Stochastic Gradient Descent on the Linear Bigram Model: Bias-Variance Scaling and Critical Batch Size

Zeyu Bian[*] and **Dmitriy Drusvyatskiy**[*] and **Tianhao Wang**[*]

Abstract

The critical batch size, beyond which data parallelism gives diminishing returns, is a key factor in language-model pretraining. Existing finite-time least-squares theory provides useful templates for predicting it, but does not directly give finite-vocabulary Zipf rates with explicit dependence on the vocabulary size. The linear bigram model, which fits a next-token transition matrix under squared loss with power-law token frequencies, provides a tractable setting for this regime. Its one-hot sampling structure also departs from prior vector analyses, and so far it has only been studied in the deterministic full-batch case. We give a finite-time analysis of mini-batch SGD on this model under power-law token distributions. Our main result is an exact, closed-form bias–variance decomposition of the expected loss, in which the bias term equals the loss of deterministic gradient descent and the variance term captures the cost of mini-batch noise. From this decomposition we obtain scaling laws for the bias and the variance, governed by a frequency cutoff that separates rows with enough effective updates from rarer under-trained rows. When this cutoff reaches the full vocabulary, the learning curve changes phase. Balancing bias and variance yields the scaling of the critical batch size. We confirm the predicted scaling on simulated bigrams and on bigram statistics estimated from OpenWebText.

1. Introduction

Large language models are trained in a regime where token frequency, optimizer choice, and data parallelism are inseparable. Token frequencies follow Zipf's law; more generally, we write $\pi_i \propto i^{-\alpha}$ for the power-law exponent α, with Zipfian text corresponding to $\alpha \approx 1$. This gives embedding and unembedding layers highly nonuniform updates: common rows are sampled many times, while rare rows move on a much slower clock. This row-wise imbalance is one reason SGD is inefficient for language-model embeddings, while Adam-type adaptive methods are often preferred in practice [3, 4, 10]. Recent theory captures the deterministic part of this story: in the linear bigram model, rare tokens produce slow active-vocabulary growth and make gradient descent especially costly in the Zipfian case $\alpha = 1$ [3].

Practical pretraining, however, uses mini-batches. Mini-batching adds a stochastic variance term on top of the deterministic rare-token bias. In our decomposition, bias is the squared error of the mean averaged iterate, while variance is the fluctuation induced by mini-batch sampling. The active-vocabulary cutoff separates the dominant scaling contributions: rows beyond the cutoff remain bias-limited, whereas activated rows contribute the usual mini-batch variance floor; in heavier-tailed regimes, the inactive tail also contributes to variance. This distinction is central for predicting the critical batch size (CBS), the point beyond which increasing data parallelism gives diminishing returns [1, 5, 6, 11].

[*] University of California, San Diego. Email: {zebian, ddrusvyatskiy, tianhaowang}@ucsd.edu

© .

General least-squares SGD theory provides sharp bias–variance decompositions and excess-risk bounds in terms of the full covariance spectrum [2, 7, 8, 12]. However, existing scaling-law analyses of stochastic least squares commonly turn these tools into clean power-law rates under trace-class or fast-decay regimes, and do not directly yield finite-V, normalized Zipf rates with explicit vocabulary dependence at the boundary $\alpha = 1$ or in heavier-tailed regimes $\alpha \leq 1$. The obstacle is not merely that the objective is quadratic. In standard linear regression, the second-moment dynamics involve fourth-moment operators such as $\mathbb{E}[xx^\top \Sigma xx^\top]$, so eigendirections are coupled through aggregate covariance or effective-dimension quantities. Thus current theory does not explain how mini-batch noise, finite-time optimization, and batch-size gains scale in the regime most relevant to language-model embeddings.

We address this gap in the linear bigram model of Kunstner and Bach [3]. The model maps the current one-hot token $x = e_i$ to a next-token prediction $W^\top x$, with token frequencies $\pi_i \propto i^{-\alpha}$ and exponent $\alpha > 0$. Its sampling structure closes the stochastic dynamics row-by-row, reducing the mean and second-moment evolution to one-dimensional recursions for each token row. These scalar recursions let us obtain an exact Polyak bias–variance formula and then split the risk at the active-vocabulary cutoff. We extend the deterministic picture to mini-batch SGD with Polyak averaging [9] and compute the finite-time stochastic risk exactly.

