PROVABLE BENEFIT OF SIGNGD: A MINIMAL MODEL UNDER HEAVY-TAILED CLASS IMBALANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive and non-Euclidean optimizers often outperform stochastic gradient descent (SGD) in language modeling by a large margin. Existing theory usually explains this gap by assuming favorable smoothness geometry or noise structure tailored to the specific optimizer. We instead ask whether such geometry can be induced from a concrete learning problem. Starting from an optimizer gap on realistic language-modeling experiments, we progressively remove sequence dependence, architectural complexity, and stochasticity. We find that the gap exists in a minimal setting: the softmax unigram model with heavy-tailed class imbalance. We prove that GD learns rare tokens slowly because the corresponding logits receive small updates, while SignGD removes this magnitude dependence and moves rare and common coordinates on a more comparable scale. We make this precise with upper and lower bounds for the convergence rate of GD and upper bounds for the convergence of SignGD. In the stochastic setting, we prove that mini-batch noise masks rather than creates this advantage, with sufficiently large batches yielding a separation between SignSGD and GD.

1 INTRODUCTION

Adaptive coordinate-wise methods are widely used for modern deep learning problems (Bernstein & Newhouse, 2024). In particular, the Adam optimizer (Kingma & Ba, 2017) and its variants (Loshchilov & Hutter, 2019) are prevalent in LLM pretraining, where they significantly surpass the performance of conventional SGD methods (Xie et al., 2025a; Kunstner et al., 2024). Despite this remarkable empirical success, we still lack a complete theoretical understanding of why adaptive coordinate-wise methods converge faster than SGD for language-modeling tasks.

A growing theoretical literature explains the gap between adaptive or non-Euclidean methods and Euclidean SGD by postulating optimizer-friendly geometry. These works derive convergence results under assumptions that include coordinate-wise or anisotropic smoothness (Liu et al., 2025; Bernstein & Newhouse, 2024; Jiang et al., 2025), adaptive smoothness (Xie et al., 2025a), and favorable stochastic-gradient geometry (Yu et al., 2026; Bernstein et al., 2018). These results clarify when particular algorithms benefit from specific geometric assumptions, but they leave open a basic question: how do such optimizer-friendly geometries arise in concrete learning problems? In particular, it remains unclear how the structure of LLM pretraining objective gives rise to geometry that favors non-Euclidean methods.

Our investigation is motivated by the recent finding that, under heavy-tailed token frequencies, adaptive or sign-based methods can outperform Euclidean methods even in simplified transformer models (Kunstner et al., 2024; 2023). This gap persists even in the full-batch setting, suggesting that it is not merely an artifact of stochastic gradient noise. At the same time, these simplified settings still retain several architectural and modeling ingredients, making it difficult to isolate the source of the gap between different optimizers. This leads us to ask:

> *Can we develop a minimal language model to isolate and rigorously explain the advantage of sign-based methods under heavy-tailed class imbalance?*

Our answer to this question is the *softmax unigram model*. Inspired by the simple transformer setting in Kunstner et al. (2024), we design a minimal language-modeling problem in which we can prove

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.

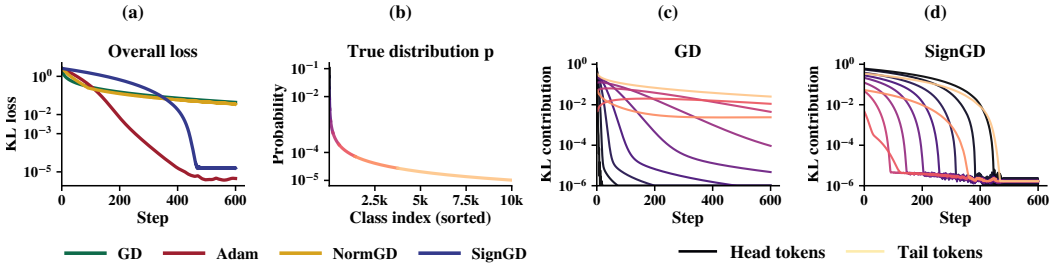


Figure 1: **GD struggles to learn the tail token probabilities on a simple deterministic softmax unigram model with heavy-tailed class imbalance**. The softmax unigram model objective is defined as $f_{\text{uni}}(\theta) = \text{KL}(p \| \text{softmax}(\theta))$ for logits θ. **(a)** Adam and SignGD outperform GD and normalized GD by a significant margin. **(b)** True token distribution for the softmax unigram model; black indicates head tokens, and yellow indicates tail tokens. **(c)** and **(d)** KL contributions grouped by token rank for GD and SignGD, respectively, where each group has mass 0.1. The contribution of token i is $p_i \log(p_i/q_i) - p_i + q_i$ for $q = \text{softmax}(\theta)$. These non-negative coordinate-wise terms, which sum to the total KL, show that GD fails to reduce loss on tail tokens, while SignGD makes more uniform progress across ranks. Although tail contributions are small, the large number of tail tokens renders them non-negligible.

that non-Euclidean steepest descent methods such as sign gradient descent (SignGD), normalized steepest descent with respect to the ℓ_∞ norm, converge faster than GD under heavy-tailed class imbalance. Despite its simplicity, the parameters of the softmax unigram model are coupled through the softmax activation, unlike the simpler separable quadratic models that are often used in prior analyses (Kunstner et al., 2024; Kunstner & Bach, 2025).

In the deterministic setting, this model isolates and reveals a simple mechanism behind the optimizer gap, as shown in Figure 1. Under heavy-tailed token frequencies, GD learns rare-token probabilities slowly because progress on tail coordinates is scaled by their small frequencies. In contrast, SignGD is less sensitive to these coordinate-wise magnitude imbalance, making progress on head and tail coordinates at a comparable rate. We formalize this contrast by deriving a lower bound for GD from its tail-learning bottleneck and an upper bound for SignGD strictly better in dimension dependence; together, these bounds establish a strict iteration-complexity separation on the softmax unigram model.

In the stochastic setting, the rate decomposes into the same deterministic component plus an additional noise term. We show that stochastic noise does not amplify the deterministic advantage of SignGD over GD for the softmax unigram model. Instead, it acts as a floor that can hide the advantage unless the noise is sufficiently small. Specifically, for sufficiently large batch sizes, we obtain an iteration-complexity separation between SignSGD with weight decay and full-batch GD (even though GD has access to the exact gradient).

Our contributions are:

1. In Section 3, we reduce a transformer language model to the softmax unigram model, preserving the optimizer gap under heavy-tailed class imbalance in a setting amenable to rigorous analysis.
2. In Section 4, we prove GD upper and lower bounds that expose its slow learning of rare tokens (Theorems 4.1 and 4.2). We prove that SignGD's ℓ_∞ geometry yields a $\widetilde{\mathcal{O}}((\log d)^2/T)$ rate, versus $\widetilde{\mathcal{O}}(d/T)$ for normalized GD both with and without weight decay .
3. In Section 5, we separate mini-batch noise from the deterministic loss geometry and derive fixed-budget rates for SignSGD, normalized SGD, and averaged-iterate SGD. With large batches, SignSGD with weight decay retains a separation from full-batch GD (Corollary 5.3).

2 Related Work

Adam, SGD, and heavy-tailed language data. Other existing explanations for the gap between Adam and SGD include heavy-tailed gradient noise, directional sharpness, and Hessian heterogeneity (Zhang et al., 2020; Pan & Li, 2023; Zhang et al., 2024). Recent work also emphasizes that batch size

is an essential hyperparameter for the Adam–SGD gap across various domains (Srećković et al., 2025). Glentis et al. (2026) emphasize effective learning rates and motivate clipping through frequency-dependent gradient bounds. We instead characterize convergence bounds under heavy-tailed tokens and prove an iteration-complexity separation between GD and SignGD.

Geometry-aware Optimizers. Many theoretical analyses of adaptive optimizers assume a favorable geometry and derive generic convergence guarantees under it. SignGD is tied to the ℓ_∞ geometry (Bernstein et al., 2018; Balles & Hennig, 2018); related non-Euclidean descent in this geometry was used for approximate max flow (Kelner et al., 2014). Adam and AdamW admit related coordinate-wise, constrained, or steepest-descent-like ℓ_∞ interpretations (Kunstner et al., 2023; Bernstein & Newhouse, 2024; Kunstner et al., 2024; Xie & Li, 2024; Xie et al., 2025a). AdaGrad-type methods exploit diagonal, anisotropic, or adaptive smoothness (Duchi et al., 2011; Ward et al., 2020; Liu et al., 2025; Jiang et al., 2025), while structured preconditioners such as Shampoo exploit richer matrix geometries (Gupta et al., 2018; Xie et al., 2025b). Recent work also studies adaptive variance and preconditioning, normalized steepest descent, and trust-region methods in general norms (Xie et al., 2026; Kovalev, 2026; Cutkosky & Mehta, 2020; Kovalev, 2025). The broader Frank–Wolfe family includes classical, stochastic, and momentum analyses (Jaggi, 2013; Hazan & Luo, 2016; Li et al., 2021; Pethick et al., 2025; Nazykov et al., 2024).

Simplified models. Prior works have developed simplified models to understand the benefits of adaptive optimizers such as Adam compared to SGD. For instance, Kunstner & Bach (2025) study SignGD and GD under a diagonal quadratic model (bigram model with MSE loss) with heavy-tailed class imbalance. Other quadratic models are inspired by transformer Hessian structure or batch-size effects (Zhang et al., 2024; Srećković et al., 2025). Gradient flow and continuous-time SignGD have been studied on a simple imbalanced problem (Kunstner et al., 2024). Instead of using a quadratic proxy, we introduce a new model by directly simplifying the transformer architecture and preserving the softmax objective. This gives a simplified language-modeling problem in which the softmax still couples coordinates, allowing us to isolate the effect of heavy-tailed class imbalance without introducing additional separable structure or using continuous-time analysis.

3 From Transformers to the Softmax Unigram Model

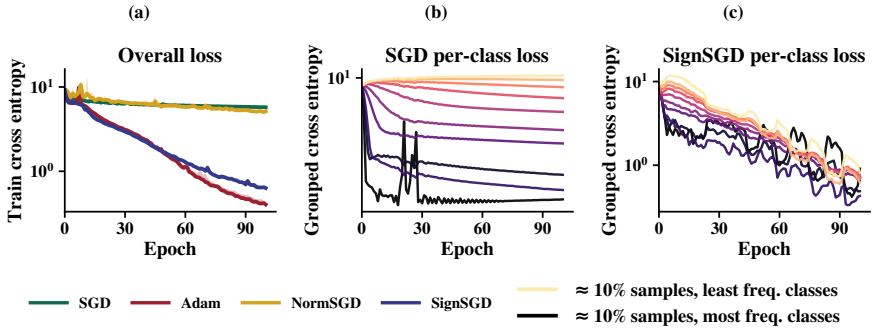
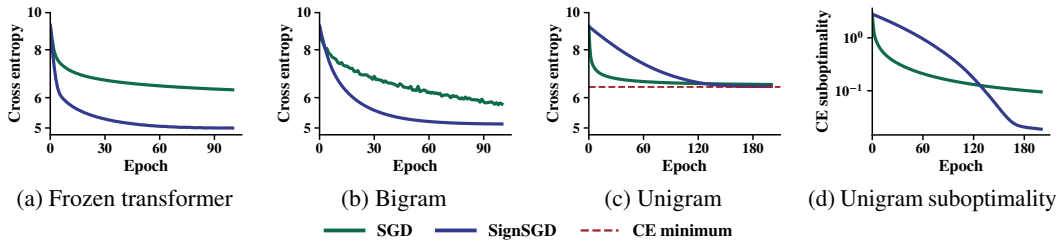


Figure 2: **SignSGD makes more uniform progress across token frequencies than SGD on a one-layer PTB transformer. (a)** Overall training loss. **(b,c)** Loss by ten frequency-sorted token groups for SGD and SignSGD, respectively (dark: most frequent; yellow: least frequent). Each group contains $\approx 10\%$ of samples; the overall losses in **(a)** average the corresponding groups in **(b,c)**.

In this section, we progressively simplify the transformer language modeling setup to the softmax unigram model. As a starting point, Figure 2 shows that adaptive and sign-based methods outperform Euclidean methods under heavy-tailed class imbalance, replicating the observations in Kunstner et al. (2024). However, this setting still involves many components: the embeddings, attention, classification head, sequence dependence, and mini-batch noise make it difficult to isolate the source of the gap and build a precise theory.

Therefore, we simplify the setup in two stages, verifying that the gap persists after every modification. First, we freeze every component except the classification head and remove attention. Then, we collapse all token embeddings to a single shared vector. Second, we remove mini-batch noise and train the same model in the full-batch setting. The full experimental details are in Appendix L.

3.1 Progressive Architectural Simplification



(a) Frozen transformer (b) Bigram (c) Unigram (d) Unigram suboptimality

Figure 3: The SignSGD–SGD gap persists as the model is simplified. **(a)** Freeze embeddings and attention; train only the classification head. **(b)** Remove attention to obtain a bigram model. **(c)** Share one embedding across tokens to obtain a unigram model; the dashed line marks token-distribution entropy. **(d)** Unigram suboptimality: cross-entropy in **(c)** minus that entropy.

Our starting point is the one-layer transformer of Figure 2 trained on Penn Treebank (PTB) (Marcus et al., 1999). As a first step, we freeze the token embeddings and the attention layer and train only the classification head using mini-batch SignSGD and SGD (Figure 3a). This setting stays close to the experiment of Figure 2: the representation is fixed, but the model must still learn the final softmax readout under heavy-tailed language data. Consistent with Kunstner et al. (2024), who compare sign descent with GD in a similar frozen-head setting under full-batch training, SignSGD retains a large advantage over SGD. This already shows that jointly training all layers of the network is not necessary for the gap.

Therefore, in our next step, we simplify the architecture itself. Specifically, removing the attention layer yields a model with a frozen embedding layer and a trainable classification head. This is a *bigram language model*. It eliminates sequence-dependent attention and most architectural complexity. As shown in Figure 3b, the gap between SignSGD and SGD still exists in this bigram setting.

Finally, we collapse the token representation so that every token maps to the same fixed embedding vector. This removes the remaining sequence dependence and leaves a model whose task is simply to fit the marginal token distribution with the softmax activation, i.e., a *unigram model*. Even after this simplification, a clear gap remains in Figure 3c. SGD plateaus above the minimum achievable loss, while SignSGD continues to approach it. The sub-optimality plot in Figure 3d shows this gap directly. This collapsed model therefore isolates the optimizer separation from architectural complexity, and we use it as the basis for the theory that follows.

3.2 Removing Stochasticity: The Deterministic Softmax Unigram Model

After simplifying the architecture, the remaining major complication is mini-batch noise. It can obscure whether the advantage of sign-based methods comes from the stochastic noise or from the underlying deterministic geometry of the loss. Therefore, we remove stochasticity and train the model in the full-batch setting, which amounts to optimizing the following objective with exact gradients.

$$\boxed{f_{\text{uni}}(\theta) = \mathrm{KL}(p \| \operatorname{softmax}(\theta))} \tag{1}$$

where $\mathrm{KL}(p\|q) = \sum_{i=1}^{d} p_i \log \frac{p_i}{q_i}$ is the Kullback–Leibler divergence between two discrete distributions p and q. We call this the *softmax unigram model*. Given a vocabulary of size d, we learn logits $\theta \in \mathbb{R}^d$ to match the target token distribution p, where p_k is the frequency of the k-th token in the dataset. We further simplify this setup by replacing empirical token frequencies with the power law distribution p defined in Definition 3.1, a common approximation to token frequencies in natural language (Piantadosi, 2014)

Definition 3.1 (Zipfian distribution)**.** *The Zipfian distribution p on $[d]$ is defined by $p_k = k^{-1}/\sum_{j=1}^{d} j^{-1}$ for $k \in [d]$, so that $p_k \propto 1/k$.*

As shown in Figure 1, even in the deterministic setting, coordinate-wise methods such as Adam and SignGD significantly outperform GD on the softmax unigram model. To understand why GD

struggles, we inspect the loss for each token. We can express the KL loss in per-coordinate non-negative terms: $f_{\text{uni}}(\theta) = \sum_{i=1}^{d} p_i \left(\log \frac{p_i}{q_i} - 1 + \frac{q_i}{p_i} \right)$ where $q = \text{softmax}(\theta)$. The sum equals the KL because p and q each sum to one, and each term is non-negative because $\log x \geq 1 - 1/x$. Thus each term $\log \frac{p_i}{q_i} - 1 + \frac{q_i}{p_i}$ reflects how well q_i matches the target probability p_i for each coordinate i, which lets us inspect progress on different parts of the distribution. Figure 1(c) and Figure 1(d) reveal that GD makes especially slow progress on the low-probability tail tokens, shown by the yellow curves, whereas SignGD makes much more uniform progress across the distribution. The same qualitative contrast between SGD and SignSGD appears when training the transformer language model in Figure 2(b,c).

The softmax unigram model is convex and smooth, and it simply learns the token frequencies. At the same time, it retains the key optimization difficulty faced by GD under heavy-tailed class imbalance. Furthermore, the coordinates remain coupled through the softmax activation, unlike the simpler separable quadratic models often used in prior analyses. In this sense, the model stays meaningfully connected to the language-modeling experiments above, while being tractable to analyze.

4 Deterministic Softmax Unigram Analysis

In Figure 1, the deterministic experiments already show that GD learns tail tokens much more slowly than SignGD. We formalize this bottleneck through upper and lower bounds on GD. We then show that SignGD avoids this bottleneck and achieves faster convergence, yielding a separation between SignGD and GD. Before proceeding to our analysis, we provide some notation.

Notation. We say a function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth w.r.t. norm $\|\cdot\|$ if for all $x, y \in \mathbb{R}^d$ we have $\|\nabla f(x) - \nabla f(y)\|_\star \leq L\|x - y\|$ where $\|\cdot\|_\star$ denotes the dual norm. Let the smallest such constant be denoted by $L_{\|\cdot\|}(f)$. For a fixed norm $\|\cdot\|$, any $x \in \mathbb{R}^d$, and any nonempty set $\mathcal{S} \subseteq \mathbb{R}^d$, define

$$\text{dist}_{\|\cdot\|}(x, \mathcal{S}) := \inf_{y \in \mathcal{S}} \|x - y\|.$$

When $\mathcal{X}^\star := \arg\min_{z \in \mathbb{R}^d} f(z) \neq \emptyset$, we let $x_\star \in \mathcal{X}^\star$ denote a minimizer of minimum norm, i.e. $\|x_\star\| = \min_{z \in \mathcal{X}^\star} \|z\|$. We write $f^\star := \min_{z \in \mathbb{R}^d} f(z)$ for the minimum value of f. We apply $\text{sign}(\cdot)$ coordinate-wise to vectors, with the convention $\text{sign}(0) = 0$. The updates for GD, SignGD, and normalized GD are given in Algorithms 1 to 3 in the appendix. For a fixed norm $\|\cdot\|$, normalized steepest descent (NSD) chooses $u_t \in \arg\max_{\|y\| \leq 1} \langle \nabla f(x_{t-1}), y \rangle$ and sets $x_t = x_{t-1} - \eta u_t$. SignGD and normalized GD are its ℓ_∞ and ℓ_2 instances, respectively. NSD with weight decay $\lambda > 0$ (NSD-WD) uses the same direction and sets $x_t = (1 - \lambda\eta)x_{t-1} - \eta u_t$.

4.1 Failure of GD: The Tail Learning Bottleneck

The gradient of f_{uni} defined in Equation (1) is $\nabla f_{\text{uni}}(\theta) = \text{softmax}(\theta) - p$. Thus, on tail coordinates where both the target and model probabilities are small, the residual $|\text{softmax}(\theta)_k - p_k|$ is small, and GD makes only small updates to the corresponding logits. However, when starting from uniform initialization, reaching the true tail probabilities requires making large updates to the model tail logits. This mismatch causes GD to make slow progress, which we formalize with the following upper and lower bounds.

Theorem 4.1 (GD upper bound)**.** *Consider optimizing $f_{\text{uni}}(\theta)$ in Equation* (1) *using GD, initialized at $\theta_0 = 0$ with constant step size $\eta = 2$. Then for any $T \geq 1$, it holds that*

$$f_{\text{uni}}(\theta_T) - f^\star_{\text{uni}} \leq \frac{d}{4T}.$$

The upper bound in Theorem 4.1 is linear in d, suggesting slow convergence and a vacuous convergence guarantee when $T \ll d$. This matches the empirical behavior of GD in Figure 1 where $d = 10^4$. The proof of Theorem 4.1 is in Appendix E.4. Next, the following lower bound shows that such slow convergence is unavoidable for GD.

Theorem 4.2 (GD lower bound)**.** *There exist absolute constants $c, C > 0$ such that the following holds for all sufficiently large d. Suppose we run GD initialized at $\theta_0 = 0$ with a constant step size*

$0 < \eta < d/2$ *on* $f_{\mathrm{uni}}(\theta)$ *in Equation* (1)*. Then, whenever* $C \log d \le \eta T \le \frac{d \log d}{C}$*, it holds that*

$$f_{\mathrm{uni}}(\theta_T) - f^\star_{\mathrm{uni}} \ge \frac{c}{\log d} \log\left(\frac{d \log d}{\eta T}\right).$$

Consequently, there exist $\epsilon_0 \in (0,1)$ *such that for any* $\epsilon \in (0, \epsilon_0)$*, it holds for all sufficiently large* d *that the number of iterations required for GD to reach* ϵ*-suboptimality is lower bounded by*

$$\min\{T : f_{\mathrm{uni}}(\theta_T) - f^\star_{\mathrm{uni}} \le \epsilon\} \ge \frac{1}{\eta} d^{1-\epsilon} \log d.$$

We remark that this lower bound covers a broad range of learning rates $\eta \in (0, d/2)$. Indeed, one can show that when using a learning rate $\eta \gtrsim \log d$, GD would become unstable near a minimizer and hence does not converge; see Theorem F.4 for details. The proof of Theorem 4.2 is in Appendix F.1. The key idea is to track the marginal distribution for coordinate pairs $(i, 2i)$. For each such pair, the target logit gap is $\log(p_i/p_{2i}) = \log 2$, while the initial logit gap is 0 at the initialization $\theta_0 = 0$; meanwhile, GD can only grow the logit gap very slowly on tail coordinates because the corresponding updates are themselves tiny. As a result, many tail pairs remain far from their target relative probabilities for polynomially many iterations. Converting these unresolved tail errors into the KL loss yields the lower bound in Theorem 4.2.

4.2 Success of SignGD

The SignGD update for the softmax unigram objective is $\theta_t = \theta_{t-1} - \eta \,\mathrm{sign}(\mathrm{softmax}(\theta_{t-1}) - p)$. So every coordinate with a nonzero gradient entry can move at a comparable speed even when the original magnitude $|[\mathrm{softmax}(\theta)]_k - p_k|$ is small. In this sense, SignGD can move tail coordinates efficiently even when their gradient entries are small.

We derive the deterministic convergence rate of SignGD from a generic guarantee for NSD, defined above. For comparison, we also derive convergence guarantees for normalized GD, illustrating the difference between the ℓ_∞ and ℓ_2 geometries. The following theorem gives a generic convergence guarantee for NSD under a bound on the distance of the iterates to the minimizers.

Theorem 4.3 (Conditional convergence of NSD)**.** *Suppose* $f : \mathbb{R}^d \to \mathbb{R}$ *is convex and* L*-smooth with respect to a norm* $\|\cdot\|$*, with* $\mathcal{X}^\star \neq \emptyset$*. Fix a radius* $R > 0$*. Suppose we run NSD for this norm* $\|\cdot\|$ *with horizon* $T \ge 3$ *and learning rate* $\eta = \frac{R \log T}{T}$*. If the iterates satisfy* $\mathrm{dist}_{\|\cdot\|}(x_{t-1}, \mathcal{X}^\star) \le R$ *for* $1 \le t \le T$*, then*

$$f(x_T) - f^\star \le \frac{LR^2 \log T}{T}.$$

The proof is in Appendix C.1. To apply this conditional guarantee to obtain the convergence rate of NSD, the main technical obstacle is to establish such a uniform distance bound. This condition is not automatically guaranteed for general functions, but for the softmax unigram model, we can exploit the structure of the softmax function and the exact gradient to show that the iterates of both SignGD and normalized GD remain at a uniformly bounded distance from the minimizer set.

