac{ze^{-z}}{1-e^{-z}} = \frac{z}{e^z - 1} \le 1.$$

Combining all cases, we conclude that:

$$|h(z)| \le \log(1 + |c(y)|) + 1.$$

and since $c(y) + 1 \le d$, we have:

$$|h(z)| \le \log(d) + 1.$$

Substituting $h(z) \le \log(d) + 1$ and $g(z,y) \le 1$ back into the product rule bounds the partial derivative:

$$\left|\frac{\partial F_d}{\partial y}(x,y)\right| \le \frac{\log(1+d)}{xy^2H_d^2} + \frac{1}{xyH_d^2}\frac{\log(d)+1+1}{y} = O\left(\frac{\log(d)}{xy^2H_d^2}\right).$$

Step 5: Bounding the Remainders via Exact Differentiation

With the partial derivative w.r.t. y bounded by $O\left(\frac{\log(d)}{xy^2H_d^2}\right)$, we integrate over y to bound R_x:

$$|R_x| \le \int_1^d \frac{\log(d)}{xy^2H_d^2}\,\mathrm{d}y = \frac{\log(d)}{xH_d^2}\left[-y^{-1}\right]_1^d \le \frac{\log(d)}{xH_d^2}.$$

This allows us to immediately evaluate the specific remainders:

$$|R_1| \leq \frac{\log(d)}{H_d^2} = O(1) \quad \text{and} \quad \left| \int_1^d R_x \, \mathrm{d}x \right| \leq \frac{\log(d)}{H_d^2} \int_1^d x^{-1} \, \mathrm{d}x = \frac{\log^2(d)}{H_d^2} = O(1).$$

Finally, to bound the global remainder R, we must explicitly differentiate $F_d(x, y)$ with respect to x:

$$\frac{\partial F_d}{\partial x}(x, y) = -\frac{1}{x^2 y H_d^2} \log A(x, y) + \frac{1}{xyH_d^2} \frac{1}{A(x, y)} \frac{\partial A(x, y)}{\partial x}.$$

Differentiating $A(x, y) = 1 + e^{-\frac{d^\tau}{xy}} c(y)$ with respect to x yields:

$$\frac{\partial A(x, y)}{\partial x} = c(y) e^{-\frac{d^\tau}{xy}} \left(\frac{d^\tau}{x^2 y} \right) = \frac{1}{x} \left(\frac{d^\tau}{xy} \right) e^{-z} c(y) = \frac{1}{x} z e^{-z} c(y).$$

Substituting this into the logarithmic derivative, we recover the exact same function $h(z)$ we analyzed:

$$\frac{1}{A(x, y)} \frac{\partial A(x, y)}{\partial x} = \frac{1}{x} \left[\frac{z e^{-z} c(y)}{1 + e^{-z} c(y)} \right] = \frac{1}{x} h(z).$$

Since we already proved that $|h(z)| \leq \log(d) + 1$, and knowing $|\log A(x, y)| \leq \log(1 + d) \sim \log(d)$, we bound the partial derivative:

$$\left| \frac{\partial F_d}{\partial x}(x, y) \right| \leq \frac{\log(d + 1)}{x^2 y H_d^2} + \frac{1}{xyH_d^2} \frac{\log(d) + 1}{x} = O\left(\frac{\log(d)}{x^2 y H_d^2} \right).$$

To bound the final cross-remainder R, summing this derivative over k evaluates to the harmonic number H_d:

$$\sum_{k=1}^{d} \left| \frac{\partial F_d}{\partial x}(x, k) \right| \leq O\left(\frac{\log(d)}{x^2 H_d^2} \sum_{k=1}^{d} \frac{1}{k} \right) = O\left(\frac{\log(d)}{x^2 H_d^2} H_d \right) = O\left(\frac{\log(d)}{x^2 H_d} \right) = O\left(\frac{1}{x^2} \right).$$

Integrating this bound over x yields:

$$|R| \leq \int_1^d O\left(\frac{1}{x^2} \right) \mathrm{d}x = O\left(\left[-\frac{1}{x} \right]_1^d \right) = O(1).$$