In this work, we make the following contributions:

1. **Exact bias–variance decomposition for averaged SGD** (Theorem 3). We derive a closed-form expression for the expected excess risk of the Polyak-averaged iterate, separating contributions from a deterministic bias term and a stochastic variance term.

2. **Two-phase scaling laws** (Theorem 4). We derive scaling laws for the bias and variance terms, organized by the active-vocabulary cutoff that reflects the effective dimension and the critical time at which this cutoff reaches the full vocabulary. Beyond this critical time, the bias and variance terms recover the statistically optimal $\Theta(1/T^2)$ and $\Theta(1/(bT))$ rates.

3. **Fixed-compute critical batch size** (Proposition 6). We derive the scaling of $b_{\text{crit}}(N)$ in V, α, and compute budget $N = bT$. Across all α, the compute dependence is $b_{\text{crit}}(N) \propto \sqrt{N}$, matching the empirically observed square-root law for language-model pretraining [11].

2. Model and SGD Dynamics

Let $[V] = \{1, \ldots, V\}$. Let $W^\star$ denote the population optimum, and define the iterate error $\Delta_t := W_t - W^\star$, with (i, j)-coordinate $\Delta_{ij,t}$. We assume power laws for the marginal token frequencies π_i and the sorted conditional next-token frequencies $\pi_{j|i}$:

Assumption 1 (Power-law marginals and conditionals) *Given vocabulary size V, for all $i, j \in [V]$, we have $\pi_i = z^{-1} i^{-\alpha}, \pi_{\rho_i(j)|i} = z^{-1} j^{-\alpha}$, where $z = \sum_{k=1}^{V} k^{-\alpha}$ is the power-law normalizing constant, and ρ_i is a sorting permutation that ranks the conditional frequencies in decreasing order.*

We remark that the assumption does not require all conditionals to follow the same distribution. It only states that the sorted frequencies follow a power law with the same exponent α. Our experiments on OpenWebText (Figure 3 and Figure 4) suggest this provides a reasonable coarse-grained approximation of real-world language data.

Lemma 2 (Population structure) *Let $W \in \mathbb{R}^{V\times V}$, and let x, y be one-hot vectors in $\mathbb{R}^V$. Then the linear predictor $f_W(x) = W^\top x$ trained with squared loss $\mathcal{L}(W) = \frac{1}{2}\mathbb{E}\|W^\top x - y\|_2^2$, where $\mathbb{P}(x = e_i) = \pi_i$ and $\mathbb{P}(y = e_j \mid x = e_i) = \pi_{j|i}$, has population minimizer $W^\star_{ij} = \pi_{j|i}$, diagonal second moment $D = \mathrm{diag}(\pi_1, \dots, \pi_V)$, and excess risk*

$$\mathcal{E}(W) = \mathcal{L}(W) - \mathcal{L}(W^\star) = \tfrac{1}{2}\sum\nolimits_{i=1}^{V} \pi_i \|W_{i,:} - W^\star_{i,:}\|_2^2. \tag{1}$$

Thus the problem is a quadratic objective in the row-vectorized parameter W, with Hessian matrix $D \otimes I_V$, and the loss decouples row-wise with weight π_i.

Let $\{W_t\}_{t=0}^T$ be iterates generated by online mini-batch SGD with batch size b: $W_{t+1} = W_t - \eta G_t$ where $G_t = \frac{1}{b}\sum_{\ell=1}^{b} x_t^{(\ell)}(W_t^\top x_t^{(\ell)} - y_t^{(\ell)})^\top$. We study the Polyak average $\bar{W}_T = T^{-1}\sum_{t=0}^{T-1} W_t$ with initialization at $W_0 = 0$.