For the initialization and learning rates used below, Appendices C.2 and C.4 establish the uniform bounds $\mathrm{dist}_{\|\cdot\|_\infty}(\theta_t, \arg\min f_{\mathrm{uni}}) \le \frac{\log d}{2}$ for SignGD and $\mathrm{dist}_{\|\cdot\|_2}(\theta_t, \arg\min f_{\mathrm{uni}}) \le 2\sqrt{d}$ for normalized GD. To obtain the explicit convergence rates of SignGD and normalized GD, it then remains to determine the smoothness constants of f_{uni} with respect to the ℓ_∞ and ℓ_2 norms.

For a generic function, the ℓ_∞-smoothness constant can be up to d times larger than the ℓ_2-smoothness constant. For the softmax unigram model, however, Lemma B.2 gives $L_{\|\cdot\|_\infty}(f_{\mathrm{uni}}) = 1$ and $L_{\|\cdot\|_2}(f_{\mathrm{uni}}) = 1/2$, so the two constants differ only by a factor of two. Therefore, the different dimensional dependence of the two upper bounds instead comes from how far the iterates can be from the minimizer set: the bounds are $\mathcal{O}(\sqrt{d})$ in the ℓ_2 norm and $\mathcal{O}(\log d)$ in the ℓ_∞ norm. Substituting these smoothness constants and distance bounds into Theorem 4.3 gives the following convergence rates for SignGD and normalized GD.

Corollary 4.4 (SignGD upper bound)**.** *Suppose we run SignGD on* f_{uni} *initialized at* $\theta_0 = 0$ *with horizon* $T \ge 3$ *and learning rate* $\eta = \frac{\log d \log T}{2T}$*. Then*

$$f_{\mathrm{uni}}(\theta_T) - f^\star_{\mathrm{uni}} \le \frac{(\log d)^2 \log T}{4T}.$$

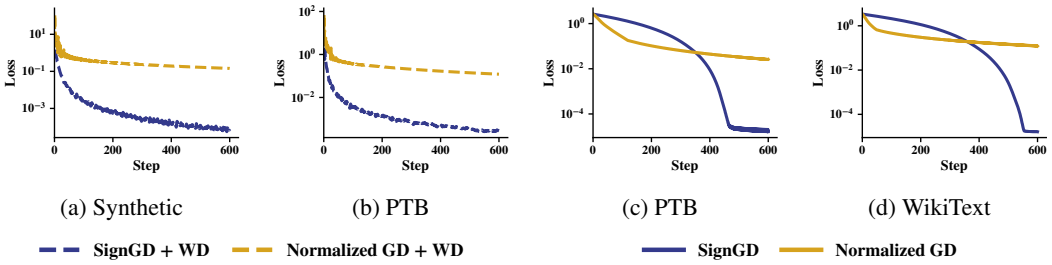


Figure 4: **(a,b)** Weight-decay runs on a synthetic distribution and the PTB dataset, using the learning rates prescribed by Corollary D.2 and Corollary D.3. **(c,d)** Best-performing runs with non-weight decay optimizers on the PTB and WikiText-2 datasets.

Corollary 4.5 (Normalized GD upper bound)**.** *Suppose we run normalized GD on* f_{uni} *with* $\theta_0 = 0$*, horizon* $T \geq 3$*, and fixed learning rate* $\eta = \frac{2\sqrt{d}\,\log T}{T}$*. Then*

$$f_{\text{uni}}(\theta_T) - f^\star_{\text{uni}} \leq \frac{2d \log T}{T}.$$

The SignGD upper bound scales as $\widetilde{\mathcal{O}}((\log d)^2/T)$, whereas the normalized-GD upper bound scales as $\widetilde{\mathcal{O}}(d/T)$. Thus, the SignGD upper bound is smaller by a factor of order $d/(\log d)^2$. Furthermore, SignGD reaches suboptimality below ϵ in $\widetilde{\mathcal{O}}((\log d)^2/\epsilon)$ iterations, while Theorem 4.2 shows that GD requires $\text{poly}(d)$ many iterations, specifcally, $\Omega(d^{1-\epsilon}\log d)$ for every fixed sufficiently small ϵ. This gives a strict separation between SignGD and GD on the softmax unigram model.

Additionally, we analyze SignGD and normalized GD with weight decay in Appendix D. Weight decay contracts the iterates and directly bounds their norms, eliminating the need for a separate bound on their distance to the minimizer set. For each norm, the convergence bound depends on its smoothness constant and the corresponding minimum-norm minimizer $\|\theta_\star\|$. This yields the same $\widetilde{\mathcal{O}}((\log d)^2/T)$ and $\widetilde{\mathcal{O}}(d/T)$ upper bounds for SignGD and normalized GD, respectively. Similarly, the difference in dimensional dependence comes from the minimizer norm, which is $\Theta(\log d)$ in ℓ_∞ and $\Theta(\sqrt{d})$ in ℓ_2. We verify our theory on synthetic and real-data token distributions in Figure 4.

5 Stochastic Softmax Unigram Analysis

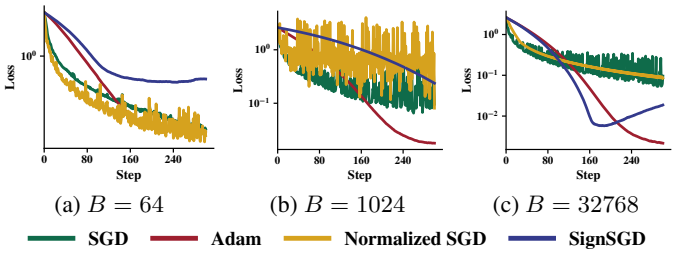
In Section 4, we worked in the deterministic setting to isolate the core mechanism that separates SignGD from GD. Here, we reintroduce mini-batch noise and study how it interacts with the deterministic loss geometry. We find that stochastic noise does not amplify the advantage of sign-based methods in the deterministic setting; rather, the separation becomes visible once the noise is sufficiently small. We formalize this for the softmax unigram model by providing an upper bound for SignSGD with weight decay and proving a separation from full-batch GD when the batch size for SignSGD is large enough.

5.1 SignSGD under Mini-Batch Noise

At iteration t, draw a mini-batch of tokens $\xi_{t,1}, \ldots, \xi_{t,B} \overset{\text{iid}}{\sim} \text{Cat}(p)$ and use the stochastic gradient:

$$g_t := \nabla \widehat{f}_{\text{uni},B}(\theta_t; \xi_{t,1:B}) = \text{softmax}(\theta_t) - \hat{p}_t, \qquad \hat{p}_t := \frac{1}{B}\sum_{b=1}^{B} e_{\xi_{t,b}}. \tag{2}$$

This is the gradient of the mini-batch cross-entropy loss, whose expectation differs from the KL objective only by a constant entropy term. Since $\mathbb{E}[\hat{p}_t] = p$, the mini-batch gradient $g_t = \text{softmax}(\theta_t) - \hat{p}_t$ is also an unbiased stochastic gradient of f_{uni}. The noise $\varepsilon_t = g_t - \nabla f_{\text{uni}}(\theta_t) = p - \hat{p}_t$ depends only on the sampled mini-batch and is independent of θ_t.



(a) $B = 64$ (b) $B = 1024$ (c) $B = 32768$

Figure 5: Effect of batch size on the softmax unigram model ($d = 10^4$, Zipfian p). The same qualitative effect appears in the transformer experiment (Figure 7).

Now we state a generic NSD-WD convergence bound for convex, smooth objectives in the stochastic setting. We instantiate it below for SignSGD with weight decay and later in Section 5.2 for normalized SGD with weight decay. The proof is given in Appendix G.

Theorem 5.1 (Stochastic NSD-WD). *Suppose f is convex and L-smooth with respect to a norm $\|\cdot\|$. Suppose the stochastic gradient g_t satisfies $\mathbb{E}_t[g_t] = \nabla f(x_t)$ and $\mathbb{E}_t[\|g_t - \nabla f(x_t)\|_\star] \le \sigma$. For a horizon $T \ge 2$, run stochastic NSD-WD with weight decay factor $\lambda \le 1/\|x_\star\|$ and fixed learning rate $\eta = \frac{\log T}{\lambda T}$. Then*

$$\mathbb{E}[f(x_T) - f^\star] \le \frac{2L(1+\lambda D)^2 \log T}{\lambda^2 T} + \frac{2\sigma}{\lambda},$$

where $D = \max\{\|x_0\|, 1/\lambda\}$. In particular, choosing $x_0 = 0$ and $\lambda = 1/\|x_\star\|$ yields

$$\mathbb{E}[f(x_T) - f^\star] \le \frac{8L\|x_\star\|^2 \log T}{T} + 2\|x_\star\|\sigma.$$

When $\sigma = 0$, Theorem 5.1 recovers the deterministic rate in Theorem D.1. Its stochastic term is a constant error floor proportional to σ, so achieving small error requires sufficiently large batch size. For the softmax unigram gradient, Lemma B.4 gives

$$\mathbb{E}\|g_t - \nabla f_{\text{uni}}(\theta_t)\|_2 \le \frac{1}{\sqrt{B}}, \qquad \mathbb{E}\|g_t - \nabla f_{\text{uni}}(\theta_t)\|_1 \le \min\left\{2, \frac{\|\sqrt{p}\|_1}{\sqrt{B}}\right\},$$

where $\sqrt{\cdot}$ is applied element-wise. Since the noise is independent of θ_t, the bounds above are derived by bounding the distance from the estimator $\hat{p}_t$ to the true p in each norm. We now use Theorem 5.1 and the ℓ_1 stochastic-gradient error bound to obtain an upper bound for SignSGD with weight decay.

Corollary 5.2 (SignSGD with weight decay). *For all sufficiently large d, consider applying SignSGD to f_{uni} using the mini-batch gradient g_t defined in Equation (2), with initialization $\theta_0 = 0$, batch size B, horizon $T \ge 2$, weight-decay parameter $\lambda = 2/\log d$, and fixed learning rate $\eta = \frac{\log d \log T}{2T}$. Then*

$$\mathbb{E}[f_{\text{uni}}(\theta_T) - f^\star_{\text{uni}}] \le \frac{2(\log d)^2 \log T}{T} + \underbrace{\min\left\{2\log d, 4\sqrt{\frac{d\log d}{B}}\right\}}_{=:N_B(d)}.$$

The $N_B(d)$ term in Corollary 5.2 depends only on the batch size and acts as an error floor in this guarantee. When $B < d$, we have $N_B(d) \ge \min\{2\log d, 4\sqrt{\log d}\}$, so the bound is essentially vacuous in this regime. For example, when $B = 1$, the error floor $N_1(d) = 2\log d$ exceeds the initialization loss $0.5\log d - \log\log d + \mathcal{O}(1)$ and therefore does not certify progress from initialization. Our experiments in Figure 5 show that SignSGD attains higher loss than SGD and normalized SGD at $B = 2^6$ and $B = 2^{10}$, but lower loss at a larger batch size $B = 2^{15} > d$. At sufficiently large batch sizes, SignSGD's deterministic advantage becomes visible. We therefore compare its iteration complexity directly with full-batch GD, which has exact gradients.

Corollary 5.3 (SignSGD complexity). *Let $\theta_{T,B}$ denote the T-th SignSGD iterate from Corollary 5.2 with batch size B. For every fixed $\epsilon > 0$ and all sufficiently large d, for any sufficiently large batch size satisfying $B \ge \frac{64 d \log d}{\epsilon^2}$, the number of iterations required to reach ϵ-accuracy is bounded by*

$$\min\{T : \mathbb{E}[f_{\text{uni}}(\theta_{T,B}) - f^\star_{\text{uni}}] \le \epsilon\} \le \left\lceil \frac{8(\log d)^2}{\epsilon} \log\left(\frac{8(\log d)^2}{\epsilon}\right)\right\rceil.$$

Corollary 5.3 shows that, with a sufficiently large batch size, SignSGD with weight decay reaches fixed accuracy in $\mathcal{O}((\log d)^2 \log\log d)$ iterations. In contrast, Theorem 4.2 shows that full-batch GD requires $\Omega(d^{1-\epsilon}\log d)$ iterations for every fixed sufficiently small ϵ. Together, these results give an iteration-complexity separation between large-batch SignSGD with weight decay and full-batch GD, despite the latter having access to exact gradients.

5.2 Batch-Size Tradeoffs under a Fixed Token Budget

The previous subsection showed that a sufficiently large batch size can reveal SignSGD's deterministic advantage without constraining the total number of sampled tokens. We now fix the token budget $N = BT$ and compare SignSGD with normalized SGD and vanilla SGD under this constraint. Recall Theorem 5.1 for generic stochastic NSD-WD which we now instantiate for normalized SGD with weight decay, i.e., NSD-WD under the ℓ_2 norm.

Corollary 5.4 (Normalized SGD with weight decay)**.** *Consider applying normalized SGD to f_{uni} using the mini-batch gradient g_t defined in Equation* (2)*, with initialization $\theta_0 = 0$, batch size B, horizon $T \geq 2$, weight-decay parameter $\lambda = 1/\sqrt{d}$, and fixed learning rate $\eta = \frac{\sqrt{d}\,\log T}{T}$. Then*

$$\mathbb{E}[f_{\text{uni}}(\theta_T) - f^\star_{\text{uni}}] \leq \frac{4d\log T}{T} + 2\sqrt{\frac{d}{B}}.$$

The stochastic terms in Corollaries 5.2 and 5.4 are both of order $\mathcal{O}(B^{-1/2})$. Their deterministic terms are $2(\log d)^2 \log T/T$ and $4d\log T/T$, respectively, preserving the same dimensional separation. Comparison between these two upper bounds suggests that *stochastic noise only determines how large the batch size B must be before this separation becomes visible*. We next fix the total sample budget $N = BT$ and optimize over the choice of batch size. This yields the batch-size choices and resulting bounds in Corollaries 5.5 and 5.6.

Corollary 5.5 (Tuned Batch Size for SignSGD w/ WD)**.** *Fix a token budget $N = BT$. For all sufficiently large d, if $N \geq d$, and $B = \widetilde{\Theta}\Big(\frac{(dN^2)^{1/3}}{\log d}\Big)$, then the iterates in Corollary 5.2 satisfy,*

$$\mathbb{E}[f_{\text{uni}}(\theta_T) - f^\star_{\text{uni}}] = \widetilde{\mathcal{O}}\left(\frac{d^{1/3}\log d}{N^{1/3}}\right).$$

Corollary 5.6 (Tuned Batch Size for normalized SGD w/ WD)**.** *Fix a token budget $N = BT$. For all sufficiently large d, if $N \geq 4\sqrt{d}\log N$, and $B = \widetilde{\Theta}\Big(\frac{N^{2/3}}{d^{1/3}}\Big)$, then the iterates in Corollary 5.4 satisfy,*

$$\mathbb{E}[f_{\text{uni}}(\theta_T) - f^\star_{\text{uni}}] = \widetilde{\mathcal{O}}\left(\frac{d^{2/3}}{N^{1/3}}\right).$$

Under a fixed sample budget, the resulting SignSGD bound is smaller than the normalized SGD bound by a factor of order $d^{1/3}/\log d$. The corresponding optimal batch size for SignSGD is larger by a factor of $\Theta(d^{2/3}/\log d)$. Thus, SignSGD uses a larger batch size to reveal its deterministic advantage. Appendix J gives analogous rates for the momentum variants and yields the same separation.

In Appendix H, we compare SignSGD with weight decay against SGD under a fixed token budget. When the batch size is independently tuned for each method, SGD's upper bound is minimized at $B = 1$ and is smaller than the tuned SignSGD bound for sufficiently large N/d. However, comparing the bounds at a fixed batch size, increasing B suppresses SignSGD's stochastic term, allowing its better deterministic dependence on d to yield a smaller upper bound.

6 Conclusion and Future Direction

The optimizer gap seen in transformer language modeling persists in the softmax unigram model, where we identify a tail learning bottleneck for GD under heavy-tailed class imbalance: logits for rare tokens receive tiny updates despite target log odds of order one. SignGD avoids this dependence on gradient magnitude and achieves faster deterministic rates. Mini-batch noise does not strengthen this advantage; it determines when the advantage becomes visible. We provide preliminary results on Adam in Appendix K. Extending to transformers with learned embeddings remains open, as does proving a lower bound for normalized GD on the softmax unigram model. The latter is challenging because global normalization couples coordinates and can repeatedly reverse progress on tail ratios.

AI USE STATEMENT

We used generative AI tools to assist with running numerical experiments and to identify potentially relevant literature. We also used them to improve the clarity of the writing and to help with LaTeX formatting, figure placement, and appendix organization.We reviewed the AI-assisted material, including suggested references and experimental outputs, before incorporating it into the paper. We take responsibility for the final text, claims, figures, and reported results.

REFERENCES

Lukas Balles and Philipp Hennig. Dissecting Adam: The Sign, Magnitude and Variance of Stochastic Gradients. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 404–413. PMLR, July 2018. URL https://proceedings.mlr.press/v80/balles18a.html. ISSN: 2640-3498.

Jeremy Bernstein and Laker Newhouse. Old Optimizer, New Norm: An Anthology, December 2024. URL http://arxiv.org/abs/2409.20325. arXiv:2409.20325 [cs].

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed Optimisation for Non-Convex Problems. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 560–569. PMLR, July 2018. URL https://proceedings.mlr.press/v80/bernstein18a.html. ISSN: 2640-3498.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2260–2268. PMLR, July 2020. URL https://proceedings.mlr.press/v119/cutkosky20b.html.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL https://www.jmlr.org/papers/v12/duchi11a.html.

Athanasios Glentis, Dawei Li, Chung-Yiu Yau, and Mingyi Hong. Revisiting the Adam–SGD gap in LLM pre-training: The role of large effective learning rates, 2026. URL https://arxiv.org/abs/2605.17787.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned Stochastic Tensor Optimization. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1842–1850. PMLR, July 2018. URL https://proceedings.mlr.press/v80/gupta18a.html. ISSN: 2640-3498.

Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1263–1271. PMLR, June 2016. URL https://proceedings.mlr.press/v48/hazana16.html.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 427–435. PMLR, June 2013. URL https://proceedings.mlr.press/v28/jaggi13.html.

Ruichen Jiang, Devyani Maladkar, and Aryan Mokhtari. Provable Complexity Improvement of AdaGrad over SGD: Upper and Lower Bounds in Stochastic Non-Convex Optimization, June 2025. URL http://arxiv.org/abs/2406.04592. arXiv:2406.04592 [math].

Jonathan A. Kelner, Yin Tat Lee, Lorenzo Orecchia, and Aaron Sidford. An almost-linear-time algorithm for approximate max flow in undirected graphs, and its multicommodity generalizations. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 217–226, 2014. URL https://arxiv.org/abs/1304.2338.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. URL http://arxiv.org/abs/1412.6980. arXiv:1412.6980 [cs].

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization, March 2025. URL https://arxiv.org/abs/2503.12645. arXiv:2503.12645 [cs.LG].

Dmitry Kovalev. SGD with adaptive preconditioning: Unified analysis and momentum acceleration. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=XhXMzPJJ7J.

Frederik Kunstner and Francis Bach. Scaling Laws for Gradient Descent and Sign Descent for Linear Bigram Models under Zipf's Law, May 2025. URL http://arxiv.org/abs/2505.19227. arXiv:2505.19227 [cs].

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise Is Not the Main Factor Behind the Gap Between SGD and Adam on Transformers, but Sign Descent Might Be. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=a65YK0cqH8g.

Frederik Kunstner, Alan Milligan, Robin Yadav, Mark Schmidt, and Alberto Bietti. Heavy-Tailed Class Imbalance and Why Adam Outperforms Gradient Descent on Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 30106–30148, November 2024. doi: 10.52202/079017-0948. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/350e718ff74062b4bac2c6ffd9e1ac66-Abstract-Conference.html.

Bingcong Li, Alireza Sadeghi, and Georgios B. Giannakis. Heavy ball momentum for conditional gradient. In *Advances in Neural Information Processing Systems 34*, pp. 21244–21255, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/b166b57d195370cd41f80dd29ed523d9-Abstract.html.

Yuxing Liu, Rui Pan, and Tong Zhang. AdaGrad under Anisotropic Smoothness. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4GT9uTsAJE.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, January 2019. URL http://arxiv.org/abs/1711.05101. arXiv:1711.05101 [cs].

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. Treebank-3, 1999. LDC Catalog No. LDC99T42.

Ruslan Nazykov, Aleksandr Shestakov, Vladimir Solodkin, Aleksandr Beznosikov, Gauthier Gidel, and Alexander Gasnikov. Stochastic Frank-Wolfe: Unified analysis and zoo of special cases. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4870–4878. PMLR, May 2024. URL https://proceedings.mlr.press/v238/nazykov24a.html.

Yan Pan and Yuanzhi Li. Toward Understanding Why Adam Converges Faster Than SGD for Transformers, May 2023. URL https://arxiv.org/abs/2306.00204v1.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49069–49104. PMLR, July 2025. URL https://proceedings.mlr.press/v267/pethick25a.html.

Steven T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130, October 2014. ISSN 1531-5320. doi: 10.3758/s13423-014-0585-6. URL https://doi.org/10.3758/s13423-014-0585-6.

Teodora Srećković, Jonas Geiping, and Antonio Orvieto. Is your batch size the problem? revisiting the Adam-SGD gap in language modeling, June 2025. URL https://arxiv.org/abs/2506.12543. arXiv:2506.12543 [cs.LG].

Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(219):1–30, 2020. ISSN 1533-7928. URL http://jmlr.org/papers/v21/18-352.html.

Shuo Xie and Zhiyuan Li. Implicit bias of AdamW: ℓ_∞-norm constrained optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 54488–54510, Vienna, Austria, July 2024. JMLR.org.

Shuo Xie, Mohamad Amin Mohamadi, and Zhiyuan Li. Adam Exploits ℓ_∞-geometry of Loss Landscape via Coordinate-wise Adaptivity. In *International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=PUnD86UEK5.

Shuo Xie, Tianhao Wang, Sashank Reddi, Sanjiv Kumar, and Zhiyuan Li. Structured Preconditioners in Adaptive Optimization: A Unified Analysis, July 2025b. URL http://arxiv.org/abs/2503.10537. arXiv:2503.10537 [cs].

Shuo Xie, Tianhao Wang, Beining Wu, and Zhiyuan Li. A tale of two geometries: Adaptive optimizers and Non-Euclidean descent. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=iaoAKDRAJQ.

Dingzhi Yu, Hongyi Tao, Yuanyu Wan, Luo Luo, and Lijun Zhang. Sign-based optimizers are effective under heavy-tailed noise, February 2026. URL https://arxiv.org/abs/2602.07425. arXiv:2602.07425 [cs.LG].

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are Adaptive Methods Good for Attention Models? In *Advances in Neural Information Processing Systems*, volume 33, pp. 15383–15393. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/b05b57f6add810d3b7490866d74c0053-Abstract.html.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why Transformers Need Adam: A Hessian Perspective. *Advances in Neural Information Processing Systems*, 37:131786–131823, December 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ee0e45ff4de76cbfdf07015a7839f339-Abstract-Conference.html.

CONTENTS

1 Introduction **1**

2 Related Work **2**

3 From Transformers to the Softmax Unigram Model **3**

3.1 Progressive Architectural Simplification 4

3.2 Removing Stochasticity: The Deterministic Softmax Unigram Model 4

4 Deterministic Softmax Unigram Analysis **5**

4.1 Failure of GD: The Tail Learning Bottleneck 5

4.2 Success of SignGD 6

5 Stochastic Softmax Unigram Analysis **7**

5.1 SignSGD under Mini-Batch Noise 7

5.2 Batch-Size Tradeoffs under a Fixed Token Budget 9

6 Conclusion and Future Direction **9**

A Algorithm Definitions **14**

B Softmax-Unigram Constants **14**

C Proofs without Weight Decay **19**

C.1 Proof of the conditional NSD convergence theorem 24

C.2 SignGD distance bound and proof 25

C.3 Proof of the SignGD upper bound 25

C.4 Normalized-GD distance bound and proof 25

C.5 Proof of the normalized-GD upper bound 26

D Deterministic Analysis with Weight Decay **26**

E Proofs of Deterministic Rates **27**

E.1 Proof of the horizon-tuned NSD-WD theorem 27

E.2 Proof of the deterministic SignGD theorem 27

E.3 Proof of the deterministic normalized-GD theorem 28

E.4 Proof of the GD upper bound 28

F Proof of the Deterministic GD Lower Bound **29**

F.1 Proof of the GD lower bound 31

F.2 GD with Large Step Sizes is unstable 32

G Generic Stochastic (Momentum) NSD-WD Proofs **33**

G.1 Stochastic NSD-WD Proofs 35
G.2 Stochastic Momentum NSD-WD Proofs 35

H Comparison with SGD **39**
H.1 Fixed-batch SignGD vs. SGD 40

I Stochastic Softmax Unigram Analysis Proofs **41**

J Stochastic Momentum Softmax Unigram Analysis **43**

K Challenges in Extending the Analysis to Adam **45**
K.1 Proof of the Diagonal Adaptive Smoothness Bound 45

L Experimental Details for Section 3 **46**
L.1 Batch-Size Experiments 48

A Algorithm Definitions

For completeness, we give the GD, SignGD, and normalized-GD updates used in the main text.