Since all boundary and remainder terms are bounded by a constant $O(1)$, the absolute error between the sum and the integral is $O(1)$. Now, by re-normalizing by $\frac{2}{\log(d)}$, we have:

$$r(d^\tau) = I_d + O\left(\frac{1}{\log(d)} \right).$$

□

Theorem 3.5 (Asymptotic Limit of the Relative Error)**.** *Under the time scaling $t = d^\tau$ with $\tau \in [0, 2]$, the relative error converges as $d \to \infty$ to:*

$$\lim_{d \to \infty} r(\boldsymbol{W}(d^\tau)) = 2 \iint_{\tau < u + v} (1 - v) \, \mathrm{d}u \, \mathrm{d}v = \begin{cases} 1 - \tau^2 + \frac{\tau^3}{3} & \text{if } \tau \in [0, 1], \\ \frac{(2 - \tau)^3}{3} & \text{if } \tau \in [1, 2]. \end{cases}$$

Proof of Theorem 3.5. Applying the change of variables described in the main text, we obtain:

$$I_d = \int_0^1 \int_0^1 \underbrace{\frac{2}{\log(d)} \log\left(1 - e^{-d^{\tau - u - v}} + e^{-d^{\tau - u - v}} \frac{d^{1 - v}}{H_d} \right)}_{:= f_d(u, v)} \mathrm{d}u \, \mathrm{d}v.$$

We analyze the pointwise limit of the integrand $f_d(u, v)$ as $d \to \infty$ in two distinct regimes:

Case 1: $\tau > u + v$**.** In this region, $\tau - u - v > 0$, meaning $d^{\tau-u-v} \to +\infty$ as $d \to \infty$, and

$$\lim_{d\to\infty} f_d(u, v) = \lim_{d\to\infty} \frac{2}{\log(d)} \log(1) = 0.$$

Case 2: $\tau < u + v$**.** Here, $\tau - u - v < 0$ implies $d^{\tau-u-v} \to 0$ and $e^{-d^{\tau-u-v}} \to 1$. Because $v < 1$, the term proportional to d^{1-v} dominates inside the logarithm. We can factorize the argument as:

$$1 - e^{-d^{\tau-u-v}} + e^{-d^{\tau-u-v}} \frac{d^{1-v}}{H_d} = \frac{d^{1-v}}{H_d}\big(1 + o(1)\big).$$

Taking the logarithm of this dominant term yields:

$$\begin{aligned}\log\left(\frac{d^{1-v}}{H_d}\big(1 + o(1)\big)\right) &= \log\left(d^{1-v}\right) - \log(H_d) + \log\big(1 + o(1)\big) \\ &= (1 - v)\log(d) + O(\log(\log(d))).\end{aligned}$$

Multiplying by the global prefactor $2/\log(d)$, the limit of the integrand becomes:

$$\lim_{d\to\infty} f_d(u, v) = \lim_{d\to\infty} \frac{2}{\log(d)}\Big[(1 - v)\log(d) - O(\log(\log(d)))\Big] = 2(1 - v).$$

Since the integrand is bounded across the domain $[0, 1]^2$, we use the dominated convergence theorem to interchange the limit and the double integral:

$$\begin{aligned}\lim_{d\to\infty} I_d &= \int_0^1 \int_0^1 \lim_{d\to\infty} f_d(u, v)\ \mathrm{d}u\ \mathrm{d}v \\ &= \iint_{\tau<u+v} 2(1 - v)\ \mathrm{d}u\ \mathrm{d}v.\end{aligned}$$

□

D.7 Sequential learning

This section proves the extension of the argument to show sequential learning of input words.

Proposition 3.7 (Limit of the Per-Context Relative Error)**.** *Under the time scaling $t = d^\tau$, the relative error associated with input word $j = d^u$ for $u \in (0, 1)$ converges to:*

$$\lim_{d\to\infty} r_{d^u}(\boldsymbol{W}(d^\tau)) = (1 - \max(0, \tau - u))^2 \mathbb{1}_{\{\tau-u\le 1\}}.$$

Proof of Proposition 3.7. For a fixed log-rank $u_j \in (0, 1)$, consider the input word $j = \lfloor d^{u_j} \rfloor$ with log-rank $u_j = \log(j)/\log(d) = u$. We can apply Lemma 3.2 and Lemma 3.4 to the specific column, as the effective step-size for the jth column scales as $\eta\pi_j = \eta d^{-u_j} = O(\log(d)/d^{u_j})$. We may replace the exact trajectory of column j by the decoupled exponential flow and omit the $\log Z_j$ term. This yields the asymptotic equivalence:

$$r_j(\boldsymbol{W}(t)) \sim \frac{2}{\log(d)} \sum_{k=1}^{d} \frac{1}{kH_d} \log\left(1 - e^{-\frac{t}{jk}} + e^{-\frac{t}{jk}} \frac{d}{kH_d}\right),$$

Applying the 1D Euler-Maclaurin formula (following the same bounding steps as in Proposition D.17) to this sum, and setting the time scale $t = d^\tau$ and the continuous variable $k = d^v$ (with $\mathrm{d}k = \log(d)d^v\ \mathrm{d}v$), the sum transforms into an integral over $v \in [0, 1]$:

$$r_j(\boldsymbol{W}(d^\tau)) \sim \int_0^1 \frac{2}{\log(d)} \log\left(1 - e^{-d^{\tau-u_j-v}} + e^{-d^{\tau-u_j-v}} \frac{d^{1-v}}{H_d}\right) \mathrm{d}v.$$

As in the proof of Theorem 3.5, the integrand converges pointwise as $d \to \infty$. When $\tau < u_j + v$, the exponential $\exp(-d^{\tau-u_j-v}) \to 1$, and the term d^{1-v}/H_d dominates the argument of the logarithm, yielding $(2/\log(d))(1 - v)\log(d) = 2(1 - v)$. When $\tau > u_j + v$, the argument of the logarithm tends to 1, yielding 0. Thus, the entire integrand converges pointwise to $2(1 - v)\mathbf{1}_{\{\tau<u_j+v\}}$.

Since the integrand is globally bounded, we apply the dominated convergence theorem:

$$\lim_{d\to\infty} r_j(\boldsymbol{W}(d^\tau)) = \int_0^1 2(1-v)\mathbf{1}_{\{\tau<u_j+v\}}\ \mathrm{d}v = 2\int_{\max(0,\tau-u_j)}^1 (1-v)\ \mathrm{d}v.$$

Evaluating this final integral exactly gives:

$$\lim_{d\to\infty} r_j(\boldsymbol{W}(d^\tau)) = \big(1-\max(0,\tau-u_j)\big)^2\mathbf{1}_{\{\tau-u_j\le 1\}}. \qquad \square$$

E Dynamics of Sign Descent

E.1 The oscillatory phase

This section proves Proposition 4.2, describing how the weights reach the oscillatory regime.

Proposition 4.2 (Weight Dynamics)**.** *Let $\log(\pi_k) + \log Z(t)$ be the (moving) target of weight $W_k(t)$. Under sign descent, W_k always increases (or always decreases) until it reaches its target, then oscillates around it. That is, if $W_k(t) - \log(\pi_k) - \log Z(t)$ first changes sign at T_k,*

$$W_k(t) = W_0 \pm \eta t \text{ for } t \leq T_k, \qquad \text{and} \qquad |W_k(t) - \log(\pi_k) - \log Z(t)| \leq 2\eta \text{ for } t > T_k.$$

The system reaches the oscillatory regime when all weights oscillate, at $T_{\text{osc}} := \max\{T_1, \dots, T_d\}$.