Algorithm 1 Gradient Descent (GD)

Input: Initial point x_0, learning rate η, objective f, horizon T
for $t = 1$ to T **do**
  $g_t \leftarrow \nabla f(x_{t-1})$
  $x_t \leftarrow x_{t-1} - \eta g_t$
end for

Algorithm 2 Sign Gradient Descent (SignGD)

Input: Initial point x_0, learning rate η, objective f, horizon T
for $t = 1$ to T **do**
  $g_t \leftarrow \nabla f(x_{t-1})$
  $x_t \leftarrow x_{t-1} - \eta \operatorname{sign}(g_t)$
end for

Algorithm 3 Normalized Gradient Descent

Input: Initial point x_0, learning rate η, objective f, horizon T
for $t = 1$ to T **do**
  $g_t \leftarrow \nabla f(x_{t-1})$
  $u_t \leftarrow g_t / \|g_t\|_2$
  $x_t \leftarrow x_{t-1} - \eta u_t$
end for

B Softmax-Unigram Constants

Lemma B.1 (Initial Loss)**.** *For the softmax unigram objective initialized at $\theta_0 = 0$,*

$$f_{\mathrm{uni}}(0) - f^\star_{\mathrm{uni}} = \mathrm{KL}\left(p \middle\| \frac{\mathbf{1}}{d}\right) \le \log d.$$

For the Zipfian target in Definition 3.1, more precisely,

$$f_{\mathrm{uni}}(0) - f^\star_{\mathrm{uni}} = \frac{1}{2}\log d - \log\log d + \mathcal{O}(1).$$

Proof. At $\theta_0 = 0$, the softmax distribution is uniform:

$$\mathrm{softmax}(\theta_0) = u := \frac{1}{d}\mathbf{1}.$$

Since $f^\star_{\mathrm{uni}} = 0$,

$$f_{\mathrm{uni}}(0) - f^\star_{\mathrm{uni}} = \mathrm{KL}(p\|u) = \log d - H(p) \le \log d.$$

For the Zipfian target $p_i = 1/(iH_d)$, we further have

$$\mathrm{KL}(p\|u) = \sum_{i=1}^{d} p_i \log(dp_i) = \log d - \log H_d - \frac{1}{H_d}\sum_{i=1}^{d}\frac{\log i}{i}.$$

The harmonic number satisfies

$$H_d = \log d + \mathcal{O}(1), \qquad \log H_d = \log\log d + \mathcal{O}\left(\frac{1}{\log d}\right).$$

Moreover, by an integral comparison,

$$\sum_{i=1}^{d}\frac{\log i}{i} = \int_1^d \frac{\log x}{x}\,dx + \mathcal{O}(1) = \frac{1}{2}(\log d)^2 + \mathcal{O}(1).$$

Therefore,

$$\frac{1}{H_d}\sum_{i=1}^{d}\frac{\log i}{i} = \frac{1}{2}\log d + \mathcal{O}(1).$$

Substituting these estimates gives

$$f_{\mathrm{uni}}(0) - f^\star_{\mathrm{uni}} = \frac{1}{2}\log d - \log\log d + \mathcal{O}(1),$$

as claimed. □

Lemma B.2 (Smoothness constants)**.** *We have*

$$L_{\|\cdot\|_2}(f_{\mathrm{uni}}) = \frac{1}{2}, \qquad L_{\|\cdot\|_\infty}(f_{\mathrm{uni}}) = 1.$$

Proof. For any $\theta \in \mathbb{R}^d$,

$$\nabla f_{\mathrm{uni}}(\theta) = \mathrm{softmax}(\theta) - p, \text{ and } \quad \nabla^2 f_{\mathrm{uni}}(\theta) = \mathrm{diag}(\mathrm{softmax}(\theta)) - \mathrm{softmax}(\theta)\,\mathrm{softmax}(\theta)^T.$$

Therefore, for any $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$,

$$\begin{aligned} u^\top \nabla^2 f_{\mathrm{uni}}(\theta) u &= \sum_{k=1}^{d}[\mathrm{softmax}(\theta)]_k u_k^2 - \left(\sum_{k=1}^{d}[\mathrm{softmax}(\theta)]_k u_k\right)^2 \\ &= \mathrm{Var}_{I\sim\mathrm{softmax}(\theta)}[u_I]. \end{aligned}$$

By Popoviciu's inequality,

$$\mathrm{Var}_{I\sim\mathrm{softmax}(\theta)}[u_I] \le \frac{(\max_i u_i - \min_i u_i)^2}{4}.$$

Also, if $\|u\|_2 = 1$, then

$$\max_i u_i - \min_i u_i \le \sqrt{2},$$

because the difference of two coordinates is maximized by concentrating the ℓ_2 mass on those two coordinates with opposite signs. Therefore,

$$u^\top \nabla^2 f_{\mathrm{uni}}(\theta) u \le \frac{(\sqrt{2})^2}{4} = \frac{1}{2},$$

and taking the supremum over θ and $\|u\|_2 = 1$ gives

$$L_{\|\cdot\|_2}(f_{\mathrm{uni}}) \le \frac{1}{2}.$$

Similarly, for $u \in \mathbb{R}^d$ with $\|u\|_\infty = 1$,

$$u^\top \nabla^2 f_{\text{uni}}(\theta) u \le \sum_{k=1}^{d} [\text{softmax}(\theta)]_k u_k^2$$
$$\le \sum_{k=1}^{d} [\text{softmax}(\theta)]_k |u_k| \le \|\text{softmax}(\theta)\|_1 \|u\|_\infty = 1.$$

Hence $L_{\|\cdot\|_\infty}(f_{\text{uni}}) \le 1$.

LOWER BOUNDS

Next, we compute the lower bounds on $L_{\|\cdot\|_2}(f_{\text{uni}})$ and $L_{\|\cdot\|_\infty}(f_{\text{uni}})$. Consider $\theta(t) : \mathbb{R} \to \mathbb{R}^d$ as follows $\theta(t) = [t\ t\ -t\ \cdots\ -t]$. Then define

$$s(t) := [\text{softmax}(\theta(t))]_1 = [\text{softmax}(\theta(t))]_2 = \frac{e^t}{2e^t + (d-2)e^{-t}}.$$

For any $r \in \mathbb{R}^2$ with $\|r\|_2 = 1$ define $u = [r_1\ r_2\ 0\ \cdots\ 0] \in \mathbb{R}^d$. Let

$$\mathbf{A}(t) = \begin{bmatrix} s(t)(1-s(t)) & -s(t)^2 \\ -s(t)^2 & s(t)(1-s(t)) \end{bmatrix}.$$

Then $u^T \nabla^2 f_{\text{uni}}(\theta(t)) u = r^T \mathbf{A}(t) r$. Since r is arbitrary,

$$\sup_{\|u\|_2=1} u^T \nabla^2 f_{\text{uni}}(\theta(t)) u \ge \sup_{\|r\|_2=1} r^T \mathbf{A}(t) r.$$

The largest eigenvalue of $\mathbf{A}(t)$ is $s(t)$, and $\lim_{t\to\infty} s(t) = 1/2$. Therefore,

$$L_{\|\cdot\|_2}(f_{\text{uni}}) \ge \lim_{t\to\infty} \sup_{\|u\|_2=1} u^T \nabla^2 f_{\text{uni}}(\theta(t)) u$$
$$\ge \lim_{t\to\infty} \sup_{\|r\|_2=1} r^T \mathbf{A}(t) r = \frac{1}{2}.$$

For the ℓ_∞ lower bound, take $r \in \mathbb{R}^2$ with $\|r\|_\infty = 1$. Then again $u^\top \nabla^2 f_{\text{uni}}(\theta(t)) u = r^\top \mathbf{A}(t) r$, and the supremum is achieved at $\bar{r} = [1, -1]$, which gives $\bar{r}^\top \mathbf{A}(t) \bar{r} = 2s(t)$. Hence

$$L_{\|\cdot\|_\infty}(f_{\text{uni}}) \ge \lim_{t\to\infty} \sup_{\|u\|_\infty=1} u^T \nabla^2 f_{\text{uni}}(\theta(t)) u$$
$$\ge \lim_{t\to\infty} \sup_{\|r\|_\infty=1} r^T \mathbf{A}(t) r = 1.$$

Together with the upper bounds above, this proves $L_{\|\cdot\|_2}(f_{\text{uni}}) = 1/2$ and $L_{\|\cdot\|_\infty}(f_{\text{uni}}) = 1$. □

Lemma B.3 (Minimal-norm minimizers). *Among all minimizers $\theta_\star \in \arg\min f_{\text{uni}}$, we have*

$$\min_{\theta_\star \in \arg\min f_{\text{uni}}} \|\theta_\star\|_\infty = \frac{1}{2} \log d,$$
$$\min_{\theta_\star \in \arg\min f_{\text{uni}}} \|\theta_\star\|_2 = \sqrt{d\, \text{Var}_{k\sim\text{Unif}[d]}[\log k]}.$$

Moreover, for every $d \ge 2$,

$$\text{Var}_{k\sim\text{Unif}[d]}[\log k] \le 1 - \left(\frac{1+\log d}{d}\right)^2 < 1, \qquad \textit{so} \qquad \min_{\theta_\star \in \arg\min f_{\text{uni}}} \|\theta_\star\|_2 \le \sqrt{d}.$$

As $d \to \infty$, we also have

$$\text{Var}_{k\sim\text{Unif}[d]}[\log k] \to 1.$$

Proof. Of course, for a given $\theta \in \arg\min f_{\text{uni}}$ we must have that $\text{softmax}(\theta) = p$. We note that the softmax map is invariant up to a constant shift. Since $p_k = k^{-1}/\sum_i i^{-1}$, the set of optimal solutions is described by

$$\theta_k = -\log(k) + c, \qquad k \in [d],$$

where $c \in \mathbb{R}$ is a fixed constant.

To minimize $\|\theta_\star\|_\infty$, observe that $\|\theta_\star\|_\infty = \max\{|c|, |-\log(d) + c|\}$, so the optimum is attained at the midpoint $c = \frac{\log(d)}{2}$. Thus,

$$\min_{\theta_\star \in \arg\min f_{\text{uni}}} \|\theta_\star\|_\infty = \frac{\log(d)}{2}.$$

To minimize $\|\theta_\star\|_2$, define

$$\gamma(c) = \sum_{k=1}^{d} (-\log(k) + c)^2 = \|\theta_\star\|_2^2.$$

Then

$$\gamma'(c) = 2\sum_{k=1}^{d} (c - \log(k)), \qquad \gamma''(c) = 2d.$$

Solving $\gamma'(c) = 0$ gives the optimal shift $c = \frac{1}{d}\sum_{k=1}^{d} \log(k)$. Since $\gamma''(c) > 0$, this is indeed a minimum. Therefore

$$\begin{aligned}\min_{\theta_\star \in \arg\min f_{\text{uni}}} \|\theta_\star\|_2 &= \sqrt{\sum_{k=1}^{d} \left(-\log(k) + \frac{1}{d}\sum_{j=1}^{d} \log(j) \right)^2} \\ &= \sqrt{d \,\text{Var}_{k \sim \text{Unif}[d]}[\log(k)]}.\end{aligned}$$

Let $K_d \sim \text{Unif}[d]$ and define

$$X_d := \log\left(\frac{K_d}{d}\right), \qquad \mu_d := \mathbb{E}[X_d], \qquad \nu_d := \mathbb{E}[X_d^2].$$

Since variance is invariant under constant shifts,

$$\text{Var}[\log K_d] = \text{Var}[X_d] = \nu_d - \mu_d^2.$$

Because $x \mapsto \log x$ is increasing on $(0, 1]$, the right- and left-Riemann-sum bounds give

$$-1 = \int_0^1 \log x \, dx \le \mu_d \le \int_{1/d}^1 \log x \, dx = -1 + \frac{1 + \log d}{d}.$$

Since $(1 + \log d)/d < 1$ for $d \ge 2$, we have $\mu_d < 0$, and therefore

$$\mu_d^2 \ge \left(1 - \frac{1 + \log d}{d}\right)^2.$$

Also, $x \mapsto \log^2 x$ is decreasing on $(0, 1]$, so

$$\begin{aligned}\nu_d &= \frac{\log^2 d}{d} + \frac{1}{d}\sum_{k=2}^{d} \log^2\left(\frac{k}{d}\right) \\ &\le \frac{\log^2 d}{d} + \int_{1/d}^1 \log^2 x \, dx \\ &= \frac{\log^2 d}{d} + 2 - \frac{\log^2 d + 2\log d + 2}{d} \\ &= 2 - \frac{2(1 + \log d)}{d}.\end{aligned}$$

Consequently,

$$\begin{aligned}\mathrm{Var}_{k\sim\mathrm{Unif}[d]}[\log k] &= \nu_d - \mu_d^2 \\ &\le 2 - \frac{2(1+\log d)}{d} - \left(1 - \frac{1+\log d}{d}\right)^2 \\ &= 1 - \left(\frac{1+\log d}{d}\right)^2 < 1.\end{aligned}$$

This proves the explicit upper bound and the estimate $\min_{\theta_\star \in \arg\min f_{\mathrm{uni}}} \|\theta_\star\|_2 \le \sqrt{d}$.

Finally, since $\log x$ and $\log^2 x$ are improper Riemann integrable on $(0,1]$, we also have

$$\mu_d = \frac{1}{d}\sum_{k=1}^{d} \log\left(\frac{k}{d}\right) \to -1, \qquad \nu_d = \frac{1}{d}\sum_{k=1}^{d} \log^2\left(\frac{k}{d}\right) \to 2.$$

Therefore,

$$\mathrm{Var}_{k\sim\mathrm{Unif}[d]}[\log k] = \nu_d - \mu_d^2 \to 2 - (-1)^2 = 1,$$

which proves the asymptotic claim. □

Lemma B.4 (Stochastic-gradient error)**.** *For f_{uni} and its mini-batch stochastic gradient with batch size B, we have*

$$\mathbb{E}\left[\left\|\nabla \widehat{f}_{\mathrm{uni},B}(\theta;\xi_{1:B}) - \nabla f_{\mathrm{uni}}(\theta)\right\|_2\right] \le \sqrt{\frac{1}{B}},$$
$$\mathbb{E}\left[\left\|\nabla \widehat{f}_{\mathrm{uni},B}(\theta;\xi_{1:B}) - \nabla f_{\mathrm{uni}}(\theta)\right\|_1\right] \le \min\left\{2, \sqrt{\frac{1}{B}}\|\sqrt{p}\|_1\right\},$$

where the $\sqrt{\cdot}$ operator is applied element-wise. Moreover, for large d,

$$\|\sqrt{p}\|_1 \sim 2\sqrt{\frac{d}{\log d}}.$$

Proof. For one token $\xi \sim \mathrm{Cat}(p)$, differentiating its cross-entropy contribution gives $\nabla_\theta(-\log[\mathrm{softmax}(\theta)]_\xi) = \mathrm{softmax}(\theta) - e_\xi$. Averaging over B samples gives the mini-batch gradient defined in Equation (2):

$$\nabla \widehat{f}_{\mathrm{uni},B}(\theta;\xi_1,\dots,\xi_B) = \mathrm{softmax}(\theta) - \hat{p},$$

where $\hat{p} = \frac{1}{B}\sum_{b=1}^{B} e_{\xi_b}$. By definition, $\hat{p} \sim \frac{1}{B}\mathrm{Multinomial}(B,p)$.

Now, compute the covariance of the stochastic gradient. Since $\nabla \widehat{f}_{\mathrm{uni},B}(\theta;\xi_{1:B}) - \nabla f_{\mathrm{uni}}(\theta) = p - \hat{p}$,

$$\mathrm{Cov}(\hat{p}) = \frac{1}{B}\left(\mathrm{diag}(p) - pp^\top\right),$$

which follows from the covariance of the multinomial distribution. By Jensen's inequality,

$$\mathbb{E}[\|\hat{p}-p\|_2] \le \sqrt{\mathbb{E}[\|\hat{p}-p\|_2^2]} = \sqrt{\mathrm{Tr}(\mathrm{Cov}(\hat{p}))} = \sqrt{\frac{1-\|p\|_2^2}{B}} \le \sqrt{\frac{1}{B}}.$$

For the ℓ_1 noise bound, Jensen again gives

$$\mathbb{E}[\|\hat{p}-p\|_1] = \sum_{i=1}^{d}\mathbb{E}[|\hat{p}_i - p_i|] \le \sum_{i=1}^{d}\sqrt{\mathbb{E}[(\hat{p}_i-p_i)^2]}.$$

Let $K_i \sim \mathrm{Bin}(B,p_i)$ denote the ith count in the multinomial draw. Then

$$\mathbb{E}[(\hat{p}_i-p_i)^2] = \frac{1}{B^2}\mathrm{Var}[K_i] = \frac{p_i(1-p_i)}{B}.$$

Therefore,

$$\mathbb{E}[\|\hat{p} - p\|_1] \le \sum_{i=1}^{d} \sqrt{\frac{p_i(1-p_i)}{B}} \le \sqrt{\frac{1}{B}} \sum_{i=1}^{d} \sqrt{p_i} = \sqrt{\frac{1}{B}} \|\sqrt{p}\|_1.$$

Since both $\hat{p}$ and p are probability vectors, we also have

$$\|\hat{p} - p\|_1 \le \|\hat{p}\|_1 + \|p\|_1 = 2.$$

Combining the two bounds gives the stated minimum upper bound.

For large-d, note that

$$\|\sqrt{p}\|_1 = \frac{1}{\sqrt{H_d}} \sum_{i=1}^{d} \frac{1}{\sqrt{i}},$$

where $H_d = \sum_{i=1}^{d} \frac{1}{i}$. Since $x \mapsto 1/x$ is decreasing,

$$\log(d+1) \le H_d \le 1 + \log d.$$

Likewise, since $x \mapsto x^{-1/2}$ is decreasing,

$$2(\sqrt{d+1} - 1) \le \sum_{i=1}^{d} \frac{1}{\sqrt{i}} \le 1 + 2\sqrt{d} - 2.$$

Combining these bounds shows

$$\frac{2(\sqrt{d+1} - 1)}{\sqrt{1 + \log d}} \le \|\sqrt{p}\|_1 \le \frac{2\sqrt{d} - 1}{\sqrt{\log(d+1)}},$$

and hence $\left\|\sqrt{p}\right\|_1 \sim 2\sqrt{d/\log d}$. □

C Proofs without Weight Decay

Algorithm 4 (Normalized) Steepest Descent

1: **Input:** Start point x_0, learning rate η
2: **for** $t = 1$ to $\dots$ **do**
3: Compute descent direction $u_t \leftarrow \arg\max_{\|y\| \le 1} \langle \nabla f(x_{t-1}), y \rangle$
4: $x_t \leftarrow \begin{cases} x_{t-1} - \eta \|\nabla f(x_{t-1})\|_\star u_t & \text{Standard} \\ x_{t-1} - \eta u_t & \text{Normalized} \end{cases}$
5: **end for**

In this section, we prove convergence bounds for GD, SignGD, and unnormalized SignGD, all without weight decay. GD and unnormalized SignGD are instances of the standard steepest descent branch of Algorithm 4 with respect to the ℓ_2 and ℓ_∞ norms, respectively, while fixed-step SignGD is an instance of the normalized branch. We first provide generic convergence proofs for normalized steepest descent in Lemma C.1 and standard steepest descent in Lemma C.2. The generic part uses only convexity, smoothness, and a certificate for the trajectory radius, where $\mathcal{X}^\star := \arg\min_{x \in \mathbb{R}^d} f(x) \neq \emptyset$:

$$R_T := \sup_{0 \le t \le T-1} \operatorname{dist}_{\|\cdot\|}(x_t, \mathcal{X}^\star).$$

Equivalently, if the same certificate holds for the whole trajectory, one may take $R_{\mathrm{traj}} := \sup_{t \ge 0} \operatorname{dist}_{\|\cdot\|}(x_t, \mathcal{X}^\star)$; the finite-horizon definition above is the quantity needed for a T-step bound. We then certify upper bounds on this quantity for the softmax unigram objective.

Lemma C.1 (Generic fixed-step NSD bound)**.** *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth with respect to $\|\cdot\|$, and let $\mathcal{X}^\star := \arg\min_{x \in \mathbb{R}^d} f(x) \neq \emptyset$. Suppose we run normalized steepest descent with respect to $\|\cdot\|$ with a constant learning rate η. Fix a horizon T, and let*

$$R_T := \sup_{0 \le t \le T-1} \operatorname{dist}_{\|\cdot\|}(x_t, \mathcal{X}^\star).$$

If $R \ge R_T$ and $0 < \eta \le R$, then the iterates satisfy

$$f(x_T) - f^\star \le \left(1 - \frac{\eta}{R}\right)^T (f(x_0) - f^\star) + \frac{L\eta R}{2}.$$

Proof. Let $G_t := \|\nabla f(x_t)\|_\star$. By the definition of u_t, $\langle \nabla f(x_t), u_t \rangle = G_t$. Smoothness gives

$$\begin{aligned} f(x_{t+1}) &\le f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ &\le f(x_t) - \eta G_t + \frac{L\eta^2}{2}. \end{aligned}$$

Let $\delta_t := f(x_t) - f^\star$. Since $R \ge R_T$, for every $t = 0, \dots, T-1$ we can choose $x_t^\star \in \mathcal{X}^\star$ with $\|x_t - x_t^\star\| \le R$. By convexity and Hölder's inequality,

$$\delta_t \le \langle \nabla f(x_t), x_t - x_t^\star \rangle \le \|\nabla f(x_t)\|_\star \|x_t - x_t^\star\| \le R G_t.$$

Thus $G_t \ge \delta_t / R$, and

$$\delta_{t+1} \le \left(1 - \frac{\eta}{R}\right) \delta_t + \frac{L\eta^2}{2}.$$

Unrolling the recursion and using $0 < \eta \le R$ gives

$$\begin{aligned} \delta_T &\le \left(1 - \frac{\eta}{R}\right)^T \delta_0 + \frac{L\eta^2}{2} \sum_{s=0}^{T-1} \left(1 - \frac{\eta}{R}\right)^s \\ &\le \left(1 - \frac{\eta}{R}\right)^T \delta_0 + \frac{L\eta R}{2}. \end{aligned}$$

□

Lemma C.2 (Generic steepest descent bound)**.** *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth with respect to $\|\cdot\|$, and let $\mathcal{X}^\star := \arg\min_{x \in \mathbb{R}^d} f(x) \ne \emptyset$. Suppose we run the standard steepest descent branch of Algorithm 4 with the learning rate $\eta = 1/L$. Fix a horizon T, and let*

$$R_T := \sup_{0 \le t \le T-1} \operatorname{dist}_{\|\cdot\|}(x_t, \mathcal{X}^\star).$$

If $R \ge R_T$, then the iterates satisfy

$$f(x_T) - f^\star \le \frac{2LR^2}{T}.$$

Proof. Let $G_t := \|\nabla f(x_t)\|_\star$. If $G_t = 0$, then x_t is optimal and the claim is trivial. Otherwise, by the definition of u_t, $\langle \nabla f(x_t), u_t \rangle = G_t$ and $\|u_t\| \le 1$. Smoothness and $\langle \nabla f(x_t), u_t \rangle = G_t$ give

$$\begin{aligned} f(x_{t+1}) &\le f(x_t) - \frac{1}{L} G_t^2 + \frac{L}{2} \frac{1}{L^2} G_t^2 \\ &= f(x_t) - \frac{1}{2L} G_t^2. \end{aligned}$$

With $\delta_t := f(x_t) - f^\star$, the trajectory-radius bound gives the needed gradient lower bound. Since $R \ge R_T$, for every $t = 0, \dots, T-1$ we can choose $x_t^\star \in \mathcal{X}^\star$ with $\|x_t - x_t^\star\| \le R$. By convexity and Hölder's inequality,

$$\delta_t \le \langle \nabla f(x_t), x_t - x_t^\star \rangle \le \|\nabla f(x_t)\|_\star \|x_t - x_t^\star\| \le R G_t.$$

Thus $G_t \ge \delta_t / R$, hence

$$\delta_{t+1} \le \delta_t - \frac{1}{2LR^2} \delta_t^2.$$

If $\delta_t = 0$ the claim is trivial. Otherwise $\delta_{t+1} \le \delta_t$ and

$$\frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} = \frac{\delta_t - \delta_{t+1}}{\delta_t \delta_{t+1}} \ge \frac{\delta_t - \delta_{t+1}}{\delta_t^2} \ge \frac{1}{2LR^2}.$$

Summing over $t = 0, \dots, T-1$ proves the result. □

Theorem C.3 (GD as Euclidean steepest descent)**.** *Consider optimizing f_{uni} using full-batch GD initialized at $\theta_0 = 0$ with constant step size $\eta = 2$. Then, for every $T \ge 1$,*

$$f_{\text{uni}}(\theta_T) - f_{\text{uni}}^\star \le \frac{d}{T}.$$

Proof. We instantiate the generic steepest descent bound in Lemma C.2. The main task is to determine the radius R for the iterates of GD on the softmax unigram model.

Let $\theta^\star \in \arg\min f_{\text{uni}}$ be an arbitrary minimizer. Since f_{uni} is convex and L_2-smooth for any valid Euclidean smoothness constant L_2, cocoercivity gives

$$\langle \nabla f_{\text{uni}}(\theta_t), \theta_t - \theta^\star \rangle \geq \frac{1}{L_2} \|\nabla f_{\text{uni}}(\theta_t)\|_2^2.$$

Therefore

$$\begin{aligned}\|\theta_{t+1} - \theta^\star\|_2^2 &= \|\theta_t - \theta^\star - \eta \nabla f_{\text{uni}}(\theta_t)\|_2^2 \\ &\leq \|\theta_t - \theta^\star\|_2^2 - \eta \left(\frac{2}{L_2} - \eta\right) \|\nabla f_{\text{uni}}(\theta_t)\|_2^2.\end{aligned}$$

Thus, if $0 < \eta \leq 2/L_2$, the distance to every minimizer is nonincreasing. Let $\theta_2^\star$ be a minimum-ℓ_2-norm minimizer, and define the finite-horizon Euclidean trajectory radius

$$R_T^{(2)} := \sup_{0 \leq t \leq T-1} \text{dist}_2(\theta_t, \arg\min f_{\text{uni}}).$$

We now upper bound $R_T^{(2)}$ without interchanging the supremum over t and the infimum inside the distance-to-set definition. For each fixed t, the distance to the set is at most the distance to any particular minimizer. In particular,

$$\text{dist}_2(\theta_t, \arg\min f_{\text{uni}}) = \inf_{\theta^\star \in \arg\min f_{\text{uni}}} \|\theta_t - \theta^\star\|_2 \leq \|\theta_t - \theta_2^\star\|_2.$$

Taking the supremum over $t = 0, \ldots, T-1$ only after this pointwise inequality gives

$$R_T^{(2)} \leq \sup_{0 \leq t \leq T-1} \|\theta_t - \theta_2^\star\|_2.$$

Since $\theta_0 = 0$ and the distance to the fixed minimizer $\theta_2^\star$ is nonincreasing,

$$\sup_{0 \leq t \leq T-1} \|\theta_t - \theta_2^\star\|_2 \leq \|\theta_0 - \theta_2^\star\|_2 = \|\theta_2^\star\|_2 =: R_2.$$

By Lemma B.2, the exact Euclidean smoothness constant is $L_{\|\cdot\|_2}(f_{\text{uni}}) = 1/2$. By Lemma B.3,

$$R_2^2 = d \operatorname{Var}_{k \sim \text{Unif}[d]}[\log k] \leq d.$$

Applying Lemma C.2 with $L = 1/2$, hence with the standard steepest descent learning rate $\eta = 1/L = 2$, gives

$$f_{\text{uni}}(\theta_T) - f_{\text{uni}}^\star \leq \frac{R_2^2}{T} \leq \frac{d}{T}.$$

This matches the direct smooth-convex proof of Theorem 4.1. □

Lemma C.4 (Softmax residual radius for SignGD)**.** *For a fixed iterate θ_t, let*

$$q_t := \text{softmax}(\theta_t), \qquad r_{t,i} := \log \frac{q_{t,i}}{p_i}.$$

We write

$$M_t := \max_i r_{t,i}, \qquad \ell_t := \min_i r_{t,i}, \qquad V_t := M_t - \ell_t.$$

Then

$$\text{dist}_\infty(\theta_t, \arg\min f_{\text{uni}}) = \frac{V_t}{2}.$$

Moreover, if $\theta_0 = 0$, then $V_0 = \log d$.

Proof. Since $q_{t,i} - p_i = p_i(e^{r_{t,i}} - 1)$, the signs of $q_{t,i} - p_i$ and $r_{t,i}$ agree coordinate-wise.

The minimizer set is

$$\mathcal{X}^\star = \arg\min f_{\text{uni}} = \{\log p + c\mathbf{1} : c \in \mathbb{R}\}.$$

For the current iterate θ_t, define the particular shifted minimizer

$$\bar{\theta}_t := \log p + c_t \mathbf{1}, \qquad c_t := \log\left(\sum_j e^{\theta_{t,j}}\right) + \frac{M_t + \ell_t}{2}.$$

The notation $\bar{\theta}_t$ emphasizes that this optimizer may depend on t: it is not a fixed optimizer used for all times, but the shift of the minimizer set that is closest to the current iterate in ℓ_∞ norm. Indeed, using

$$r_{t,i} = \theta_{t,i} - \log p_i - \log\left(\sum_j e^{\theta_{t,j}}\right),$$

we get

$$\theta_{t,i} - \bar{\theta}_{t,i} = r_{t,i} - \frac{M_t + \ell_t}{2}.$$

Therefore the coordinates of $\theta_t - \bar{\theta}_t$ lie in the interval

$$\left[\frac{\ell_t - M_t}{2}, \frac{M_t - \ell_t}{2}\right],$$

and so

$$\left\|\theta_t - \bar{\theta}_t\right\|_\infty = \frac{V_t}{2}.$$

Conversely, fix any shift $c \in \mathbb{R}$ and write

$$z_i(c) := \theta_{t,i} - (\log p_i + c).$$

Using the definition of $r_{t,i}$, this can be rewritten as

$$z_i(c) = r_{t,i} + \log\left(\sum_j e^{\theta_{t,j}}\right) - c.$$

The last two terms are independent of i. Therefore changing the shift c only adds the same scalar to every coordinate, so it does not change the coordinate span:

$$\max_i z_i(c) - \min_i z_i(c) = \max_i r_{t,i} - \min_i r_{t,i} = V_t.$$

For any vector z, its coordinate span is bounded by twice its ℓ_∞ norm,

$$V_t = \max_i z_i - \min_i z_i \le \max_i |z_i| + \max_i |z_i| = 2\|z\|_\infty.$$

Thus every shifted minimizer $\log p + c\mathbf{1}$ is at ℓ_∞ distance at least $V_t/2$ from θ_t. Combined with the explicit choice $\bar{\theta}_t$, which achieves distance $V_t/2$, this proves

$$\mathrm{dist}_\infty(\theta_t, \mathcal{X}^\star) = \frac{V_t}{2}.$$

At initialization $\theta_0 = 0$, $q_0 = (1/d)\mathbf{1}$. Since $p_i = 1/(iH_d)$,

$$r_{0,i} = \log\frac{q_{0,i}}{p_i} = \log\left(\frac{iH_d}{d}\right), \qquad V_0 = \log d.$$

□

Theorem C.5 (Fixed-step SignGD without weight decay)**.** *Consider optimizing f_{uni} with fixed-step SignGD initialized at $\theta_0 = 0$. For any horizon $T \ge 2$, use the constant stepsize*

$$\eta = \frac{\log d \log T}{2T}.$$

Then

$$f_{\mathrm{uni}}(\theta_T) - f^\star_{\mathrm{uni}} \le \frac{\log d}{T} + \frac{(\log d)^2 \log T}{8T}.$$

Proof. Consider fixed-step SignGD

$$\theta_{t+1} = \theta_t - \eta \operatorname{sign}(\nabla f_{\mathrm{uni}}(\theta_t)).$$

The residual update has the form

$$r_{t+1,i} = r_{t,i} - \eta \operatorname{sign}(r_{t,i}) + b_t,$$

where b_t is a scalar independent of i. Thus b_t cancels when taking spans. Since $\ell_t \le 0 \le M_t$, the transformed values

$$r_{t,i} - \eta \operatorname{sign}(r_{t,i})$$

have span at most $\max\{V_t, 2\eta\}$. Therefore

$$V_{t+1} \le \max\{V_t, 2\eta\}.$$

If $0 < \eta \le \log d/2$, then induction and $V_0 = \log d$ (because we initialized $\theta_0 = 0$) imply $V_t \le \log d$ for all t. Thus the finite-horizon ℓ_∞ trajectory radius

$$R_T^{(\infty)} := \sup_{0 \le t \le T-1} \operatorname{dist}_\infty(\theta_t, \mathcal{X}^\star) \le \frac{\log d}{2}.$$

Applying Lemma C.1 with $\|\cdot\| = \|\cdot\|_\infty$, $L = L_{\|\cdot\|_\infty}(f_{\mathrm{uni}}) = 1$, and $R = \log d/2$ gives

$$f_{\mathrm{uni}}(\theta_T) - f_{\mathrm{uni}}^\star \le \left(1 - \frac{2\eta}{\log d}\right)^T (f_{\mathrm{uni}}(\theta_0) - f_{\mathrm{uni}}^\star) + \frac{\eta \log d}{4}.$$

For a prescribed horizon $T \ge 2$, choose the constant stepsize

$$\eta = \frac{\log d \log T}{2T}.$$

This choice satisfies $0 < \eta \le \log d/2$. Moreover,

$$\left(1 - \frac{2\eta}{\log d}\right)^T \le \exp\left(-\frac{2\eta T}{\log d}\right) = \frac{1}{T}.$$

By Lemma B.1, $f_{\mathrm{uni}}(\theta_0) - f_{\mathrm{uni}}^\star \le \log d$, so

$$f_{\mathrm{uni}}(\theta_T) - f_{\mathrm{uni}}^\star \le \frac{\log d}{T} + \frac{(\log d)^2 \log T}{8T}.$$

Thus, a horizon-tuned fixed stepsize preserves the polylogarithmic dependence on d, with only an additional $\log T$ factor compared with the $\mathcal{O}((\log d)^2/T)$ rate below for unnormalized SignGD. □

Theorem C.6 (Unnormalized SignGD without weight decay)**.** *Consider optimizing f_{uni} from $\theta_0 = 0$ using the update*

$$\theta_{t+1} = \theta_t - \eta \|\nabla f_{\mathrm{uni}}(\theta_t)\|_1 \operatorname{sign}(\nabla f_{\mathrm{uni}}(\theta_t))$$

with constant learning rate $\eta = 1/4$. Then, for every $T \ge 1$,

$$f_{\mathrm{uni}}(\theta_T) - f_{\mathrm{uni}}^\star \le \frac{8(\log d)^2}{7T}.$$

Proof. This is rescaled steepest descent in the ℓ_∞ norm. We first recall the softmax-specific stepsize control. Since $\sum_i (q_{t,i} - p_i) = 0$,

$$\frac{1}{2}\|\nabla f_{\mathrm{uni}}(\theta_t)\|_1 = \sum_{i: r_{t,i} > 0} (q_{t,i} - p_i).$$

For $r_{t,i} > 0$,

$$q_{t,i} - p_i = q_{t,i}(1 - e^{-r_{t,i}}) \le q_{t,i} r_{t,i} \le q_{t,i} M_t,$$

so $\|\nabla f_{\mathrm{uni}}(\theta_t)\|_1/2 \le M_t$. Applying the same argument to the negative residuals gives

$$\frac{1}{2}\|\nabla f_{\mathrm{uni}}(\theta_t)\|_1 \le \min\{M_t, -\ell_t\}.$$

If $0 < \eta \le 1/4$, then the actual coordinate step length

$$s_t := \eta \|\nabla f_{\mathrm{uni}}(\theta_t)\|_1$$

satisfies $s_t \le \frac{1}{2}\min\{M_t, -\ell_t\}$. The residual-span argument then gives

$$V_{t+1} = V_t - 2s_t \le V_t.$$

Therefore $V_t \le V_0 = \log d$, and the finite-horizon trajectory radius satisfies

$$R_T^{(\infty)} := \sup_{0 \le t \le T-1} \mathrm{dist}_\infty(\theta_t, \mathcal{X}^\star) \le \frac{\log d}{2}.$$

For rescaled SignGD, the residual-span certificate above requires the smaller learning rate $\eta = 1/4$, rather than the generic choice $1/L = 1$. With $G_t = \|\nabla f_{\mathrm{uni}}(\theta_t)\|_1$, smoothness gives

$$f_{\mathrm{uni}}(\theta_{t+1}) \le f_{\mathrm{uni}}(\theta_t) - \frac{1}{4}G_t^2 + \frac{1}{2}\frac{1}{16}G_t^2 = f_{\mathrm{uni}}(\theta_t) - \frac{7}{32}G_t^2.$$

Let $R = \log d/2$ and $\Delta_t := f_{\mathrm{uni}}(\theta_t) - f_{\mathrm{uni}}^\star$. The trajectory-radius certificate above gives $R_T^{(\infty)} \le R$, so for every $t = 0, \dots, T-1$,

$$\mathrm{dist}_\infty(\theta_t, \mathcal{X}^\star) \le R.$$

Use the shifted minimizer $\bar\theta_t$ from Lemma C.4, which satisfies $\|\theta_t - \bar\theta_t\|_\infty \le R$. By convexity and Hölder's inequality in the ℓ_∞/ℓ_1 pair,

$$\Delta_t \le \langle \nabla f_{\mathrm{uni}}(\theta_t), \theta_t - \bar\theta_t \rangle \le \|\nabla f_{\mathrm{uni}}(\theta_t)\|_1 \|\theta_t - \bar\theta_t\|_\infty \le RG_t.$$

Therefore $G_t \ge \Delta_t/R$, and

$$\Delta_{t+1} \le \Delta_t - \frac{7}{32R^2}\Delta_t^2.$$

The same reciprocal-sum argument as in Lemma C.2 gives

$$f_{\mathrm{uni}}(\theta_T) - f_{\mathrm{uni}}^\star \le \frac{(\log d)^2/4}{(7/32)T} = \frac{8(\log d)^2}{7T}.$$

□

C.1 Proof of Theorem 4.3

The proof iterates the following one-step recursion.

Lemma C.7 (One-step radius-to-rate recursion). *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth with respect to a norm $\|\cdot\|$, with $\mathcal{X}^\star := \arg\min f \ne \emptyset$, and let $x_t = x_{t-1} - \eta_t u_t$ be one step of NSD. Write $\delta_t := f(x_t) - f^\star$. If $\mathrm{dist}_{\|\cdot\|}(x_{t-1}, \mathcal{X}^\star) \le R$ and $0 < \eta_t \le R$, then*

$$\delta_t \le \left(1 - \frac{\eta_t}{R}\right)\delta_{t-1} + \frac{L\eta_t^2}{2}.$$

Proof. If $\nabla f(x_{t-1}) = 0$, then x_{t-1} is a minimizer, the algorithm stops, and there is nothing to prove. Otherwise $\|u_t\| = 1$ and $\langle \nabla f(x_{t-1}), u_t \rangle = \|\nabla f(x_{t-1})\|_\star$, so smoothness gives

$$f(x_t) \le f(x_{t-1}) - \eta_t \langle \nabla f(x_{t-1}), u_t \rangle + \frac{L\eta_t^2}{2}\|u_t\|^2 = f(x_{t-1}) - \eta_t \|\nabla f(x_{t-1})\|_\star + \frac{L\eta_t^2}{2}.$$

Since $\mathcal{X}^\star$ is closed, we may choose $x_{t-1}^\star \in \mathcal{X}^\star$ with $\|x_{t-1} - x_{t-1}^\star\| = \mathrm{dist}_{\|\cdot\|}(x_{t-1}, \mathcal{X}^\star) \le R$. Convexity and Hölder's inequality give

$$\delta_{t-1} \le \langle \nabla f(x_{t-1}), x_{t-1} - x_{t-1}^\star \rangle \le \|\nabla f(x_{t-1})\|_\star \|x_{t-1} - x_{t-1}^\star\| \le R\,\|\nabla f(x_{t-1})\|_\star,$$

so $\|\nabla f(x_{t-1})\|_\star \ge \delta_{t-1}/R$. Substituting into the previous display proves the claim. □

The schedule is constant, $\eta_t \equiv \eta = \frac{R \log T}{T}$, and $0 < \eta \le R$. Iterating Lemma C.7 for $t = 1, \dots, T$ gives

$$\delta_T \le \left(1 - \frac{\eta}{R}\right)^T \delta_0 + \frac{L\eta R}{2}.$$

The radius assumption at $t = 1$ gives $\mathrm{dist}_{\|\cdot\|}(x_0, \mathcal{X}^\star) \le R$. Thus smoothness at a closest minimizer gives $\delta_0 \le LR^2/2$. Also,

$$\left(1 - \frac{\log T}{T}\right)^T \le \frac{1}{T}.$$

Therefore,

$$\delta_T \le \frac{LR^2(1 + \log T)}{2T} \le \frac{LR^2 \log T}{T},$$

where the last inequality uses $T \ge 3$. □

C.2 SignGD distance bound and proof

Lemma C.8 (Distance-to-minimizer-set bound for SignGD)**.** *Suppose we run SignGD on f_{uni} initialized at $\theta_0 = 0$ with horizon $T \ge 2$, and learning rate $\eta = \frac{\log d \, \log T}{2T}$. Then*

$$\mathrm{dist}_{\|\cdot\|_\infty}(\theta_t, \arg\min f_{\mathrm{uni}}) \le \frac{\log d}{2} \qquad \textit{for all } 0 \le t \le T.$$

The quotient-distance identity $\mathrm{dist}_\infty(\theta_t, \mathcal{X}^\star) = V_t/2$ and the initialization value $V_0 = \log d$ are proved in Lemma C.4. It remains to prove the span recursion for an arbitrary step size $\eta_t > 0$.

Write $Z_t := \log \sum_j e^{\theta_{t,j}}$, $M_t := \max_i r_{t,i}$, and $m_t := \min_i r_{t,i}$. Since $q_{t,i} - p_i = p_i(e^{r_{t,i}} - 1)$, we have $\mathrm{sign}(q_{t,i} - p_i) = \mathrm{sign}(r_{t,i})$, so the SignGD update $\theta_t = \theta_{t-1} - \eta_t \,\mathrm{sign}(q_{t-1} - p)$ yields

$$r_{t,i} = r_{t-1,i} - \eta_t \,\mathrm{sign}(r_{t-1,i}) + b_t, \qquad b_t := Z_{t-1} - Z_t,$$

where b_t does not depend on i and therefore cancels in the span. Because q_{t-1} and p are both probability vectors, some coordinate has $q_{t-1,i} \ge p_i$ and some has $q_{t-1,i} \le p_i$, hence $M_{t-1} \ge 0 \ge m_{t-1}$. The map $r \mapsto r - \eta_t \,\mathrm{sign}(r)$ sends each positive residual into $(-\eta_t,\ M_{t-1} - \eta_t]$, each negative residual into $[m_{t-1} + \eta_t,\ \eta_t)$, and zero to zero. Every updated residual therefore lies in the interval $[\min\{m_{t-1} + \eta_t, -\eta_t\},\ \max\{M_{t-1} - \eta_t, \eta_t\}]$, so

$$V_t \le \max\{M_{t-1} - \eta_t, \eta_t\} - \min\{m_{t-1} + \eta_t, -\eta_t\} \le \max\{V_{t-1} - 2\eta_t,\ M_{t-1},\ -m_{t-1},\ 2\eta_t\} \le \max\{V_{t-1}, 2\eta_t\},$$

where the middle inequality checks the four combinations and the last uses $M_{t-1} \le V_{t-1}$ and $-m_{t-1} \le V_{t-1}$, both consequences of $M_{t-1} \ge 0 \ge m_{t-1}$.

Finally, if $\eta_t \le \frac{\log d}{2}$ for all t, then induction from $V_0 = \log d$ gives $V_t \le \log d$ for all t, and the quotient-distance identity converts this into $\mathrm{dist}_\infty(\theta_t, \mathcal{X}^\star) \le \frac{\log d}{2}$. □

C.3 Proof of Corollary 4.4

Set $R := \frac{\log d}{2}$. By Lemma C.8, the hypotheses of Theorem 4.3 hold with $L = 1$ and $\eta = \frac{R \log T}{T}$. Therefore,

$$\delta_T \le \frac{R^2 \log T}{T} = \frac{(\log d)^2 \log T}{4T}.$$ □

C.4 Normalized-GD distance bound and proof

Lemma C.9 (Distance-to-minimizer-set bound for normalized GD)**.** *Suppose we run normalized GD on f_{uni} with $\theta_0 = 0$, horizon $T \ge 2$, and fixed learning rate $\eta = \frac{2\sqrt{d} \, \log T}{T}$. Then*

$$\mathrm{dist}_{\|\cdot\|_2}(\theta_t, \arg\min f_{\mathrm{uni}}) \le 2\sqrt{d} \qquad \textit{for all } 0 \le t \le T.$$

Let $\theta^\star$ be a minimizer closest to θ_0 in ℓ_2. By Lemma B.3, $\|\theta_0 - \theta^\star\|_2^2 \le d$. For normalized GD, convexity gives $\langle \nabla f_{\text{uni}}(\theta_{t-1}), \theta_{t-1} - \theta^\star \rangle \ge 0$, and hence

$$\|\theta_t - \theta^\star\|_2^2 = \|\theta_{t-1} - \theta^\star\|_2^2 - \frac{2\eta_t \langle \nabla f_{\text{uni}}(\theta_{t-1}), \theta_{t-1} - \theta^\star \rangle}{\|\nabla f_{\text{uni}}(\theta_{t-1})\|_2} + \eta_t^2 \le \|\theta_{t-1} - \theta^\star\|_2^2 + \eta_t^2,$$

and hence $\|\theta_t - \theta^\star\|_2^2 \le d + \sum_{s \le t} \eta_s^2$.

With $\eta := \frac{2\sqrt{d}\log T}{T}$, for every $t \le T$,

$$\|\theta_t - \theta^\star\|_2^2 \le d + T\eta^2 = d\left(1 + \frac{4(\log T)^2}{T}\right) \le 4d,$$

since $\sup_{T \ge 2} (\log T)^2/T = 4/e^2 < 3/4$. Therefore $\text{dist}_2(\theta_t, \mathcal{X}^\star) \le 2\sqrt{d}$ for every $0 \le t \le T$. □

C.5 Proof of Corollary 4.5

By Lemma C.9, the trajectory radius is at most $R := 2\sqrt{d}$ and $\eta = \frac{R \log T}{T}$. Since Lemma B.2 gives $L = 1/2$, Theorem 4.3 yields

$$\delta_T \le \frac{LR^2 \log T}{T} = \frac{2d \log T}{T}.$$ □

D Deterministic Analysis with Weight Decay

The analysis in Section 4.2 establishes convergence through a uniform bound on the distance of the iterates to the minimizer set. For SignGD and normalized GD, establishing this bound exploits the exact gradient structure of the softmax unigram objective, making the argument specific to the deterministic full-gradient setting. Weight decay provides a more general alternative: the contraction in each update directly controls the iterate norms and supplies a generic substitute for the problem-specific distance certificate. We therefore analyze NSD-WD, defined in Section 4, by adapting the convergence guarantee of Xie & Li (2024) to a fixed, horizon-dependent learning rate. This generic norm control also allows the framework to extend naturally to the stochastic-gradient setting.