Proof of Proposition 4.2. As long as the weights are below (resp. above) their target, they increase (resp. decrease) by η at each step by definition of the sign update. What we need to show is that if the update on $\boldsymbol{W}_k$ crosses the target, it does not overshoot by more than 2η. This happens because the weights change by exactly η at each step, and the target can move by at most η, as

$$\log Z(t+1) = \log \sum_{k=1}^{d} e^{\boldsymbol{W}_k(t+1)} = \log \sum_{k=1}^{d} e^{\boldsymbol{W}_k(t) \pm \eta} \in [\log Z(t) - \eta, \log Z(t) + \eta]. \qquad \square$$

E.2 Picking the step-size

To pick the step-size for sign descent, we need η to be large enough so that every weight reaches the oscillatory regime, but as small as possible to limit the error due to the oscillations. To estimate the required scaling between η, T and $\log(d)$, we use continuous sign descent as an approximation and derive a lower bound on η required so that each weight reaches its target after time T. We confirm that this step-size generalizes to the discrete case experimentally in the main text.

Proposition 4.3 (Minimum step-size for sign flow)**.** *For sign flow with step-size η,*

$$\dot{\boldsymbol{w}}(t) = -\eta \operatorname{sign}(\boldsymbol{w}(t) - \log \pi - \log Z(t)),$$

the step-size needs to be at least $\eta \geq \log(d)/2T$ to hit the solution at time T.

Proof of proposition 4.3. Equivalently, time to hit the solution is $T \geq \log(d)/2\eta$. The solution of the problem is

$$\{\boldsymbol{w} \in \mathbb{R}^d : \boldsymbol{w} = \log \pi + c\mathbf{1} : c \in \mathbb{R}\}.$$

Since $|\dot{w}_k| \leq \eta$, $\|\boldsymbol{w}(t) - \boldsymbol{w}(0)\|_\infty \leq \eta t$. Hitting the solution set at time T requires

$$\eta T \geq \min_{c \in \mathbb{R}} \|\log \pi + c\mathbf{1} - \boldsymbol{w}(0)\|_\infty.$$

The solution in c^* is found at the midpoint of the largest and smallest entries of $\log \pi - \boldsymbol{w}(0)$, giving

$$\begin{aligned} c^* &= \frac{1}{2}(\max_i \log \pi_i - \boldsymbol{w}_i(0) + \min_i \log \pi_i - \boldsymbol{w}_i(0)), \\ \text{and } \eta T &\geq \min_{c \in \mathbb{R}} \|\log \pi + c\mathbf{1} - \boldsymbol{w}(0)\|_\infty, \\ &= \frac{1}{2}(\max_i \log \pi_i - \boldsymbol{w}_i(0) - \min_i \log \pi_i - \boldsymbol{w}_i(0)), \\ &= \frac{1}{2}(\max_i - \log(iH_d) - \min_i - \log(iH_d)), = \frac{1}{2}\log(d). \end{aligned} \qquad \square$$

E.3 Limit theorems

Under Assumption 4.4, $\mathbf{W}_k(T) = \log \pi_k + c + \delta_k$ with $\delta_1, \dots, \delta_d$ i.i.d. $\mathcal{U}([-\eta, \eta])$. The excess loss admits the closed form

$$\mathcal{L} - \mathcal{L}^* = \sum_{k=1}^{d} \pi_k \log\left(\frac{\pi_k \sum_j \exp(\mathbf{W}_j)}{\exp(\mathbf{W}_k)}\right) = -A_d + \log(B_d), \tag{11}$$

with:

$$A_d := \sum_{k=1}^{d} \pi_k \delta_k, \qquad B_d := \sum_{k=1}^{d} \pi_k \exp(\delta_k).$$

Since δ_k has mean 0, variance $\eta^2/3$, fourth moment $\eta^4/5$, and $\mathbb{E}[e^{\delta_k}] = \sinh(\eta)/\eta$, we have:

$$\mathbb{E}[A_d] = 0, \qquad \mathbb{E}[B_d] = \frac{\sinh \eta}{\eta}, \tag{12}$$

$$\mathrm{Var}(A_d) = \frac{\eta^2}{3} \sum_{k=1}^{d} \pi_k^2, \qquad \mathrm{Var}(B_d) = \mathrm{Var}(e^{\delta_1}) \sum_{k=1}^{d} \pi_k^2. \tag{13}$$