Theorem D.1 (Convergence rate of NSD-WD)**.** *Suppose f is convex and L-smooth with respect to $\|\cdot\|$. Given horizon $T \ge 2$, suppose NSD-WD is run with $\lambda \le 1/\|x_\star\|$ and learning rate $\eta = \frac{\log T}{\lambda T}$. Let $D = \max\{\|x_0\|, 1/\lambda\}$. Then*

$$f(x_T) - f^\star \le \frac{2L(1 + \lambda D)^2 \log T}{\lambda^2 T}.$$

In particular, choosing $x_0 = 0$ and $\lambda = 1/\|x_\star\|$ yields

$$f(x_T) - f^\star \le \frac{8L\|x_\star\|^2 \log T}{T}.$$

Under the ℓ_∞ and ℓ_2 norms, NSD-WD becomes SignGD and normalized GD with weight decay, respectively. By Lemmas B.2 and B.3, SignGD has $L = 1$ and $\|x_\star\|_\infty = \frac{1}{2}\log d$, while normalized GD has $L = 1/2$ and $\|x_\star\|_2 \le \sqrt{d}$; applying Theorem D.1 with $x_0 = 0$, $\lambda = 2/\log d$ for SignGD, and $\lambda = 1/\sqrt{d}$ for normalized GD gives the following corollaries.

Corollary D.2 (SignGD with weight decay)**.** *For all sufficiently large d, SignGD on f_{uni} with $\theta_0 = 0$, $T \ge 2$, weight decay factor $\lambda = \frac{2}{\log d}$, and fixed learning rate $\eta = \frac{\log d \log T}{2T}$ satisfies*

$$f_{\text{uni}}(\theta_T) - f_{\text{uni}}^\star \le \frac{2(\log d)^2 \log T}{T}.$$

Corollary D.3 (Normalized GD with weight decay)**.** *For all sufficiently large d, normalized GD on f_{uni} with $\theta_0 = 0$, $T \ge 2$, weight decay factor $\lambda = \frac{1}{\sqrt{d}}$, and fixed learning rate $\eta = \frac{\sqrt{d}\log T}{T}$ satisfies*

$$f_{\text{uni}}(\theta_T) - f_{\text{uni}}^\star \le \frac{4d \log T}{T}.$$

Their dimensional dependence matches the corresponding results in Section 4.2 up to constant factors: $\widetilde{\mathcal{O}}((\log d)^2/T)$ for SignGD and $\widetilde{\mathcal{O}}(d/T)$ for normalized GD. Thus, weight decay preserves the advantage of the ℓ_∞ geometry while providing generic iterate-norm control. In Figure 4, we compare SignGD and normalized GD with and without weight decay. The same gap appears for synthetic Zipfian probabilities and for token frequencies from real text datasets, and adding weight decay leaves it essentially unchanged.

E Proofs of Deterministic Rates

E.1 Proof of Theorem D.1

Let $\delta_t := f(x_t) - f^\star$. Since $\eta\lambda = \log T/T \le 1$ and $\lambda D \ge 1$, induction gives $\|x_t\| \le D$: if $\|x_{t-1}\| \le D$, then

$$\|x_t\| \le (1-\eta\lambda)\|x_{t-1}\| + \eta \le D + \eta(1-\lambda D) \le D.$$

Let u_t be the normalized steepest-descent direction. Since $\|u_t\| \le 1$, $\langle \nabla f(x_{t-1}), u_t\rangle = \|\nabla f(x_{t-1})\|_\star$, and $\lambda\|x_\star\| \le 1$, convexity gives

$$\langle \nabla f(x_{t-1}), u_t + \lambda x_{t-1}\rangle \ge \lambda \langle \nabla f(x_{t-1}), x_{t-1} - x_\star\rangle \ge \lambda\delta_{t-1}.$$

Smoothness and $\|u_t + \lambda x_{t-1}\| \le 1 + \lambda D$ therefore imply

$$\delta_t \le (1-\eta\lambda)\delta_{t-1} + \frac{L\eta^2}{2}(1+\lambda D)^2.$$

Iterating this recursion yields

$$\delta_T \le (1-\eta\lambda)^T\delta_0 + \frac{L\eta(1+\lambda D)^2}{2\lambda}.$$

Moreover, smoothness at a minimizer and $\|x_0 - x_\star\| \le D + 1/\lambda$ give

$$\delta_0 \le \frac{L}{2}\|x_0 - x_\star\|^2 \le \frac{L(1+\lambda D)^2}{2\lambda^2}.$$

With $\eta = \frac{\log T}{\lambda T}$, we have $(1-\eta\lambda)^T \le T^{-1}$. Hence

$$\delta_T \le \frac{L(1+\lambda D)^2(1+\log T)}{2\lambda^2 T} \le \frac{2L(1+\lambda D)^2\log T}{\lambda^2 T},$$

where the last inequality holds for $T \ge 2$. If $x_0 = 0$ and $\lambda = 1/\|x_\star\|$, then $D = \|x_\star\|$ and $1 + \lambda D = 2$, proving the final claim. □

E.2 Proof of Corollary D.2

We apply Theorem D.1 with respect to $\|\cdot\|_\infty$. By Lemma B.2, $L_{\|\cdot\|_\infty}(f_{\text{uni}}) = 1$. By Lemma B.3, the smallest ℓ_∞-norm minimizer has norm

$$R_\infty = \min_{\theta_\star \in \arg\min f_{\text{uni}}} \|\theta_\star\|_\infty = \frac{\log d}{2}.$$

With $\theta_0 = 0$, choose $\lambda = 1/R_\infty = 2/\log d$. The theorem prescribes the fixed learning rate

$$\eta_t \equiv \frac{\log d\,\log T}{2T}.$$

Therefore,

$$f_{\text{uni}}(\theta_T) - f^\star_{\text{uni}} \le \frac{8L_{\|\cdot\|_\infty}(f_{\text{uni}})R_\infty^2\log T}{T} = \frac{2(\log d)^2\log T}{T}.$$

E.3 Proof of Corollary D.3

Again apply Theorem D.1, now with respect to $\|\cdot\|_2$. By Lemma B.2,

$$L_{\|\cdot\|_2}(f_{\text{uni}}) = \frac{1}{2}.$$

By Lemma B.3, the smallest ℓ_2-norm minimizer satisfies

$$\min_{\theta_\star \in \arg\min f_{\text{uni}}} \|\theta_\star\|_2 \le \sqrt{d},$$

so $\lambda = 1/\sqrt{d}$ is admissible in Theorem D.1. With $\theta_0 = 0$, we have $D = \sqrt{d}$, and the theorem prescribes the fixed learning rate

$$\eta_t \equiv \frac{\sqrt{d}\,\log T}{T}.$$

Therefore,

$$f_{\text{uni}}(\theta_T) - f^\star_{\text{uni}} \le \frac{2L_{\|\cdot\|_2}(f_{\text{uni}})(1+\lambda D)^2 \log T}{\lambda^2 T} = \frac{4d\log T}{T}.$$

E.4 Proof of Theorem 4.1

Let $\theta_\star \in \arg\min f_{\text{uni}}$ minimize the ℓ_2 norm, so by Lemma B.3,

$$\|\theta_\star\|_2^2 = d\,\mathrm{Var}_{k\sim\mathrm{Unif}[d]}[\log k].$$

Let $g_t := \nabla f_{\text{uni}}(\theta_t)$ and run full-batch GD with stepsize $\eta = 2$:

$$\theta_{t+1} = \theta_t - 2g_t.$$

Since Lemma B.2 gives $L_{\|\cdot\|_2}(f_{\text{uni}}) = 1/2$, smoothness implies

$$f_{\text{uni}}(\theta_{t+1}) \le f_{\text{uni}}(\theta_t) - \|g_t\|_2^2.$$

Also,

$$\begin{aligned}
\|\theta_{t+1} - \theta_\star\|_2^2 &= \|\theta_t - \theta_\star\|_2^2 - 4\,\langle g_t, \theta_t - \theta_\star\rangle + 4\|g_t\|_2^2 \\
&\le \|\theta_t - \theta_\star\|_2^2 - 4\big(f_{\text{uni}}(\theta_t) - f^\star_{\text{uni}}\big) + 4\big(f_{\text{uni}}(\theta_t) - f_{\text{uni}}(\theta_{t+1})\big) \\
&= \|\theta_t - \theta_\star\|_2^2 - 4\big(f_{\text{uni}}(\theta_{t+1}) - f^\star_{\text{uni}}\big),
\end{aligned}$$

where we used convexity in the second line. Summing from $t = 0$ to $T - 1$ gives

$$4\sum_{t=1}^{T}\big(f_{\text{uni}}(\theta_t) - f^\star_{\text{uni}}\big) \le \|\theta_0 - \theta_\star\|_2^2 = d\,\mathrm{Var}_{k\sim\mathrm{Unif}[d]}[\log k].$$

Because the objective decreases along GD iterates, we have

$$T\big(f_{\text{uni}}(\theta_T) - f^\star_{\text{uni}}\big) \le \sum_{t=1}^{T}\big(f_{\text{uni}}(\theta_t) - f^\star_{\text{uni}}\big),$$

and therefore

$$f_{\text{uni}}(\theta_T) - f^\star_{\text{uni}} \le \frac{d\,\mathrm{Var}_{k\sim\mathrm{Unif}[d]}[\log k]}{4T}.$$

By Lemma B.3, we have $\mathrm{Var}_{k\sim\mathrm{Unif}[d]}[\log k] < 1$, so this implies the simpler main-text bound

$$f_{\text{uni}}(\theta_T) - f^\star_{\text{uni}} \le \frac{d}{4T}.$$

F Proof of the Deterministic GD Lower Bound

Throughout this section, let $q_t = \mathrm{softmax}(\theta_t)$ and let $H_d = \sum_{j=1}^{d} 1/j$. Since $\nabla f_{\mathrm{uni}}(\theta) = \mathrm{softmax}(\theta) - p$, full-batch GD with stepsize η satisfies

$$\theta_{t+1} = \theta_t + \eta(p - q_t).$$

Lemma F.1 (Dyadic log-ratio control)**.** *For $1 \le i \le \lfloor d/2 \rfloor$, define*

$$\omega_{t,i} := \log \frac{q_{t,i}}{q_{t,2i}} = \theta_{t,i} - \theta_{t,2i}.$$

If $\theta_0 = 0$ and $0 < \eta < d/2$, then for every $t \ge 0$ and every $i \le \lfloor d/2 \rfloor$ satisfying

$$\frac{\eta t}{i H_d} \le \log 2,$$

we have

$$0 \le \omega_{t,i} \le \frac{\eta t}{2 i H_d}.$$

Proof. We prove the claim by induction on t. The base case holds because $\theta_0 = 0$. Suppose the claim holds at time t. Fix an index i satisfying

$$\frac{\eta(t+1)}{i H_d} \le \log 2.$$

Then the induction hypothesis applies to the same index at time t, so

$$0 \le \omega_{t,i} \le \frac{\eta t}{2 i H_d}.$$

We first bound the mass of the pair $\{i, 2i\}$. Since

$$\sum_{k=1}^{d} (p_k - q_{s,k}) = 0$$

for every s, and $\theta_0 = 0$, we have $\sum_{k=1}^{d} \theta_{t,k} = 0$. Hence, by AM-GM,

$$\sum_{k=1}^{d} e^{\theta_{t,k}} \ge d.$$

Also, for every coordinate k,

$$\theta_{t,k} = \eta \sum_{s=0}^{t-1} (p_k - q_{s,k}) \le \eta t p_k = \frac{\eta t}{k H_d}.$$

Therefore

$$\theta_{t,i} \le \log 2, \qquad \theta_{t,2i} \le \frac{\log 2}{2},$$

and consequently

$$q_{t,i} + q_{t,2i} = \frac{e^{\theta_{t,i}} + e^{\theta_{t,2i}}}{\sum_{k=1}^{d} e^{\theta_{t,k}}} \le \frac{2 + \sqrt{2}}{d} \le \frac{4}{d}.$$

Write $\omega_t = \omega_{t,i}$. The GD update gives

$$\omega_{t+1} = \omega_t + \eta \left((p_i - p_{2i}) - (q_{t,i} - q_{t,2i}) \right).$$

Since $\omega_t \ge 0$, we have $q_{t,i} \ge q_{t,2i}$ and

$$\begin{aligned} q_{t,i} - q_{t,2i} &= (q_{t,i} + q_{t,2i}) \frac{q_{t,i} - q_{t,2i}}{q_{t,i} + q_{t,2i}} \\ &= (q_{t,i} + q_{t,2i}) \tanh(\omega_t/2) \le \frac{2}{d} \omega_t, \end{aligned}$$

where we used $\tanh x \le x$ for $x \ge 0$. Therefore

$$\begin{aligned}\omega_{t+1} &\ge \omega_t + \eta(p_i - p_{2i}) - \frac{2\eta}{d}\omega_t \\ &= \left(1 - \frac{2\eta}{d}\right)\omega_t + \frac{\eta}{2iH_d} \ge 0,\end{aligned}$$

since $\eta < d/2$.

Since $\omega_t \ge 0$, we also have $q_{t,i} - q_{t,2i} \ge 0$, so

$$\omega_{t+1} \le \omega_t + \eta(p_i - p_{2i}) = \omega_t + \frac{\eta}{2iH_d}.$$

Combining this with the induction hypothesis gives

$$\omega_{t+1} \le \frac{\eta(t+1)}{2iH_d}.$$

This proves the claim at time $t+1$, completing the induction. □

Lemma F.2 (KL chain rule)**.** *Let p, q be probability distributions on $[d]$, and let $\mathcal{P} = \{A_b\}_{b=1}^{B}$ be a partition of $[d]$ with $q(A_b) > 0$ for every b. Define $p_{\mathcal{P}}(b) = p(A_b)$ and $q_{\mathcal{P}}(b) = q(A_b)$, and let $p(\cdot \mid A_b)$ and $q(\cdot \mid A_b)$ be the conditional distributions inside A_b. Then*

$$\mathrm{KL}(p\|q) = \mathrm{KL}(p_{\mathcal{P}}\|q_{\mathcal{P}}) + \sum_{b=1}^{B} p(A_b)\,\mathrm{KL}(p(\cdot \mid A_b)\|q(\cdot \mid A_b)).$$

In particular,

$$\mathrm{KL}(p\|q) \ge \sum_{b=1}^{B} p(A_b)\,\mathrm{KL}(p(\cdot \mid A_b)\|q(\cdot \mid A_b)).$$

Proof. Expanding the KL divergence over the partition gives

$$\begin{aligned}\mathrm{KL}(p\|q) &= \sum_{b=1}^{B}\sum_{j\in A_b} p_j \log\frac{p_j}{q_j} \\ &= \sum_{b=1}^{B} p(A_b)\log\frac{p(A_b)}{q(A_b)} + \sum_{b=1}^{B} p(A_b)\sum_{j\in A_b}\frac{p_j}{p(A_b)}\log\frac{p_j/p(A_b)}{q_j/q(A_b)}.\end{aligned}$$

The first term is $\mathrm{KL}(p_{\mathcal{P}}\|q_{\mathcal{P}})$ and the second term is the conditional KL sum. □

Lemma F.3 (Odd harmonic tail)**.** *For integers $1 \le M \le n$,*

$$\sum_{\substack{M\le i\le n\\ i\ odd}} \frac{1}{i} \ge \frac{1}{2}\log\left(\frac{n+1}{M+1}\right).$$

Proof. If the interval contains no odd integer, then necessarily $M = n$ is even and the right-hand side is zero. Otherwise, let m be the first odd integer at least M and let r be the last odd integer at most n. Then $m \le M+1$ and $r+2 \ge n+1$. For each odd i,

$$\frac{1}{i} \ge \frac{1}{2}\int_i^{i+2}\frac{dx}{x}.$$

Summing over the odd integers $m, m+2, \ldots, r$ yields

$$\sum_{\substack{M\le i\le n\\ i\ \mathrm{odd}}} \frac{1}{i} \ge \frac{1}{2}\int_m^{r+2}\frac{dx}{x} = \frac{1}{2}\log\left(\frac{r+2}{m}\right) \ge \frac{1}{2}\log\left(\frac{n+1}{M+1}\right).$$

□

F.1 Proof of Theorem 4.2

Proof. For a dyadic pair $B_i = \{i, 2i\}$ with $i \le \lfloor d/2 \rfloor$, the conditional target distribution is

$$p(\cdot \mid B_i) = \left(\frac{2}{3}, \frac{1}{3}\right),$$

while the conditional model distribution is

$$q_t(\cdot \mid B_i) = \left(\frac{q_{t,i}}{q_{t,i} + q_{t,2i}}, \frac{q_{t,2i}}{q_{t,i} + q_{t,2i}}\right) = \left(\frac{e^{\omega_{t,i}}}{1 + e^{\omega_{t,i}}}, \frac{1}{1 + e^{\omega_{t,i}}}\right).$$

Define

$$\psi(\rho) := \mathrm{KL}\left(\left(\frac{2}{3}, \frac{1}{3}\right) \middle\| \left(\frac{e^{\rho}}{1 + e^{\rho}}, \frac{1}{1 + e^{\rho}}\right)\right).$$

Then $\psi(\log 2) = 0$ and $\psi'(\log 2) = 0$. Moreover,

$$\psi''(\rho) = \frac{e^{\rho}}{(1 + e^{\rho})^2} \ge \frac{2}{9}, \qquad 0 \le \rho \le \log 2.$$

Hence, by Taylor's theorem around $\log 2$,

$$\psi(\rho) \ge \frac{1}{9}(\log 2 - \rho)^2, \qquad 0 \le \rho \le \log 2. \tag{3}$$

Fix $T \le d \log d \log 2/(4\eta)$. Since $H_d \ge \log d$, this implies

$$T \le \frac{dH_d \log 2}{4\eta}.$$

Define

$$M_T := \max\left\{1, \left\lceil \frac{\eta T}{H_d \log 2} \right\rceil\right\}, \qquad n := \left\lfloor \frac{d}{2} \right\rfloor, \qquad I_T := \{i : \ M_T \le i \le n, \ i \text{ is odd}\}.$$

For every $i \in I_T$, we have $\eta T/(iH_d) \le \log 2$, so Lemma F.1 implies

$$0 \le \omega_{T,i} \le \frac{\eta T}{2iH_d} \le \frac{\log 2}{2}.$$

Combining this with (3) gives

$$\mathrm{KL}(p(\cdot \mid B_i) \| q_T(\cdot \mid B_i)) = \psi(\omega_{T,i}) \ge \frac{(\log 2)^2}{36}.$$

The pairs $\{i, 2i\}$ are disjoint for odd i. Applying the KL chain rule to the partition formed by $\{B_i : i \in I_T\}$ and the remaining indices,

$$\begin{aligned} f_{\mathrm{uni}}(\theta_T) - f^{\star}_{\mathrm{uni}} = \mathrm{KL}(p \| q_T) &\ge \sum_{i \in I_T} p(B_i)\, \mathrm{KL}(p(\cdot \mid B_i) \| q_T(\cdot \mid B_i)) \\ &\ge \sum_{i \in I_T} \left(\frac{1}{iH_d} + \frac{1}{2iH_d}\right) \frac{(\log 2)^2}{36} \\ &= \frac{(\log 2)^2}{24 H_d} \sum_{i \in I_T} \frac{1}{i}. \end{aligned}$$

By Lemma F.3, using $n + 1 \ge d/2$ and $M_T + 1 \le \eta T/(H_d \log 2) + 2$,

$$\begin{aligned} \sum_{i \in I_T} \frac{1}{i} &\ge \frac{1}{2} \log\left(\frac{n+1}{M_T + 1}\right) \\ &\ge \frac{1}{2} \log\left(\frac{d/2}{\eta T/(H_d \log 2) + 2}\right) = \frac{1}{2} \log\left(\frac{dH_d \log 2}{2\eta T + 4H_d \log 2}\right). \end{aligned}$$

This proves the displayed lower bound.

It remains to convert the loss lower bound into an iteration lower bound. Let

$$T_{\mathrm{GD}}(\epsilon) := \min\{T : f_{\mathrm{uni}}(\theta_T) - f^\star_{\mathrm{uni}} \le \epsilon\}, \qquad a := \frac{48\epsilon}{(\log 2)^2}.$$

For the stated consequence, take $A = 48/(\log 2)^2$ and restrict to any absolute $\epsilon_0 > 0$ with $A\epsilon_0 < 1$, so $a = A\epsilon < 1$. If $T_{\mathrm{GD}}(\epsilon) > d \log d \log 2/(4\eta)$, then the desired polynomial lower bound follows immediately for all sufficiently large d because $a < 1$. Otherwise, since $H_d \ge \log d$, we have $T_{\mathrm{GD}}(\epsilon) \le dH_d \log 2/(4\eta)$ and the bound above applies at $T = T_{\mathrm{GD}}(\epsilon)$, so

$$\epsilon \ge \frac{(\log 2)^2}{48H_d} \log\left(\frac{dH_d \log 2}{2\eta T_{\mathrm{GD}}(\epsilon) + 4H_d \log 2}\right).$$

Rearranging,

$$T_{\mathrm{GD}}(\epsilon) \ge \frac{H_d \log 2}{2\eta}\left(d\exp(-aH_d) - 4\right).$$

Since $H_d \le 1 + \log d$,

$$d\exp(-aH_d) \ge e^{-a} d^{1-a}.$$

Because $a < 1$, the term $e^{-a}d^{1-a}$ dominates the constant 4 for all sufficiently large d. Using also $H_d \ge \log d$, there are constants $c_\epsilon, d_\epsilon > 0$ such that for every $d \ge d_\epsilon$,

$$T_{\mathrm{GD}}(\epsilon) \ge \frac{c_\epsilon}{\eta} d^{1-\frac{48\epsilon}{(\log 2)^2}} \log d.$$

This completes the proof. □

F.2 GD WITH LARGE STEP SIZES IS UNSTABLE

Theorem F.4 (Large GD steps are locally unstable)**.** *Let $p_i = 1/(iH_d)$, where $H_d = \sum_{j=1}^d 1/j$, and let $\theta_\star \in \arg\min f_{\mathrm{uni}}$. Define*

$$H_\star := \nabla^2 f_{\mathrm{uni}}(\theta_\star).$$

Then $\lambda_{\max}(H_\star) = \Theta(1/\log d)$. Moreover, consider full-batch GD with constant stepsize η initialized near $\theta_\star$. The linearized perturbation dynamics around $\theta_\star$ are

$$\delta_{t+1} = (I - \eta H_\star)\delta_t.$$

If

$$\eta > \frac{2}{\lambda_{\max}(H_\star)},$$

then the linearized map has spectral radius greater than one, and perturbations with a nonzero component along a top eigenvector grow to first order. In particular, for all sufficiently large d, every stepsize larger than a sufficiently large constant multiple of $\log d$ is locally unstable for full-batch GD.

Proof. At any minimizer $\theta_\star$, we have $\mathrm{softmax}(\theta_\star) = p$. Therefore

$$H_\star = \nabla^2 f_{\mathrm{uni}}(\theta_\star) = \mathrm{diag}(p) - pp^\top.$$

For any unit vector $u \in \mathbb{R}^d$,

$$u^\top H_\star u = \sum_{i=1}^d p_i u_i^2 - \left(\sum_{i=1}^d p_i u_i\right)^2 \le \sum_{i=1}^d p_i u_i^2 \le p_1 \|u\|_2^2 = \frac{1}{H_d}.$$

Hence $\lambda_{\max}(H_\star) \le 1/H_d$. For the lower bound, take $u = (e_1 - e_2)/\sqrt{2}$. Since $p_1 = 1/H_d$ and $p_2 = 1/(2H_d)$,

$$\begin{aligned} u^\top H_\star u &= \frac{p_1 + p_2}{2} - \frac{(p_1 - p_2)^2}{2} \\ &= \frac{3}{4H_d} - \frac{1}{8H_d^2}. \end{aligned}$$

This proves the claimed two-sided bound on $\lambda_{\max}(H_\star)$. Since $H_d = \Theta(\log d)$, we have $\lambda_{\max}(H_\star) = \Theta(1/\log d)$. Now consider full-batch GD near $\theta_\star$:

$$\theta_{t+1} = \theta_t - \eta \nabla f_{\text{uni}}(\theta_t).$$

Write a nearby iterate as

$$\theta_t = \theta_\star + \delta_t.$$

Since $\nabla f_{\text{uni}}(\theta_\star) = 0$, Taylor expansion gives

$$\nabla f_{\text{uni}}(\theta_\star + \delta_t) = H_\star \delta_t + o(\|\delta_t\|_2).$$

Substituting this into the GD update,

$$\theta_\star + \delta_{t+1} = \theta_\star + \delta_t - \eta \left(H_\star \delta_t + o(\|\delta_t\|_2)\right).$$

Thus

$$\delta_{t+1} = (I - \eta H_\star)\delta_t + o(\|\delta_t\|_2).$$

This is the linearization of GD around $\theta_\star$.

Let v be a unit eigenvector of $H_\star$ with eigenvalue $\lambda_{\max}(H_\star)$. If $\delta_t = \alpha v$ with α small, then the linearized update gives

$$\delta_{t+1} \approx \alpha(1 - \eta \lambda_{\max}(H_\star))v.$$

Therefore this perturbation grows to first order whenever

$$|1 - \eta \lambda_{\max}(H_\star)| > 1,$$

which is equivalent to

$$\eta > \frac{2}{\lambda_{\max}(H_\star)}.$$

Since $\lambda_{\max}(H_\star) = \Theta(1/\log d)$, this threshold is $\Theta(\log d)$. Hence any stepsize larger than a sufficiently large constant multiple of $\log d$ is locally unstable for GD. □

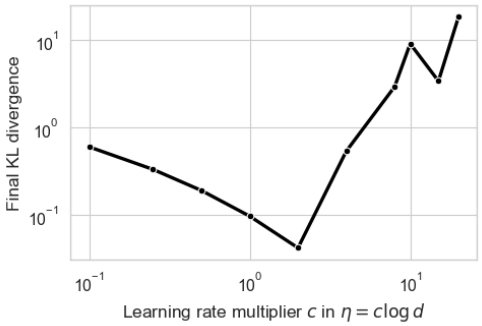


Figure 6: Grid search with vanilla GD in the intermediate- and large-step-size regimes for the softmax unigram model with $d = 10^4$. Each point reports the final KL divergence for a learning rate $\eta = c \log d$. Large learning-rate multipliers lead to instability, consistent with Theorem F.4.

G Generic Stochastic (Momentum) NSD-WD Proofs

In this section, we will present the generic convergence proof of stochastic NSD-WD in Theorem 5.1 and also provide a convergence proof of the momentum version in Theorem G.3 and Corollary G.4. For completeness, the stochastic NSD-WD algorithm with its momentum version is presented in Algorithm 5.

First, we present a descent lemma for stochastic NSD-WD that is independent of whether momentum is used. Lemma G.1 only relies on the fact that the descent direction u_t is computed directly from v_t. It does not use any information about the structure of the current estimator v_t. The estimator can be the current stochastic gradient (the standard stochastic case) or an exponential moving average (the momentum case).

Algorithm 5 Stochastic (Momentum) NSD with Weight Decay

1: **Input:** Start point x_0, initial momentum v_0 (default $v_0 = 0$), weight decay λ, learning rate η_t, momentum ρ_t
2: **for** $t = 1$ to $\dots$ **do**
3: Compute stochastic estimator $v_t = \begin{cases} g_{t-1} & \text{Vanilla Stochastic} \\ (1-\rho_t)v_{t-1} + \rho_t g_{t-1} & \text{Momentum} \end{cases}$
4: $u_t \leftarrow \arg\max_{\|y\|\leq 1} \langle v_t, y\rangle$
5: $x_t \leftarrow (1-\lambda\eta_t)x_{t-1} - \eta_t u_t$
6: **end for**

Lemma G.1 (Stochastic Descent Lemma)**.** *Suppose f is convex and L-smooth with respect to $\|\cdot\|$. Let $D = \max\{\|x_0\|, \frac{1}{\lambda}\}$. If $\lambda \leq 1/\|x_\star\|$ and $\eta_t \leq 1/\lambda$, then the iterates $\{x_t\}$ of stochastic NSD-WD satisfy,*

$$\delta_t \leq (1-\eta_t\lambda)\delta_{t-1} + 2\eta_t\mathbb{E}[\|\nabla f(x_{t-1}) - v_t\|_\star] + \frac{L\eta_t^2}{2}(1+\lambda D)^2,$$

where $\delta_t = \mathbb{E}[f(x_t) - f^\star]$.

Proof. First, we note that the iterates remain in the radius-D ball. Indeed, if $\|x_{t-1}\| \leq D$, then using $\|u_t\| \leq 1$ and $\eta_t \leq 1/\lambda$,

$$\begin{aligned}\|x_t\| &\leq (1-\lambda\eta_t)\|x_{t-1}\| + \eta_t\|u_t\| \\ &\leq (1-\lambda\eta_t)D + \eta_t \leq D,\end{aligned}$$

where the last inequality follows from $\lambda D \geq 1$. Since $\|x_0\| \leq D$, induction gives $\|x_t\| \leq D$ for all t.

Now apply L-smoothness:

$$f(x_t) \leq f(x_{t-1}) + \langle \nabla f(x_{t-1}), x_t - x_{t-1}\rangle + \frac{L}{2}\|x_t - x_{t-1}\|^2.$$

The quadratic term is bounded by the triangle inequality and $\|u_t\| \leq 1$:

$$\begin{aligned}\|x_t - x_{t-1}\|^2 &= \|-\lambda\eta_t x_{t-1} - \eta_t u_t\|^2 \\ &= \|\lambda\eta_t x_{t-1} + \eta_t u_t\|^2 \\ &\leq (\lambda\eta_t\|x_{t-1}\| + \eta_t\|u_t\|)^2 \\ &\leq \eta_t^2(\lambda D + 1)^2.\end{aligned}$$

It remains to bound the inner product term. Note that the inner product can be expressed as $-\eta_t\langle \nabla f(x_{t-1}), \lambda x_{t-1} + u_t\rangle$.

Since $u_t \in \arg\max_{\|u\|\leq 1}\langle v_t, u\rangle$, we have that $\langle v_t, u_t\rangle = \|v_t\|_\star$. Therefore,

$$\begin{aligned}\langle \nabla f(x_{t-1}), u_t\rangle &= \langle v_t, u_t\rangle + \langle \nabla f(x_{t-1}) - v_t, u_t\rangle \\ &\geq \|v_t\|_\star - \|\nabla f(x_{t-1}) - v_t\|_\star\|u_t\| \\ &\geq \|v_t\|_\star - \|\nabla f(x_{t-1}) - v_t\|_\star\end{aligned}$$

By the reverse triangle inequality, $\|v_t\|_\star = \|\nabla f(x_{t-1}) - (\nabla f(x_{t-1}) - v_t)\|_\star \geq \|\nabla f(x_{t-1})\|_\star - \|\nabla f(x_{t-1}) - v_t\|_\star$. Thus, $\langle \nabla f(x_{t-1}), u_t\rangle \geq \|\nabla f(x_{t-1})\|_\star - 2\|\nabla f(x_{t-1}) - v_t\|_\star$. Next, by the convexity of f,

$$f(x_{t-1}) - f^\star \leq \langle \nabla f(x_{t-1}), x_{t-1} - x_\star\rangle = \langle \nabla f(x_{t-1}), x_{t-1}\rangle - \langle \nabla f(x_{t-1}), x_\star\rangle.$$

Using Cauchy Schwarz with dual norm, $-\langle \nabla f(x_{t-1}), x_\star\rangle \leq \|\nabla f(x_{t-1})\|_\star\|x_\star\|$. Since $\lambda \leq 1/\|x_\star\|$, we have $\|x_\star\| \leq 1/\lambda$, and hence

$$f(x_{t-1}) - f^\star \leq \langle \nabla f(x_{t-1}), x_{t-1}\rangle + \frac{1}{\lambda}\|\nabla f(x_{t-1})\|_\star.$$

Multiplying by λ gives

$$\lambda(f(x_{t-1}) - f^\star) \leq \lambda\langle \nabla f(x_{t-1}), x_{t-1}\rangle + \|\nabla f(x_{t-1})\|_\star.$$

Combining this inequality with the lower bound on $\langle \nabla f(x_{t-1}), u_t \rangle$, we get

$$\begin{aligned}\langle \nabla f(x_{t-1}), \lambda x_{t-1} + u_t \rangle &= \lambda \langle \nabla f(x_{t-1}), x_{t-1} \rangle + \langle \nabla f(x_{t-1}), u_t \rangle \\ &\geq \lambda \langle \nabla f(x_{t-1}), x_{t-1} \rangle + \|\nabla f(x_{t-1})\|_\star - 2\|\nabla f(x_{t-1}) - v_t\|_\star \\ &\geq \lambda (f(x_{t-1}) - f^\star) - 2\|\nabla f(x_{t-1}) - v_t\|_\star.\end{aligned}$$

Substituting into the smoothness inequality yields

$$\begin{aligned}f(x_t) - f^\star &\leq f(x_{t-1}) - f^\star - \eta_t \lambda (f(x_{t-1}) - f^\star) + 2\eta_t \|\nabla f(x_{t-1}) - v_t\|_\star + \frac{L\eta_t^2}{2}(1 + \lambda D)^2 \\ &= (1 - \eta_t \lambda)(f(x_{t-1}) - f^\star) + 2\eta_t \|\nabla f(x_{t-1}) - v_t\|_\star + \frac{L\eta_t^2}{2}(1 + \lambda D)^2.\end{aligned}$$

Finally, taking expectation gives

$$\delta_t \leq (1 - \eta_t \lambda)\delta_{t-1} + 2\eta_t \mathbb{E}\|\nabla f(x_{t-1}) - v_t\|_\star + \frac{L\eta_t^2}{2}(1 + \lambda D)^2.$$

□

G.1 Stochastic NSD-WD Proofs

Here, we prove Theorem 5.1 by solving the recursion in Lemma G.1 with a horizon-tuned fixed learning rate.

Proof of Theorem 5.1. Let $\delta_t := \mathbb{E}[f(x_t) - f^\star]$, $C := (1 + \lambda D)^2$, and $\eta = \frac{\log T}{\lambda T}$. Since $T \geq 2$, $\eta \leq 1/\lambda$. For the vanilla estimator $v_t = g_{t-1}$, the noise assumption and Lemma G.1 give

$$\delta_t \leq (1 - \eta\lambda)\delta_{t-1} + 2\eta\sigma + \frac{L\eta^2 C}{2}.$$

Iterating and summing the geometric series yields

$$\delta_T \leq (1 - \eta\lambda)^T \delta_0 + \frac{2\sigma}{\lambda} + \frac{L\eta C}{2\lambda}.$$

Smoothness at a minimizer and $\|x_0 - x_\star\| \leq D + 1/\lambda$ imply

$$\delta_0 \leq \frac{L}{2}\|x_0 - x_\star\|^2 \leq \frac{LC}{2\lambda^2}.$$

Since $(1 - \eta\lambda)^T \leq e^{-\eta\lambda T} = T^{-1}$,

$$\delta_T \leq \frac{LC(1 + \log T)}{2\lambda^2 T} + \frac{2\sigma}{\lambda} \leq \frac{2LC \log T}{\lambda^2 T} + \frac{2\sigma}{\lambda}.$$

If $x_0 = 0$ and $\lambda = 1/\|x_\star\|$, then $D = \|x_\star\|$ and $C = 4$, which gives the second conclusion. □

G.2 Stochastic Momentum NSD-WD Proofs

In this subsection, we present a generic convergence bound for stochastic momentum NSD-WD in Theorem G.3. Then, we tune this bound in terms of the momentum parameter ρ to provide a convergence rate for momentum NSD-WD in Corollary G.4. Corollary G.4 will be used in Appendix J to compare upper bounds for momentum SignSGD and momentum normalized SGD on the stochastic softmax unigram model. For the proof of Theorem G.3, we first need to bound the momentum error in the dual norm which is presented in Lemma G.2.

Lemma G.2 (Momentum estimator error)**.** *Suppose f is L-smooth with respect to $\|\cdot\|$. Suppose the stochastic estimator g_t satisfies $\mathbb{E}_t[g_t] = \nabla f(x_t)$ and has bounded Euclidean second moment, $\mathbb{E}_t\left[\|g_t - \nabla f(x_t)\|_2^2\right] \leq \nu_2^2$. Let $\kappa := \sup_{z \neq 0} \|z\|_\star / \|z\|_2$ and $D = \max\{\|x_0\|, 1/\lambda\}$. Suppose the algorithm uses fixed $0 < \eta \leq 1/\lambda$ and $\rho \in (0, 1]$, so that $v_t = (1 - \rho)v_{t-1} + \rho g_{t-1}$ for $t \geq 1$. Then*

$$\mathbb{E}[\|\nabla f(x_{t-1}) - v_t\|_\star] \leq (1 - \rho)^t \mathbb{E}[\|\nabla f(x_0) - v_0\|_\star] + \kappa\nu_2\sqrt{\rho} + \frac{L\eta(1 + \lambda D)}{\rho}.$$

Proof. Let $\xi_t = g_t - \nabla f(x_t)$. Then $\mathbb{E}_t[\xi_t] = 0$, and the second-moment assumption gives $\mathbb{E}_t[\|\xi_t\|_2^2] \le \nu_2^2$. The momentum estimator can be expressed as

$$v_t = (1-\rho)v_{t-1} + \rho g_{t-1}.$$

Define $a_0 := v_0 - \nabla f(x_0)$ and, for $t \ge 1$, define the error $a_t := v_t - \nabla f(x_{t-1})$. For $t = 1$,

$$a_1 = (1-\rho)a_0 + \rho\xi_0.$$

For $t \ge 2$,

$$\begin{aligned} a_t &= (1-\rho)v_{t-1} + \rho g_{t-1} - \nabla f(x_{t-1}) \\ &= (1-\rho)a_{t-1} + \rho\big(g_{t-1} - \nabla f(x_{t-1})\big) + (1-\rho)\big(\nabla f(x_{t-2}) - \nabla f(x_{t-1})\big) \\ &= (1-\rho)a_{t-1} + (1-\rho)\big(\nabla f(x_{t-2}) - \nabla f(x_{t-1})\big) + \rho\xi_{t-1}. \end{aligned}$$

Let $\beta = 1-\rho$. Unrolling the recursion gives, for every $t \ge 1$,

$$a_t = \beta^t a_0 + \rho\sum_{k=0}^{t-1}\beta^{t-1-k}\xi_k + \sum_{k=1}^{t-1}\beta^{t-k}\big(\nabla f(x_{k-1}) - \nabla f(x_k)\big).$$

By applying the triangle inequality, we can get an upper bound on $\mathbb{E}\|a_t\|_\star$ by computing the dual norm of each term. We start with the last term. Note that we can apply smoothness,

$$\|\nabla f(x_k) - \nabla f(x_{k-1})\|_\star \le L\|x_k - x_{k-1}\| \le L\eta(1+\lambda\|x_{k-1}\|) \le L\eta(1+\lambda D).$$

Then,

$$\left\|\sum_{k=1}^{t-1}(1-\rho)^{t-k}\big(\nabla f(x_{k-1}) - \nabla f(x_k)\big)\right\|_\star \le L\eta(1+\lambda D)\sum_{j=1}^{\infty}(1-\rho)^j$$

Then,

$$L\eta(1+\lambda D)\sum_{j=1}^{\infty}(1-\rho)^j \le L\eta(1+\lambda D)\frac{1-\rho}{\rho} \le \frac{L\eta(1+\lambda D)}{\rho}$$

Now, we handle the stochastic variance term in the middle, which requires converting to the ℓ_2 norm and incurring the norm penalty κ. So,

$$\mathbb{E}\left\|\rho\sum_{k=0}^{t-1}(1-\rho)^{t-1-k}\xi_k\right\|_\star \le \kappa\mathbb{E}\left\|\rho\sum_{k=0}^{t-1}(1-\rho)^{t-1-k}\xi_k\right\|_2 \le \kappa\sqrt{\mathbb{E}\left\|\rho\sum_{k=0}^{t-1}(1-\rho)^{t-1-k}\xi_k\right\|_2^2}$$

The last inequality is obtained via Jensen's inequality.

Now compute the second moment. Let

$$\mathbb{E}\|S_t\|_2^2 = \mathbb{E}\left\|\rho\sum_{k=0}^{t-1}(1-\rho)^{t-1-k}\xi_k\right\|_2^2.$$

Expanding the square gives

$$\mathbb{E}\|S_t\|_2^2 = \rho^2\sum_{k=0}^{t-1}(1-\rho)^{2(t-1-k)}\mathbb{E}\|\xi_k\|_2^2 + 2\rho^2\sum_{0\le k<\ell\le t-1}(1-\rho)^{t-1-k}(1-\rho)^{t-1-\ell}\mathbb{E}\langle\xi_k,\xi_\ell\rangle.$$

The cross terms vanish because $(\xi_k)_{k\ge0}$ is a martingale difference sequence. Indeed, for $k < \ell$, since ξ_k is $\mathcal{F}_{\ell-1}$-measurable,

$$\mathbb{E}\langle\xi_k,\xi_\ell\rangle = \mathbb{E}\left[\mathbb{E}\left[\langle\xi_k,\xi_\ell\rangle \mid \mathcal{F}_{\ell-1}\right]\right] = \mathbb{E}\left[\langle\xi_k,\mathbb{E}[\xi_\ell \mid \mathcal{F}_{\ell-1}]\rangle\right] = 0.$$

Therefore,

$$\mathbb{E}\|S_t\|_2^2 = \rho^2\sum_{k=0}^{t-1}(1-\rho)^{2(t-1-k)}\mathbb{E}\|\xi_k\|_2^2 \le \rho^2\nu_2^2\sum_{k=0}^{t-1}(1-\rho)^{2(t-1-k)}$$

Re-indexing the sum with $j = t - 1 - k$ gives

$$\sum_{k=0}^{t-1}(1-\rho)^{2(t-1-k)} = \sum_{j=0}^{t-1}(1-\rho)^{2j} \leq \sum_{j=0}^{\infty}(1-\rho)^{2j} = \frac{1}{1-(1-\rho)^2} = \frac{1}{\rho(2-\rho)}.$$

Thus,

$$\sqrt{\mathbb{E}\|S_t\|_2^2} \leq \sqrt{\frac{\nu_2^2\rho}{2-\rho}} \leq \nu_2\sqrt{\frac{\rho}{2-\rho}} \leq \nu_2\sqrt{\rho}.$$

The last inequality comes from the fact that $0 < \rho \leq 1$ so $2 - \rho \geq 1$. Therefore,

$$\mathbb{E}\left\|\rho\sum_{k=0}^{t-1}(1-\rho)^{t-1-k}\xi_k\right\|_\star \leq \kappa\nu_2\sqrt{\rho}.$$

Putting it all together,

$$\mathbb{E}\|\nabla f(x_{t-1}) - v_t\|_\star = \mathbb{E}\|a_t\|_\star \leq (1-\rho)^t\mathbb{E}\|\nabla f(x_0) - v_0\|_\star + \kappa\nu_2\sqrt{\rho} + \frac{L\eta(1+\lambda D)}{\rho}.$$

□

Although the stochastic momentum variant of Algorithm 5 can be used with time-varying ρ_t and η_t, for our analysis, we use fixed ρ and η in Lemma G.2 and Theorem G.3. We later set ρ and η in terms of the horizon T to achieve a convergence rate for stochastic momentum NSD-WD in Corollary G.4. Using a fixed ρ allows us to tune the bound in Theorem G.3 and set ρ in terms of problem-dependent constants such as σ and $\|x_\star\|$ to achieve a tighter bound. If we naively use a time-varying ρ_t, the upper bound we get will not be tight enough to obtain a separation between momentum SignSGD and normalized SGD. As a technicality, to complete the proof of Theorem G.3 using a fixed ρ we must also use a fixed η.

Theorem G.3 (Stochastic Momentum NSD-WD)**.** *Suppose f is convex and L-smooth with respect to $\|\cdot\|$. Suppose the stochastic estimator g_t satisfies $\mathbb{E}_t[g_t] = \nabla f(x_t)$ and it has bounded Euclidean second moment $\mathbb{E}_t\left[\|g_t - \nabla f(x_t)\|_2^2\right] \leq \nu_2^2$. Suppose we run stochastic momentum NSD with weight decay $\lambda \leq \frac{1}{\|x_\star\|}$. Let $D = \max\{\|x_0\|, 1/\lambda\}$ and $\kappa = \sup_{z\neq 0}\frac{\|z\|_\star}{\|z\|_2}$. For fixed $\eta \leq 1/\lambda$ and $\rho \in (0,1]$, the iterates satisfy*

$$\mathbb{E}[f(x_T) - f^\star] \leq (1-\lambda\eta)^T(f(x_0) - f^\star) + \frac{2\eta\mathbb{E}[\|\nabla f(x_0) - v_0\|_\star]}{\rho} + \frac{2\kappa\nu_2\sqrt{\rho}}{\lambda} + \frac{3L\eta(1+\lambda D)^2}{2\rho\lambda}.$$

Proof. Let $\delta_t := \mathbb{E}[f(x_t) - f^\star], \tau := 1 - \lambda\eta$, $r := 1 - \rho$, and

$$A := \kappa\nu_2\sqrt{\rho} + \frac{L\eta(1+\lambda D)}{\rho}, \qquad B := \frac{L\eta^2(1+\lambda D)^2}{2}.$$

Combining Lemmas G.1 and G.2 gives, for $t = 1, \ldots, T$,

$$\delta_t \leq \tau\delta_{t-1} + 2\eta\left(r^t\mathbb{E}[\|\nabla f(x_0) - v_0\|_\star] + A\right) + B.$$

Unrolling the recursion,

$$\begin{aligned}
\delta_T &\leq \tau^T\delta_0 + 2\eta\mathbb{E}[\|\nabla f(x_0) - v_0\|_\star]\sum_{t=1}^{T}\tau^{T-t}r^t + (2\eta A + B)\sum_{t=1}^{T}\tau^{T-t} \\
&\leq \tau^T\delta_0 + 2\eta\mathbb{E}[\|\nabla f(x_0) - v_0\|_\star]\sum_{t=0}^{\infty}r^t + (2\eta A + B)\sum_{t=0}^{\infty}\tau^t \\
&= \tau^T\delta_0 + \frac{2\eta\mathbb{E}[\|\nabla f(x_0) - v_0\|_\star]}{\rho} + \frac{2A}{\lambda} + \frac{B}{\lambda\eta}.
\end{aligned}$$

Substituting the definitions of A and B gives

$$\delta_T \leq \tau^T\delta_0 + \frac{2\eta\mathbb{E}[\|\nabla f(x_0) - v_0\|_\star]}{\rho} + \frac{2\kappa\nu_2\sqrt{\rho}}{\lambda} + \frac{L\eta(1+\lambda D)}{\lambda}\left(\frac{2}{\rho} + \frac{1+\lambda D}{2}\right).$$

Since $\lambda D \geq 1$, we have $1 + \lambda D \geq 2$. Also, $\rho \leq 1$. Therefore,

$$\frac{2}{\rho} + \frac{1+\lambda D}{2} \leq \frac{2}{\rho} + \frac{1+\lambda D}{2\rho} = \frac{4 + (1+\lambda D)}{2\rho} \leq \frac{3(1+\lambda D)}{2\rho},$$

which proves the claim. □

Corollary G.4 (Optimized Stochastic Momentum NSD-WD)**.** *Under the assumptions of Theorem G.3, initialize at $x_0 = 0$, and choose the optimal weight decay $\lambda = 1/\|x_\star\|$. For horizon $T \geq 3$, set the fixed learning rate and momentum*

$$\eta = \frac{\|x_\star\| \log T}{T}, \qquad \rho = \left(\frac{4L\|x_\star\|}{\kappa\nu_2}\right)^{\frac{2}{3}} \left(\frac{\log T}{T}\right)^{\frac{2}{3}}.$$

If the horizon satisfies $T = \tilde{\Omega}\left(\frac{L\|x_\star\|}{\kappa\nu_2}\right)$ so that $\rho \leq 1$, then the expected suboptimality satisfies

$$\mathbb{E}[f(x_T) - f^\star] \leq \left[f(x_0) - f^\star + 2L\|x_\star\|^2\right] \frac{\log T}{T} + \frac{2(\kappa\nu_2)^{\frac{2}{3}}\|x_\star\|^{\frac{1}{3}}}{L^{\frac{2}{3}}} \left[\mathbb{E}[\|\nabla f(x_0) - v_0\|_\star] + 3L\|x_\star\|\right] \left(\frac{\log T}{T}\right)^{\frac{1}{3}}.$$

Proof. We now make our choice for learning rate η and momentum ρ. Specifically, let $\eta = \frac{\log T}{\lambda T}$. Then,

$$(1 - \lambda\eta)^T \leq e^{-\lambda\eta T} = e^{-\log T} = \frac{1}{T} \leq \frac{\log T}{T}.$$