This gives:

$$\mathrm{Var}(A_d) \sim \frac{\pi^2}{72\, T^2} \sim \frac{\pi^2 \eta^2}{18\, H_d^2}. \tag{14}$$

Theorem 4.5 (Asymptotic Relative Error of Sign Descent)**.** *Under Assumption 4.4, sign descent with step-size $\eta = \log(d)/(2T)$, as $d \to \infty$, the relative error satisfies*

$$r(\boldsymbol{W}(T)) \;=\; \frac{2}{\log(d)} \log\!\left(\frac{\sinh \eta}{\eta}\right) \big(1 + o_{\mathbb{P}}(1)\big).$$

Proof of Theorem 4.5. By equation 11,

$$\mathcal{L} - \mathcal{L}^* = -A_d + \log(B_d).$$

Any sequence $T = T(d)$ falls, along a subsequence, into one of two regimes: either η is bounded below by some $\eta_0 > 0$, or $\eta \to 0$. We prove $\mathcal{L} - \mathcal{L}^* = \log(\sinh \eta/\eta)(1 + o_{\mathbb{P}}(1))$ in each case, then normalize by $\mathcal{L}(\boldsymbol{W}(0)) - \mathcal{L}^* \sim \log(d)/2$.

Regime A: η bounded below by some $\eta_0 > 0$. We show

$$\frac{\mathcal{L} - \mathcal{L}^*}{\log(\sinh \eta/\eta)} = -\frac{A_d}{\log(\sinh \eta/\eta)} + \frac{\log(B_d)}{\log(\sinh \eta/\eta)} \xrightarrow{\mathbb{P}} 0 + 1 = 1.$$

By Chebyshev, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{A_d}{\log(\sinh \eta/\eta)}\right| > \varepsilon\right) \le \frac{\mathrm{Var}(A_d)}{\varepsilon^2 \log^2(\sinh \eta/\eta)} = O\left(\frac{\eta^2}{H_d^2 \log^2(\sinh \eta/\eta)}\right).$$

If $\eta \to \infty$, $\log(\sinh \eta/\eta) \sim \eta$, so the ratio is $O(1/H_d^2) \to 0$. If η remains bounded, η^2 and $\log^2(\sinh \eta/\eta)$ are both bounded away from 0 and ∞, so the ratio is again $O(1/H_d^2) \to 0$. Either way, $A_d/\log(\sinh \eta/\eta) \xrightarrow{\mathbb{P}} 0$.

Writing

$$\frac{\mathrm{Var}(B_d)}{(\mathbb{E}[B_d])^2} = \left[\frac{\sinh(2\eta)}{2\eta\,(\sinh \eta/\eta)^2} - 1\right] \sum_k \pi_k^2,$$

we check the bracket in each subcase:

- If $\eta \to \infty$: $\sinh(2\eta)/(2\eta) \sim e^{2\eta}/(4\eta)$ strictly dominates $(\sinh \eta/\eta)^2 \sim e^{2\eta}/(4\eta^2)$, so the bracket is $\sim \eta$, giving $\mathrm{Var}(B_d)/(\mathbb{E}[B_d])^2 = O(\eta/H_d^2) \to 0$.
- If η remains bounded, the bracket is bounded (and $(\mathbb{E}[B_d])^2$ is bounded away from 0), so $\mathrm{Var}(B_d)/(\mathbb{E}[B_d])^2 = O(1/H_d^2) \to 0$.

By Chebyshev, $B_d/\mathbb{E}[B_d] \xrightarrow{\mathbb{P}} 1$. So, $\log(B_d) = \log(\sinh \eta/\eta) + o_{\mathbb{P}}(1)$ and $\log(B_d)/\log(\sinh \eta/\eta) \xrightarrow{\mathbb{P}} 1$.