Using the bound from Theorem G.3, we obtain

$$\delta_T \leq \frac{\delta_0}{T} + \frac{2\mathbb{E}[\|\nabla f(x_0) - v_0\|_\star]\log T}{\rho\lambda T} + \frac{2\kappa\nu_2\sqrt{\rho}}{\lambda} + \frac{2L(1+\lambda D)\log T}{\lambda^2\rho T} + \frac{L(1+\lambda D)^2 \log T}{2\lambda^2 T}$$

To choose ρ, we balance the two dominant ρ-dependent contributions

$$g(\rho) = \frac{2\kappa\nu_2\sqrt{\rho}}{\lambda} + \frac{2L(1+\lambda D)\log T}{\lambda^2\rho T}.$$

Solving $g'(\rho) = 0$ gives the tuning rule $\rho_\star = \alpha\left(\frac{\log T}{T}\right)^{2/3}$, where $\alpha = \left(\frac{2L(1+\lambda D)}{\kappa\nu_2\lambda}\right)^{2/3}$, and hence

$$g(\rho_\star) = 3 \cdot 2^{1/3} \frac{(\kappa\nu_2)^{2/3}\big(L(1+\lambda D)\log T\big)^{1/3}}{\lambda^{4/3} T^{1/3}}.$$

We must have $\rho \leq 1$, so this implies that $T \geq \log T \alpha^{3/2}$. Since we are interested in regimes where our budget and iterations are large, we suppress $\log T$ factors to simplify our analysis and let $T = \tilde{\Omega}(\alpha^{3/2})$. Next, we simply substitute ρ to obtain the bound,

$$\delta_T \leq \left[\delta_0 + \frac{L(1+\lambda D)^2}{2\lambda^2}\right] \frac{\log T}{T} + \left[\frac{2^{\frac{1}{3}}(\kappa\nu_2)^{2/3}\mathbb{E}[\|\nabla f(x_0) - v_0\|_\star]}{\lambda^{\frac{1}{3}}(L(1+\lambda D))^{\frac{2}{3}}} + 3 \cdot 2^{1/3} \frac{(\kappa\nu_2)^{\frac{2}{3}}\big(L(1+\lambda D)\big)^{\frac{1}{3}}}{\lambda^{4/3}}\right] \left(\frac{\log T}{T}\right)^{\frac{1}{3}}.$$

The remaining term

$$\frac{2\mathbb{E}[\|\nabla f(x_0) - v_0\|_\star]\log T}{\rho_\star\lambda T}$$

also scales as $(\log T/T)^{1/3}$ under this choice, which is why it appears in the same bracket above after substitution. When $x_0 = 0$ and $\lambda = 1/\|x_\star\|$, we have $D = \|x_\star\|$ and $1 + \lambda D = 2$, so

$$\rho_\star = \left(\frac{4L\|x_\star\|}{\kappa\nu_2}\right)^{\frac{2}{3}} \left(\frac{\log T}{T}\right)^{\frac{2}{3}}.$$

The condition $\rho_\star \le 1$ is therefore equivalent, up to logarithmic factors, to $T = \tilde{\Omega}\left(\frac{L\|x_\star\|}{\kappa\nu_2}\right)$. After this substitution, the two coefficients needed below satisfy $2^{-1/3} \le 2$ and $3 \cdot 2^{2/3} \le 6$. Thus

$$\delta_T \le \left[\delta_0 + 2L\|x_\star\|^2\right]\frac{\log T}{T} + \frac{2(\kappa\nu_2)^{\frac{2}{3}}\|x_\star\|^{\frac{1}{3}}}{L^{\frac{2}{3}}}\left[\mathbb{E}[\|\nabla f(x_0) - v_0\|_\star] + 3L\|x_\star\|\right]\left(\frac{\log T}{T}\right)^{\frac{1}{3}},$$

as claimed. □

H Comparison with SGD

We compare SignSGD with weight decay against standard SGD under a fixed token budget $N = BT$. The following averaged-iterate guarantee uses a constant stepsize tuned to the horizon and batch size.

Theorem H.1 (Averaged-iterate SGD)**.** *Consider optimizing f_{uni} with SGD initialized at $\theta_0 = 0$, batch size $B \ge 1$, and $T \ge 1$ iterations, for a total token budget $N = BT$. Let $\theta_\star$ be the minimum-ℓ_2-norm minimizer, set $D = \|\theta_0 - \theta_\star\|_2$, and use the constant stepsize $\eta = \min\{1, D\sqrt{B/T}\}$. Then the average iterate $\bar{\theta}_T = T^{-1}\sum_{t=0}^{T-1}\theta_t$ satisfies*

$$\mathbb{E}[f_{\mathrm{uni}}(\bar{\theta}_T) - f^\star_{\mathrm{uni}}] \le \frac{d}{T} + 2\sqrt{\frac{d}{BT}} = \frac{dB}{N} + 2\sqrt{\frac{d}{N}}.$$

The proof follows below. At fixed N, the SGD bound is smallest at $B = 1$, yielding $\mathcal{O}(\sqrt{d/N})$ when $N \ge d$. For sufficiently large N relative to d, this is smaller than the tuned SignSGD upper bound $\widetilde{\mathcal{O}}((d/N)^{1/3})$ from Corollary 5.5.

For a fixed common batch size, the comparison changes. The SignSGD bound has a noise term of order $\sqrt{d\log d/B}$, which decreases with B, whereas its deterministic term has better dimension dependence than SGD's. Under $N \ge d$ and $(\log d)^2 \log N = o(d)$, Appendix H.1 and lemma H.2 shows that a conservative SignSGD upper bound becomes smaller than the SGD bound once B exceeds a threshold of order $(N^2 \log d/d)^{1/3}$.

Proof of Theorem H.1. Let $\theta_\star$ be the minimum-ℓ_2-norm minimizer of f_{uni}, and set

$$D := \|\theta_0 - \theta_\star\|_2 = \|\theta_\star\|_2.$$

By Lemma B.3, we have the exact bound

$$D \le \sqrt{d}.$$

Let

$$\bar{g}_t = \nabla\widehat{f}_{\mathrm{uni},B}(\theta_t; \xi_{t,1:B}) = \mathrm{softmax}(\theta_t) - \hat{p}_t$$

be the mini-batch gradient used by SGD. Since

$$\bar{g}_t - \nabla f_{\mathrm{uni}}(\theta_t) = p - \hat{p}_t,$$

the covariance computation from the proof of Lemma B.4 gives

$$\mathbb{E}_t[\bar{g}_t] = \nabla f_{\mathrm{uni}}(\theta_t), \qquad \mathbb{E}_t\left[\|\bar{g}_t - \nabla f_{\mathrm{uni}}(\theta_t)\|_2^2\right] = \frac{1 - \|p\|_2^2}{B} \le \frac{1}{B}.$$

Choose the constant stepsize

$$\eta = \min\left\{1, \frac{D\sqrt{B}}{\sqrt{T}}\right\}.$$

The update is $\theta_{t+1} = \theta_t - \eta\bar{g}_t$. For any minimizer $\theta^\star$, the squared-distance recursion gives

$$\begin{aligned}\mathbb{E}_t\|\theta_{t+1} - \theta^\star\|_2^2 &= \|\theta_t - \theta^\star\|_2^2 - 2\eta\langle\nabla f_{\mathrm{uni}}(\theta_t), \theta_t - \theta^\star\rangle + \eta^2\mathbb{E}_t\|\bar{g}_t\|_2^2 \\ &\le \|\theta_t - \theta^\star\|_2^2 - 2\eta\left(f_{\mathrm{uni}}(\theta_t) - f^\star_{\mathrm{uni}}\right) + \eta^2\|\nabla f_{\mathrm{uni}}(\theta_t)\|_2^2 + \frac{\eta^2}{B}.\end{aligned}$$

By Lemma B.2, $L_{\|\cdot\|_2}(f_{\mathrm{uni}}) = 1/2$, so convex smoothness gives

$$\|\nabla f_{\mathrm{uni}}(\theta_t)\|_2^2 \le f_{\mathrm{uni}}(\theta_t) - f_{\mathrm{uni}}^\star.$$

Since $\eta \le 1$,

$$\mathbb{E}_t\|\theta_{t+1} - \theta^\star\|_2^2 \le \|\theta_t - \theta^\star\|_2^2 - \eta\,(f_{\mathrm{uni}}(\theta_t) - f_{\mathrm{uni}}^\star) + \frac{\eta^2}{B}.$$

Summing from $t = 0$ to $T - 1$ and using Jensen's inequality for $\bar\theta_T = T^{-1}\sum_{t=0}^{T-1}\theta_t$ yields

$$\mathbb{E}[f_{\mathrm{uni}}(\bar\theta_T) - f_{\mathrm{uni}}^\star] \le \frac{D^2}{\eta T} + \frac{\eta}{B}.$$

Let $a = 1$ and $b = D\sqrt{B/T}$, so $\eta = \min\{a, b\}$. Since $1/\eta \le 1/a + 1/b$ and $\eta \le b$,

$$\begin{aligned}\frac{D^2}{\eta T} + \frac{\eta}{B} &\le \frac{D^2}{T}\left(1 + \frac{\sqrt{T}}{D\sqrt{B}}\right) + \frac{D}{\sqrt{BT}} \\ &= \frac{D^2}{T} + \frac{2D}{\sqrt{BT}} \\ &\le \frac{d}{T} + \frac{2\sqrt{d}}{\sqrt{BT}},\end{aligned}$$

which proves the theorem; the identity

$$\frac{d}{T} + \frac{2\sqrt{d}}{\sqrt{BT}} = \frac{dB}{N} + 2\sqrt{\frac{d}{N}}$$

follows immediately from $N = BT$. □

H.1 Fixed-batch SignGD vs. SGD

Under $N = BT$, the SignSGD bound contains both a minimum and $\log T = \log(N/B)$. For a simple conservative comparison, we use its Jensen branch and $\log T \le \log N$, defining

$$\Delta_{\mathrm{SGD}}(B) := \frac{dB}{N} + 2\sqrt{\frac{d}{N}}, \qquad \Delta_{\mathrm{Sign}}(B) := \frac{2B(\log d)^2 \log N}{N} + 4\sqrt{\frac{d\log d}{B}}.$$

The following lemma formalizes this fixed-batch comparison.

Lemma H.2 (Crossing batch size for SignSGD vs. SGD)**.** *Suppose d is sufficiently large, $N \ge d$, and $(\log d)^2 \log N = o(d)$. There is a threshold batch size $B_{\mathrm{cross}} = \Theta\left(\left(\frac{N^2 \log d}{d}\right)^{1/3}\right)$ such that $B_{\mathrm{cross}} \in [1, N/2]$ and*

$$\Delta_{\mathrm{Sign}}(B) < \Delta_{\mathrm{SGD}}(B) \iff B > B_{\mathrm{cross}}.$$

Proof. Let $\ell = \log d$, $K = \log N$, and define the difference between the two upper-bound curves

$$H(B) := \Delta_{\mathrm{SGD}}(B) - \Delta_{\mathrm{Sign}}(B) = \frac{(d - 2\ell^2 K)B}{N} + 2\sqrt{\frac{d}{N}} - 4\sqrt{\frac{d\ell}{B}}.$$

Since $\ell^2 K = o(d)$,

$$H'(B) = \frac{d - 2\ell^2 K}{N} + 2\sqrt{d\ell}\,B^{-3/2} > 0.$$

Thus H is strictly increasing. Since $N \ge d$,

$$H(1) \le 3 - 4\sqrt{d\ell} < 0$$

for large d, while

$$H(N/2) = \frac{d - 2\ell^2 K}{2} + 2\sqrt{\frac{d}{N}} - 4\sqrt{\frac{2d\ell}{N}} > 0$$

for large d. Hence there is a unique root $B_{\text{cross}} \in [1, N/2]$, and $\Delta_{\text{Sign}}(B) < \Delta_{\text{SGD}}(B)$ exactly when $H(B) > 0$, equivalently when $B > B_{\text{cross}}$.

It remains to identify the scale of this crossing. Define

$$\widehat{B}_{\text{cross}} := \left(\frac{16N^2\ell}{d}\right)^{1/3}, \qquad A := \frac{d\widehat{B}_{\text{cross}}}{N} = 4\sqrt{\frac{d\ell}{\widehat{B}_{\text{cross}}}}.$$

For any fixed $r > 0$,

$$\frac{H(r\widehat{B}_{\text{cross}})}{A} = r - r^{-1/2} - \frac{2\ell^2 K}{d}r + \frac{2\sqrt{d/N}}{d\widehat{B}_{\text{cross}}/N}.$$

The last two terms are $o(1)$ because $\ell^2 K/d \to 0$ and

$$\frac{2\sqrt{d/N}}{d\widehat{B}_{\text{cross}}/N} = \frac{2}{16^{1/3}d^{1/6}N^{1/6}\ell^{1/3}} = o(1).$$

Therefore $H(r\widehat{B}_{\text{cross}})/A = r - r^{-1/2} + o(1)$. Since $r - r^{-1/2}$ has its unique zero at $r = 1$, the unique root satisfies

$$B_{\text{cross}} = (1 + o(1))\widehat{B}_{\text{cross}} = (1 + o(1))\left(\frac{16N^2 \log d}{d}\right)^{1/3}.$$

□

Post-crossing regimes. Define $B_{\text{sat}} := \frac{(4N^2 d)^{1/3}}{\log d\,(\log N)^{2/3}}$. This is the scale where the two terms in Δ_{Sign} are equal, since $4\sqrt{\frac{d\log d}{B_{\text{sat}}}} = \frac{2B_{\text{sat}}(\log d)^2 \log N}{N}$. Also, Lemma H.2 gives $B_{\text{cross}} = \Theta\left(\left(\frac{N^2 \log d}{d}\right)^{1/3}\right)$, so $\frac{B_{\text{sat}}}{B_{\text{cross}}} = \Theta\left(\left(\frac{d}{(\log d)^2 \log N}\right)^{2/3}\right) \to \infty$. Now suppose $B_{\text{cross}} < B \ll B_{\text{sat}}$. Since $B \geq B_{\text{cross}}$,

$$\frac{\sqrt{d/N}}{dB/N} \leq \frac{\sqrt{d/N}}{dB_{\text{cross}}/N} = \Theta\left(\frac{(d/N)^{1/6}}{(\log d)^{1/3}}\right) = o(1),$$

$$\frac{2B(\log d)^2 \log N/N}{4\sqrt{d\log d/B}} = \left(\frac{B}{B_{\text{sat}}}\right)^{3/2} = o(1).$$

Hence $\Delta_{\text{SGD}}(B) = \Theta(dB/N)$ and $\Delta_{\text{Sign}}(B) = \Theta(\sqrt{d\log d/B})$, so

$$\frac{\Delta_{\text{SGD}}(B)}{\Delta_{\text{Sign}}(B)} = \Theta\left(\left(\frac{B}{B_{\text{cross}}}\right)^{3/2}\right).$$

Finally, if $B_{\text{sat}} \ll B \leq N/2$, then $\Delta_{\text{SGD}}(B) = \Theta(dB/N)$ and $\Delta_{\text{Sign}}(B) = \Theta(B(\log d)^2 \log N/N)$, hence

$$\frac{\Delta_{\text{SGD}}(B)}{\Delta_{\text{Sign}}(B)} = \Theta\left(\frac{d}{(\log d)^2 \log N}\right).$$

I Stochastic Softmax Unigram Analysis Proofs

Proof of Corollary 5.2. We apply Theorem 5.1 with respect to $\|\cdot\|_\infty$, whose dual norm is $\|\cdot\|_1$. By Lemma B.2,

$$L_{\|\cdot\|_\infty}(f_{\text{uni}}) = 1.$$

By Lemma B.3, the minimum ℓ_∞-norm solution has norm $\log(d)/2$. Thus, with $\lambda = 2/\log d$ and $\theta_0 = 0$, we have

$$D = 1/\lambda = \log(d)/2 \qquad \text{and} \qquad 1 + \lambda D = 2.$$

Also, by Lemma B.4, for all sufficiently large d,

$$\sigma_\infty \le \min\left\{2, 4\sqrt{\frac{d}{B\log d}}\right\}.$$

Applying Theorem 5.1 yields

$$\begin{aligned}\mathbb{E}[f_{\mathrm{uni}}(\theta_T) - f^\star_{\mathrm{uni}}] &\le \frac{2(1+\lambda D)^2 \log T}{\lambda^2 T} + \frac{2\sigma_\infty}{\lambda} \\ &\le \frac{2(\log d)^2 \log T}{T} + \min\left\{2\log d, 4\sqrt{\frac{d\log d}{B}}\right\}.\end{aligned}$$

□

Proof of Corollary 5.3. Let $\ell := \log d$ and

$$T_0 := \frac{8\ell^2}{\epsilon}\log\left(\frac{8\ell^2}{\epsilon}\right).$$

For sufficiently large d, the map $T \mapsto \log T/T$ is decreasing at T_0, and

$$\frac{2\ell^2 \log T_0}{T_0} \le \frac{\epsilon}{2}.$$

Indeed, with $A := 4\ell^2/\epsilon$, we have $T_0 = 2A\log(2A)$ and $\log T_0 \le 2\log(2A)$. Moreover, the batch-size assumption gives

$$N_B(d) \le 4\sqrt{\frac{d\ell}{B}} \le \frac{\epsilon}{2}.$$

Applying Corollary 5.2 at $T = \lceil T_0 \rceil$ proves the claim. □

Proof of Corollary 5.4. We apply Theorem 5.1 with respect to $\|\cdot\|_2$. By Lemma B.2,

$$L_{\|\cdot\|_2}(f_{\mathrm{uni}}) = \frac{1}{2}.$$

By Lemma B.3, the minimum ℓ_2-norm solution has norm of order $\sqrt{d}$, and the stated choice $\lambda = 1/\sqrt{d}$ gives

$$D = \max\{\|\theta_0\|_2, 1/\lambda\} = \sqrt{d}$$

for $\theta_0 = 0$. Finally, Lemma B.4 gives the dual-norm stochastic-gradient error

$$\sigma_2 \le \frac{1}{\sqrt{B}}.$$

Substituting these bounds into Theorem 5.1,

$$\begin{aligned}\mathbb{E}[f_{\mathrm{uni}}(\theta_T) - f^\star_{\mathrm{uni}}] &\le \frac{(1+\lambda D)^2 \log T}{\lambda^2 T} + \frac{2\sigma_2}{\lambda} \\ &= \frac{4d\log T}{T} + 2\sqrt{\frac{d}{B}}.\end{aligned}$$

□

Proof of Corollary 5.5. Let $\ell := \log d$ and $K := \log N$. Under $N = BT$, define

$$F_\infty(T) := \frac{2\ell^2 \log T}{T} + \min\left\{2\log d, 4\sqrt{\frac{Td\log d}{N}}\right\}, \qquad 2 \le T \le N.$$

By Corollary 5.2, the error is upper bounded by $F_\infty(T)$. Since $\log T \le K$, define the upper envelope

$$\widetilde{F}_\infty(T) := \frac{2\ell^2 K}{T} + 4\sqrt{\frac{Td\ell}{N}}.$$

Optimizing this envelope gives

$$\widetilde{F}'_\infty(T) = 2\sqrt{\frac{d\ell}{NT}} - \frac{2\ell^2 K}{T^2}, \qquad T^\star_\infty = \ell K^{2/3}\left(\frac{N}{d}\right)^{1/3}.$$

When $N \geq d$ and d is large, this choice lies in $[2, N]$. Therefore, choosing

$$T = T^\star_\infty, \qquad B = \frac{N}{T^\star_\infty} = \frac{(dN^2)^{1/3}}{\ell K^{2/3}},$$

yields

$$\begin{aligned} F_\infty(T^\star_\infty) &\leq \widetilde{F}_\infty(T^\star_\infty) \\ &= \frac{6d^{1/3}\ell K^{1/3}}{N^{1/3}}. \end{aligned}$$

This proves the claimed result. □

Proof of Corollary 5.6. Let $K := \log N$. Since $\log T \leq K$, Corollary 5.4 is upper bounded by

$$\widetilde{F}_2(T) := \frac{4dK}{T} + 2\sqrt{\frac{dT}{N}}, \qquad 2 \leq T \leq N.$$

Its derivative is

$$\widetilde{F}'_2(T) = \sqrt{\frac{d}{NT}} - \frac{4dK}{T^2}.$$

The stationary point satisfies

$$T^{3/2} = 4K\sqrt{dN},$$

and hence

$$T^\star_2 = 2^{4/3}K^{2/3}(dN)^{1/3}.$$

This is the unique minimizer, and $T^\star_2 \geq 2$. Moreover,

$$T^\star_2 \leq N \iff 2^{4/3}K^{2/3}(dN)^{1/3} \leq N \iff 16dK^2 \leq N^2,$$

which follows from $N \geq 4\sqrt{d}\log N$. Therefore,

$$\widetilde{F}_2(T^\star_2) = \frac{3 \cdot 2^{2/3}d^{2/3}K^{1/3}}{N^{1/3}}.$$

The corresponding batch size is

$$B^\star_2 = \frac{N}{T^\star_2} = \frac{N^{2/3}}{2^{4/3}d^{1/3}K^{2/3}},$$

which gives the claimed result. □

J Stochastic Momentum Softmax Unigram Analysis

Theorem J.1 (Stochastic momentum SignGD rate under a fixed sample budget)**.** *Consider optimizing f_{uni} using the single-sample objective $\widehat{f}_{\mathrm{uni},1}$ with large d, total sample budget N, and initializations $\theta_0 = 0$ and $v_0 = 0$. Suppose the momentum ρ is chosen as in Corollary G.4. Then stochastic momentum SignGD with weight decay $\lambda_\infty = 2/\log d$ and learning rate $\eta = \frac{\log d \log N}{2N}$ satisfies*

$$\mathbb{E}[f_{\mathrm{uni}}(\theta_N) - f^\star_{\mathrm{uni}}] = O\left(d^{1/3}(\log d)^{4/3}\left(\frac{\log N}{N}\right)^{1/3}\right).$$

Proof. Let $u = (1/d)\mathbf{1}$ be the uniform distribution over the vocabulary. At initialization $\theta_0 = 0$, we have $\mathrm{softmax}(\theta_0) = u$ and

$$\delta_0 = f_{\mathrm{uni}}(\theta_0) - f^\star_{\mathrm{uni}} = \mathrm{KL}(p\|u) \le \log d.$$

In the single-sample regime $B = 1$, the covariance computation in the proof of Lemma B.4 gives

$$\mathbb{E}\left[\left\|\nabla \widehat{f}_{\mathrm{uni},1}(\theta;\xi) - \nabla f_{\mathrm{uni}}(\theta)\right\|_2^2\right] \le 1,$$

so in Theorem G.3 we may take $\nu_2 = 1$ and set $T = N$.

Now consider stochastic momentum SignGD, i.e. NSD with respect to $\|\cdot\|_\infty$. Here the dual norm is $\|\cdot\|_1$, so

$$\kappa = \sup_{z\neq 0} \frac{\|z\|_1}{\|z\|_2} = \sqrt{d}.$$

By Lemmas B.2 and B.3, $L = 1$ and the choice $\lambda_\infty = 2/\log d$ is optimal for the minimum ℓ_∞-norm solution. Thus $D = 1/\lambda_\infty = \log d/2$ and $1 + \lambda_\infty D = 2$. Since $v_0 = 0$,

$$\mathbb{E}[\|\nabla f_{\mathrm{uni}}(\theta_0) - v_0\|_1] = \|u - p\|_1 \le 2.$$

The horizon condition becomes

$$T = \tilde{\Omega}\left(\frac{1 + \lambda_\infty D}{\lambda_\infty \kappa \nu_2}\right) = \tilde{\Omega}\left(\frac{\log d}{\sqrt{d}}\right).$$

For sufficiently large d, this lower bound is at most a constant, so it is automatically satisfied in the regime considered here. Applying Corollary G.4 with $T = N$,

$$\delta_N \le \left[\log d + \frac{(\log d)^2}{2}\right]\frac{\log N}{N} + \left[4d^{1/3}(\log d)^{1/3} + 3d^{1/3}(\log d)^{4/3}\right]\left(\frac{\log N}{N}\right)^{1/3}.$$

The stepsize is $\eta = \log N/(\lambda_\infty N) = \log d \log N/(2N)$, completing the proof. □

Theorem J.2 (Stochastic momentum normalized-GD rate under a fixed sample budget)**.** *Consider optimizing f_{uni} using the single-sample objective $\widehat{f}_{\mathrm{uni},1}$ with large d, total sample budget $N = \tilde{\Omega}(\sqrt{d})$, and initializations $\theta_0 = 0$ and $v_0 = 0$. Suppose the momentum ρ is chosen as in Corollary G.4. Then stochastic momentum normalized GD with weight decay $\lambda_2 = 1/\sqrt{d}$, learning rate $\eta = \frac{\sqrt{d}\log N}{N}$, and batch size $B = 1$ satisfies*

$$\mathbb{E}[f_{\mathrm{uni}}(\theta_N) - f^\star_{\mathrm{uni}}] = O\left(d^{2/3}\left(\frac{\log N}{N}\right)^{1/3}\right).$$

Proof. Let $u = (1/d)\mathbf{1}$ be the uniform distribution over the vocabulary. At initialization $\theta_0 = 0$, we have $\mathrm{softmax}(\theta_0) = u$ and

$$\delta_0 = f_{\mathrm{uni}}(\theta_0) - f^\star_{\mathrm{uni}} = \mathrm{KL}(p\|u) \le \log d.$$

In the single-sample regime $B = 1$, the covariance computation in the proof of Lemma B.4 gives

$$\mathbb{E}\left[\left\|\nabla \widehat{f}_{\mathrm{uni},1}(\theta;\xi) - \nabla f_{\mathrm{uni}}(\theta)\right\|_2^2\right] \le 1,$$

so in Theorem G.3 we may take $\nu_2 = 1$ and set $T = N$.