Combining, $(\mathcal{L} - \mathcal{L}^*)/\log(\sinh \eta/\eta) \xrightarrow{\mathbb{P}} 1$, so $\mathcal{L} - \mathcal{L}^* = \log(\sinh \eta/\eta)(1 + o_{\mathbb{P}}(1))$.

Regime B: $\eta \to 0$**.** Now $\log(\sinh\eta/\eta) \sim \eta^2/6 \to 0$. Approximating $\log(B_d)$ by $\log(\sinh\eta/\eta) + o_{\mathbb{P}}(1)$ as in Regime A no longer suffices.

Since $|\delta_k| \le \eta$, the bound $|e^x - 1 - x - x^2/2| \le e^{\eta}|x|^3/6$ gives:

$$B_d = 1 + A_d + \frac{C_d}{2} + R_d, \qquad C_d := \sum_{k=1}^{d} \pi_k \delta_k^2, \qquad |R_d| \le \frac{e^{\eta}\eta^3}{6} = O(\eta^3). \tag{15}$$

Since $A_d, C_d/2, R_d$ are all $o_{\mathbb{P}}(1)$, $B_d - 1 = o_{\mathbb{P}}(1)$, and $\log(1+x) = x - x^2/2 + O(x^3)$ gives

$$\log(B_d) = A_d + \frac{C_d}{2} - \frac{A_d^2}{2} + O_{\mathbb{P}}(\eta^3).$$

The linear term cancels in the excess loss:

$$\mathcal{L} - \mathcal{L}^* = -A_d + \log(B_d) = \frac{C_d}{2} - \frac{A_d^2}{2} + O_{\mathbb{P}}(\eta^3). \tag{16}$$

We now control the three terms after dividing by $\eta^2/6$.

Control of $(C_d/2)/(\eta^2/6)$. $\mathbb{E}[C_d/2] = \eta^2/6$, and using $\mathrm{Var}(\delta_1^2) = \eta^4/5 - (\eta^2/3)^2 = 4\eta^4/45$,

$$\mathrm{Var}(C_d/2) = \frac{\mathrm{Var}(\delta_1^2)}{4} \sum_k \pi_k^2 = O\left(\frac{\eta^4}{H_d^2}\right), \qquad \frac{\mathrm{Var}(C_d/2)}{(\mathbb{E}[C_d/2])^2} = O(H_d^{-2}) \to 0.$$

By Chebyshev, $(C_d/2)/(\eta^2/6) \xrightarrow{\mathbb{P}} 1$.

Control of $(A_d^2/2)/(\eta^2/6)$. From equation 14, $\mathbb{E}[A_d^2] = \mathrm{Var}(A_d) = O(\eta^2/H_d^2)$, so

$$\mathbb{P}\left(\frac{A_d^2/2}{\eta^2/6} > \varepsilon\right) \le \frac{\mathbb{E}[A_d^2]}{\varepsilon \cdot \eta^2/3} = O(H_d^{-2}) \to 0.$$

Control of $R_d/(\eta^2/6)$. Since $R_d = O(\eta^3)$ deterministically and $\eta \to 0$, $R_d/(\eta^2/6) = O(\eta) \to 0$.

Combining, $(\mathcal{L} - \mathcal{L}^*)/(\eta^2/6) \xrightarrow{\mathbb{P}} 1$, so $\mathcal{L} - \mathcal{L}^* = (\eta^2/6)(1 + o_{\mathbb{P}}(1)) = \log(\sinh\eta/\eta)(1 + o_{\mathbb{P}}(1))$, using $\log(\sinh\eta/\eta) = \eta^2/6 + O(\eta^4)$ as $\eta \to 0$.

Conclusion. In both regimes, $\mathcal{L} - \mathcal{L}^* = \log(\sinh\eta/\eta)(1 + o_{\mathbb{P}}(1))$. Dividing by $\mathcal{L}(\boldsymbol{W}(0)) - \mathcal{L}^* \sim \log(d)/2$ concludes the proof. □