Now consider stochastic momentum normalized GD, i.e. NSD with respect to $\|\cdot\|_2$. By Lemma B.2, $L = 1/2$. For large d, Lemma B.3 justify the choice $\lambda_2 = 1/\sqrt{d}$, and with $\theta_0 = 0$ we have $D = 1/\lambda_2 = \sqrt{d}$ and $1 + \lambda_2 D = 2$. The dual norm is again $\|\cdot\|_2$, so $\kappa = 1$. Since $v_0 = 0$,

$$\mathbb{E}[\|\nabla f_{\mathrm{uni}}(\theta_0) - v_0\|_2] = \|u - p\|_2 \le 1.$$

The horizon condition in Corollary G.4 is

$$N = \tilde{\Omega}\left(\frac{1 + \lambda_2 D}{\lambda_2 \kappa \nu_2}\right) = \tilde{\Omega}\left(\sqrt{d}\right),$$

which is exactly the stated assumption. Substituting the above constants into Corollary G.4 with $T = N$ gives

$$\delta_N \le [\log d + d]\frac{\log N}{N} + \left[4d^{1/6} + 5d^{2/3}\right]\left(\frac{\log N}{N}\right)^{1/3}.$$

The stepsize in Corollary G.4 is $\eta = \log N/(\lambda_2 N) = \sqrt{d}\log N/N$. □

K Challenges in Extending the Analysis to Adam

Although our theory so far focuses on SignGD, Adam exhibits the same qualitative advantage over Euclidean methods in both the deterministic softmax unigram experiment in Figure 1 and the transformer experiment in Figure 2. This motivates asking whether existing theory for Adam can explain its performance on the softmax unigram model. One proposed explanation is that the commonly assumed ℓ_2 smoothness is too coarse to capture the optimization benefits of structured preconditioning (Xie et al., 2025a;b). This line of work generalizes the notion of smoothness to adaptive methods with structured preconditioners, including e.g., Adam, blockwise Adam, and one-sided Shampoo. Each preconditioner structure induces a corresponding adaptive geometry. More specifically, let $\mathcal{H}$ be a family of positive semidefinite matrices encoding the preconditioner structure. The corresponding adaptive smoothness is the infimum of the traces of matrices in $\mathcal{H}$ that dominate the loss Hessian:

$$\mathcal{L}_{\mathcal{H}}(f) := \inf_{\mathbf{H}\in\mathcal{H}:\forall x, \nabla^2 f(x)\preceq\mathbf{H}} \mathrm{Tr}(\mathbf{H}).$$

For Adam, the admissible preconditioners are diagonal, so we specialize $\mathcal{H}$ to the family of diagonal positive semidefinite matrices. This gives the following definition of diagonal adaptive smoothness.

Definition K.1 (Diagonal Adaptive Smoothness)**.** *The diagonal adaptive smoothness of a function f is defined as the minimal trace of the diagonal matrix that dominates the Hessian:*

$$\mathcal{L}_{\mathrm{diag}}(f) = \min_{\mathrm{diag}\ \mathbf{H}:\forall x, \nabla^2 f(x)\preceq\mathbf{H}} \mathrm{Tr}(\mathbf{H}). \tag{4}$$

The convergence rates in this framework are controlled by the product of the adaptive smoothness and the squared radius of the optimization domain under the norm associated with the preconditioner family (Xie et al., 2025b). For Adam, the diagonal preconditioner family gives smoothness $\mathcal{L}_{\mathrm{diag}}(f)$ and the ℓ_∞ domain norm. For GD, the isotropic family gives smoothness $dL_{\|\cdot\|_2}(f)$ and the normalized Euclidean domain norm $\|\cdot\|_2/\sqrt{d}$. Thus, when the domain radii under these two norms are comparable, the theory predicts that Adam is faster than GD when $\mathcal{L}_{\mathrm{diag}}(f) \ll dL_{\|\cdot\|_2}(f)$.

We next evaluate the diagonal adaptive smoothness of f_{uni}. The proof of the following lemma is in Appendix K.1.

Lemma K.1. *The diagonal adaptive smoothness of the softmax unigram objective f_{uni} defined in Equation* (1) *satisfies $\mathcal{L}_{\mathrm{diag}}(f_{\mathrm{uni}}) \geq d/2$.*

By Lemma B.2, $L_{\|\cdot\|_2}(f_{\mathrm{uni}}) = 1/2$, so Lemma K.1 implies $\mathcal{L}_{\mathrm{diag}}(f_{\mathrm{uni}}) \geq dL_{\|\cdot\|_2}(f_{\mathrm{uni}})$. Moreover, Lemma B.3 gives $\|\theta_\star\|_\infty^2 = (\log d)^2/4$ for the minimum-ℓ_∞-norm solution and $\|\theta_\star\|_2^2 = \Theta(d)$ for the minimum-ℓ_2-norm solution. Therefore, the quantity controlling the Adam upper bound is at least of order $d(\log d)^2$, while the corresponding quantity for the GD upper bound is of order d. In particular, the Adam upper bound has at least the same polynomial dependence on d as the GD upper bound and is not smaller. Thus, the existing adaptive-smoothness theory in (Xie et al., 2025b) cannot explain the empirical advantage of Adam in Figure 1.

Indeed, this is a key distinction between the softmax unigram model and potentially simpler diagonal-quadratic setups. For instance, consider the simple diagonal quadratic example proposed in Kunstner et al. (2024), $g(\theta) = \sum_{k=1}^d p_k\theta_k^2$, where $p \in \mathbb{R}^d$ is the Zipfian distribution of Definition 3.1. Its Hessian is $\nabla^2 g(\theta) = 2\,\mathrm{diag}(p)$, so $\mathcal{L}_{\mathrm{diag}}(g) = 2$, whereas $dL_{\|\cdot\|_2}(g) = 2d/H_d = \Theta(d/\log d)$. Thus, adaptive smoothness is much smaller than $dL_{\|\cdot\|_2}(g)$ for large d, predicting faster convergence of Adam over SGD for diagonal quadratics with heavy-tailed class imbalance. This separation does not hold for the softmax unigram model, highlighting the difference between our setup and quadratic models. We leave the theoretical explanation of why Adam outperforms GD empirically on this softmax unigram model to future work.

K.1 Proof of the Diagonal Adaptive Smoothness Bound

Proof of Lemma K.1. Let $\mathbf{H} = \mathrm{diag}(\alpha_1, \ldots, \alpha_d)$ be any diagonal positive semidefinite matrix such that

$$\nabla^2 f_{\mathrm{uni}}(\theta) \preceq \mathbf{H} \qquad \text{for all } \theta \in \mathbb{R}^d.$$

We will show that $\mathrm{Tr}(\mathbf{H}) \geq d/2$.

Fix any pair $i < j$. For $s > 0$, define $\theta^{(s,i,j)} \in \mathbb{R}^d$ by

$$\theta_i^{(s,i,j)} = \theta_j^{(s,i,j)} = s, \qquad \theta_k^{(s,i,j)} = -s \quad \text{for } k \notin \{i, j\}.$$

Then

$$[\mathrm{softmax}(\theta^{(s,i,j)})]_i = [\mathrm{softmax}(\theta^{(s,i,j)})]_j = \frac{1}{2 + (d-2)e^{-2s}},$$

while for $k \notin \{i, j\}$,

$$[\mathrm{softmax}(\theta^{(s,i,j)})]_k = \frac{e^{-2s}}{2 + (d-2)e^{-2s}}.$$

Hence, as $s \to \infty$,

$$\mathrm{softmax}(\theta^{(s,i,j)}) \to \frac{1}{2}e_i + \frac{1}{2}e_j.$$

Since the Hessian map

$$q \mapsto \mathrm{diag}(q) - qq^\top$$

is continuous in q, we obtain

$$\nabla^2 f_{\mathrm{uni}}(\theta^{(s,i,j)}) \to H^{(i,j)},$$

where $H^{(i,j)}$ is the matrix whose only nonzero entries lie in the (i, j) principal block:

$$H^{(i,j)}_{\{i,j\},\{i,j\}} = \begin{pmatrix} \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} \end{pmatrix}.$$

Because $\mathbf{H} - \nabla^2 f_{\mathrm{uni}}(\theta^{(s,i,j)}) \succeq 0$ for every s, and the cone of positive semidefinite matrices is closed, we also have

$$\mathbf{H} - H^{(i,j)} \succeq 0.$$

Now test this PSD inequality with the vector $x = e_i - e_j$. Then

$$x^\top \mathbf{H} x = \alpha_i + \alpha_j, \qquad x^\top H^{(i,j)} x = 1.$$

Therefore,

$$\alpha_i + \alpha_j \geq 1 \qquad \text{for every } 1 \leq i < j \leq d.$$

Summing this inequality over all pairs (i, j) gives

$$\sum_{1 \leq i < j \leq d} (\alpha_i + \alpha_j) \geq \binom{d}{2}.$$

Each coordinate α_k appears in exactly $d - 1$ such pairs, so

$$(d-1)\,\mathrm{Tr}(\mathbf{H}) = \sum_{1 \leq i < j \leq d} (\alpha_i + \alpha_j) \geq \frac{d(d-1)}{2}.$$

Hence,

$$\mathrm{Tr}(\mathbf{H}) \geq \frac{d}{2}.$$

Since this holds for every diagonal PSD matrix $\mathbf{H}$ dominating the Hessian of f_{uni}, we conclude that

$$\mathcal{L}_{\mathrm{diag}}(f_{\mathrm{uni}}) \geq \frac{d}{2}.$$

□

L Experimental Details for Section 3

This appendix provides the model, data, optimization, and reproducibility details for the experiments in Section 3 and Figures 1–3. Unless stated otherwise, SGD and SignSGD use neither momentum nor weight decay, and all models are trained using the cross-entropy objective.

Data and preprocessing. The stochastic language-modeling experiments use the Penn Treebank (PTB) corpus and the basic-English word tokenizer. The vocabulary is constructed from the PTB training split and contains $d = 9{,}922$ tokens, including the unknown-token symbol. To reduce computational cost, the experiments denoted by `TinyPTB` use the PTB validation split for training and the PTB test split for evaluation, while retaining the vocabulary constructed from the full training split.

A mini-batch contains $1{,}024$ parallel token sequences. The transformer, frozen-transformer, and bigram experiments use a sequence length of 35, corresponding to at most $35{,}840$ next-token targets per optimization step. The collapsed unigram experiment uses a sequence length of 32, corresponding to at most $32{,}768$ targets per step. The final batch in an epoch can contain fewer targets.

Models and training horizons. Table 1 summarizes the sequence of model simplifications. "Trainable parameters" counts only parameters updated by the optimizer.

Experiment	Model	Mini-batch	Horizon and seeds
Figure 2	One-layer causal transformer with embedding dimension 512, 8 attention heads, feed-forward width 256, and dropout 0.2. The embedding, transformer layer, and output classifier are all trained. The model has approximately 11.49M trainable parameters.	$1{,}024 \times 35$	100 epochs; seeds $0, 1, 2$ for the optimizer comparison and seed 0 for the per-frequency-group runs.
Figure 3(a)	One-layer transformer with embedding dimension 256, one attention head, feed-forward width 128, and dropout 0.2. The token embedding and complete transformer encoder are frozen; only the softmax classification head is trained. The model has approximately 5.42M parameters, of which 2.55M are trainable.	$1{,}024 \times 35$	100 epochs; seeds $0, 1, 2$.
Figure 3(b)	Bigram model obtained by removing attention. It consists of a fixed $9{,}922 \times 256$ token-embedding matrix followed by a trainable, bias-free $256 \times 9{,}922$ output classifier. It has approximately 5.08M parameters, of which 2.54M are trainable.	$1{,}024 \times 35$	100 epochs; seeds $0, 1, 2$.
Figure 3(c,d)	Collapsed softmax unigram model. Every input token is replaced by the same index, leaving one trainable vector of $9{,}922$ logits. The logits are initialized to zero.	$1{,}024 \times 32$	200 epochs; seeds $0, 1, 2$.
Figure 1	Deterministic softmax unigram model with $d = 10{,}000$ trainable logits, initialized to zero, and target probabilities $p_i \propto i^{-1}$.	Full batch	600 gradient steps; seed 0.

Table 1: Model and training details for the experiments in Section 3.

Initialization. For Figure 2, the token embeddings are initialized independently from $\mathcal{N}(0, 0.02^2)$, while the output classifier weights and biases are initialized to zero. In Figure 3(a), the transformer weights and embeddings are initialized with standard deviation 0.02 before the embedding and encoder are frozen. In Figure 3(b), the fixed embedding and initial classifier weights are sampled from $\mathcal{N}(0, 0.02^2)$. The collapsed unigram logits in Figure 3(c,d) and the deterministic logits in Figure 1 are initialized to zero.

Learning-rate tuning. Learning rates are tuned independently for each optimizer using logarithmic grids. Let

$$\mathcal{G}(a, b, \delta) := \left\{10^a, 10^{a+\delta}, \ldots, 10^b\right\}.$$

For each optimizer, we select the learning rate giving the lowest training loss at the tuning horizon. No learning-rate schedule is used. The grids and selected values are reported in Table 2. Numerical approximations are included for convenience.

For the per-frequency-group panels of Figure 2, SGD uses $10^{-3/4} \approx 0.178$ and SignSGD uses $10^{-11/4} \approx 0.00178$. The unigram learning-rate grid was run for 500 epochs; the selected configura-

Experiment	Optimizer	Search grid	Selected learning rate
Figure 2	SGD	$\mathcal{G}(-3, 0, 1/4)$	$10^{-1/2} \approx 0.316$
	Adam	$\mathcal{G}(-4, -2, 1/4)$	$10^{-11/4} \approx 0.00178$
	normalized SGD	$\mathcal{G}(-2, 1, 1/4)$	$10^{1/4} \approx 1.78$
	SignSGD	$\mathcal{G}(-5, -2, 1/4)$	$10^{-5/2} \approx 0.00316$
Figure 3(a)	SGD	$\mathcal{G}(-3, 0, 1/2)$	$10^{-1/2} \approx 0.316$
	SignSGD	$\mathcal{G}(-5, -2, 1/2)$	$10^{-5/2} \approx 0.00316$
Figure 3(b)	SGD	$\mathcal{G}(0, 3, 1/2)$	$10^{3/2} \approx 31.6$
	SignSGD	$\mathcal{G}(-3, -1, 1/2)$	$10^{-2} = 0.01$
Figure 3(c,d)	SGD	$\mathcal{G}(-1, 2, 1/2)$	$10^{3/2} \approx 31.6$
	SignSGD	$\mathcal{G}(-4, -1, 1/2)$	$10^{-2} = 0.01$
Figure 1	GD	$\mathcal{G}(0, 3, 1/4)$	$10^{5/4} \approx 17.8$
	Adam	$\mathcal{G}(-2, -1, 1/4)$	$10^{-3/2} \approx 0.0316$
	normalized GD	$\mathcal{G}(-1, 1, 1/4)$	$10^{-1/4} \approx 0.562$
	SignGD	$\mathcal{G}(-3, -1, 1/4)$	$10^{-2} = 0.01$

Table 2: Per-optimizer learning-rate tuning for the Section 3 experiments. For Adam, we use $\beta_1 = 0$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$.

tions were then re-run for the 200 epochs displayed in Figure 3(c,d). All other stochastic grids use the same horizon as their corresponding plotted experiment.

Seeds and curve selection. For Figure 2(a), every learning rate is run with seeds 0, 1, and 2. For each optimizer, the learning rate is selected by the median final training cross-entropy across the three seeds; the plotted curve is the pointwise median and the shaded region spans the minimum and maximum across seeds. The corresponding per-class SGD and SignSGD runs in Figure 2(b,c) use seed 0. For every panel of Figure 3, we run seeds 0, 1, and 2 for each optimizer. The plotting procedure selects, separately for each optimizer, the seed with the lowest final training cross-entropy. The displayed seeds are: seed 1 for SGD and seed 0 for SignSGD in Figure 3(a), and seed 0 for both optimizers in Figure 3(b–d). The deterministic softmax-unigram experiment uses seed 0, although the zero initialization and exact gradients make the optimization trajectory deterministic.

Per-frequency-group losses. For Figure 2(b,c), token classes are ordered by their empirical PTB frequency and partitioned into ten groups, each containing approximately one tenth of the empirical target mass. We compute the conditional cross-entropy for each token and average it within each group. The per-class statistics are obtained from SGD and SignSGD runs with the same architecture and horizon as Figure 2(a), using the per-frequency-group learning rates reported above and seed 0.

For Figure 1(c,d), the 10,000 classes are similarly partitioned into ten groups of equal target-probability mass. If $q = \mathrm{softmax}(\theta)$, the contribution assigned to token i is

$$p_i \log \frac{p_i}{q_i} - p_i + q_i.$$

These terms are non-negative and sum exactly to $\mathrm{KL}(p\|q)$. Each curve reports the sum of these contributions over one frequency group.

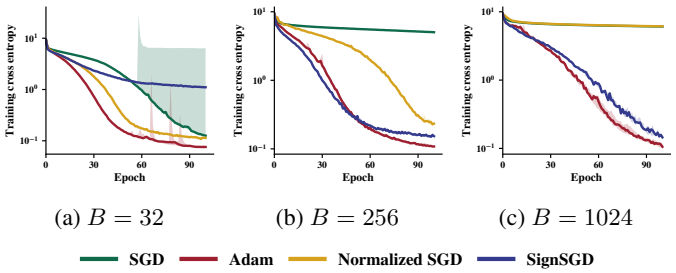
L.1 Batch-Size Experiments

This subsection provides the model, data, optimization, and curve-selection details for Figures 5 and 7.

Shared optimizer settings. We compare SGD, Adam, normalized SGD, and SignSGD. SGD, normalized SGD, and SignSGD use zero momentum. No optimizer uses weight decay or a learning-rate schedule. Adam uses $\beta_1 = 0$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. At each iteration, normalized SGD divides the complete gradient vector by its global ℓ_2 norm, whereas SignSGD applies the sign operation coordinate-wise. All experiments use 32-bit floating-point arithmetic and PyTorch 2.5.1.

Batch size	SGD	Adam	Normalized SGD	SignSGD
64	10 (2)	$10^{-3/2}$ (2)	$10^{1/2}$ (2)	$10^{-3/2}$ (2)
1024	$10^{3/2}$ (0)	$10^{-3/2}$ (0)	10 (2)	10^{-2} (0)
32768	$10^{3/2}$ (0)	$10^{-3/2}$ (0)	1 (0)	$10^{-3/2}$ (1)

Table 3: Selected learning rates for the stochastic softmax unigram experiment. Selected seeds are shown in parentheses.



(a) $B = 32$ (b) $B = 256$ (c) $B = 1024$

Figure 7: Batch-size comparison for a two-layer transformer trained on TinyPTB. The panels show training cross-entropy for SGD, Adam, normalized SGD, and SignSGD; curves are medians over three seeds and shaded regions show the minimum-to-maximum range.

Stochastic softmax unigram model (Figure 5). We use a softmax unigram model with $d = 10{,}000$ classes the Zipfian distribution in Definition 3.1.

At every optimization step, a mini-batch $\xi_1, \ldots, \xi_B \overset{\text{i.i.d.}}{\sim} \mathrm{Cat}(p)$ is sampled with replacement. The stochastic training loss is

$$-\frac{1}{B}\sum_{b=1}^{B} \log[q_\theta]_{\xi_b}.$$

The plotted metric is the exact population loss $\mathrm{KL}(p\|q_\theta)$, evaluated against the complete target distribution after every update.

We consider batch sizes $B \in \{64, 1024, 32768\}$. Every configuration is run for 300 optimization steps with seeds 0, 1, and 2. Thus, the methods are compared at a fixed iteration horizon, and a run with batch size B processes $300B$ sampled tokens.

For compactness, define the logarithmic learning-rate grid $\mathcal{G}(a, b, \delta) = \{10^a, 10^{a+\delta}, \ldots, 10^b\}$. The learning-rate grids are $\mathcal{G}(-1, 3, 1/2)$ for SGD, $\mathcal{G}(-3, 0, 1/2)$ for Adam, $\mathcal{G}(-2, 2, 1/2)$ for normalized SGD, and $\mathcal{G}(-3, 0, 1/2)$ for SignSGD.

For each optimizer and batch size, the displayed curve is the individual learning-rate and seed combination with the lowest population KL divergence after 300 steps. The selected learning rates are reported in Table 3, with the selected seed shown in parentheses.

Transformer language model (Figure 7). We use the TinyPTB setup described above. The vocabulary is constructed from the Penn Treebank training split using the basic-English word tokenizer and contains 9,922 tokens, including the unknown-token symbol. The PTB validation split is used for training and the PTB test split is used for evaluation.

The model is a two-layer causal transformer with trainable token embeddings, sinusoidal positional encodings, eight attention heads per layer, embedding dimension 512, feed-forward width 256, ReLU activations, post-layer normalization, an untied output classifier, and zero dropout. The model has approximately 12.80 million trainable parameters.

Embedding, feed-forward, and classifier weights are initialized independently from $\mathcal{N}(0, 0.02^2)$, and the corresponding biases are initialized to zero. Layer-normalization weights are initialized from $\mathcal{N}(1, 0.02^2)$, with zero biases. The attention projection weights use the default PyTorch initialization.

Batch size	SGD	Adam	Normalized SGD	SignSGD
32	10^{-1}	10^{-4}	$10^{-1/2}$	$10^{-7/2}$
256	10^{-1}	$10^{-7/2}$	1	10^{-3}
1024	10^{-1}	10^{-3}	10^{-1}	10^{-3}

Table 4: Selected learning rates for the TinyPTB transformer experiment.

We use $B \in \{32, 256, 1024\}$ parallel token sequences with sequence length 35. Thus, an optimization step contains at most $35B$ next-token targets. The final sequence chunk of an epoch can be shorter. Every configuration is trained for 100 epochs with seeds 0, 1, and 2 using the mean next-token cross-entropy objective. Because every run makes the same number of passes through the training corpus, larger sequence batches result in fewer optimizer updates per epoch.

For $B = 32$, the learning-rate grids are $\mathcal{G}(-2, 2, 1/2)$ for SGD and normalized SGD, $\mathcal{G}(-4, -1, 1/2)$ for SignSGD, and $\mathcal{G}(-5, -1, 1/2)$ for Adam. For $B = 256$, the grids are $\mathcal{G}(-2, 1, 1/2)$ for SGD and normalized SGD and $\mathcal{G}(-4, -1, 1/2)$ for Adam and SignSGD. For $B = 1024$, the grids are $\mathcal{G}(-3, -1, 1/2)$ for SGD and normalized SGD and $\mathcal{G}(-4, -1, 1/2)$ for Adam and SignSGD.

For each optimizer and batch size, the learning rate is selected by the lowest median training cross-entropy after 100 epochs across seeds 0, 1, and 2. The plotted curve is the pointwise median across the three seeds, and the shaded region spans their pointwise minimum and maximum. The selected learning rates are reported in Table 4.