3. Main Results

Our first result gives the bias–variance decomposition of the expected excess risk. The derivation has two steps. First, conditional on the current iterate W_t, the mini-batch gradient is unbiased, so the mean dynamics recover gradient descent: $\mathbb{E}[\Delta_{ij,t+1} \mid W_t] = \phi_i \Delta_{ij,t}$, where $\phi_i := 1 - \eta\pi_i$ is the per-row iterate contraction factor. Second, the mini-batch noise enters through the second moment of the iterate error, $\mathbb{E}[\Delta^2_{ij,t+1} \mid W_t] = a_i \Delta^2_{ij,t} + \{\text{noise floor}\}$, where $a_i := \phi_i^2 + \frac{\eta^2}{b}\pi_i(1-\pi_i)$ is the stochastic second-moment contraction factor.

Theorem 3 (Exact bias–variance decomposition) *Under Assumption 1, suppose $W_0 = 0$ and choose a stable step size $\eta < 2bz/(b+z-1)$ for batch size b. Define the per-row contraction factor $\phi_i = 1 - \eta\pi_i$ and second-moment contraction factor $a_i = \phi_i^2 + \frac{\eta^2}{b}\pi_i(1-\pi_i)$. Then for every $T \geq 1$,*

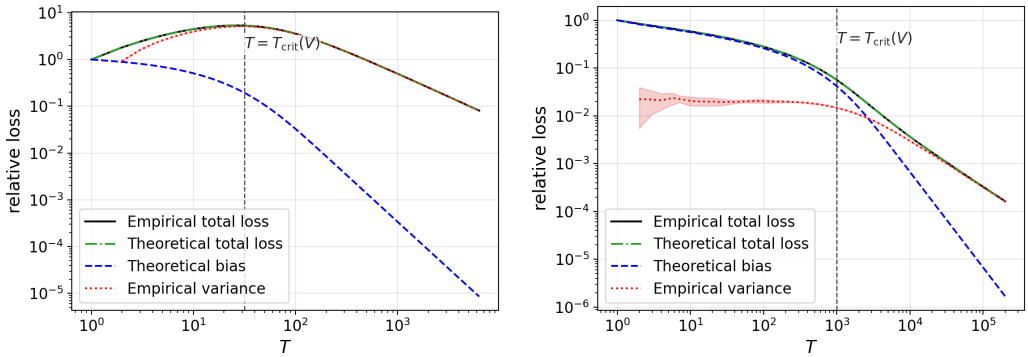
$$\mathbb{E}[\mathcal{E}(\bar{W}_T)] = \underbrace{\frac{\mathcal{E}(0)}{T^2}\sum_{i=1}^{V}\pi_i\Gamma_i(T)}_{\mathcal{E}_{\mathrm{bias}}(T)} + \underbrace{\frac{\mathcal{E}(0)}{T^2}\sum_{i=1}^{V}\pi_i\big(S_i^{(a)}(T) - \Gamma_i(T)\big) + \frac{\eta^2(1-2\mathcal{E}(0))}{2bT^2}\sum_{i=1}^{V}\frac{\pi_i^2}{1-a_i}S_i^{(1-a)}(T)}_{\mathcal{L}_{\mathrm{var}}(T)},$$

where $\Gamma_i(T) := \big(\frac{1-\phi_i^T}{1-\phi_i}\big)^2$, and $S_i^{(a)}(T)$ and $S_i^{(1-a)}(T)$ are the weighted temporal sums w.r.t. a_i:

$$S_i^{(a)}(T) := \sum_{s=0}^{T-1} w_{i,s}a_i^s, \quad S_i^{(1-a)}(T) := \sum_{s=0}^{T-1} w_{i,s}(1-a_i^s) \quad \textit{with weights } w_{i,s} := 1 + 2\sum_{k=1}^{T-1-s}\phi_i^k.$$

The advantage of the linear bigram model is that its learning dynamics separate cleanly across frequency modes. Each row-coordinate has a scalar mean contraction ϕ_i and scalar second-moment contraction a_i, so the excess loss can be written exactly as a sum of modal contributions. This is the technical resolution of the least-squares coupling challenge: in standard SGD linear regression, the second-moment recursion is governed by fourth-moment operators and each eigendirection is affected by aggregate covariance across directions [7]. The row-wise recursion here lets us directly identify which frequencies have been learned, which remain underfit, and how different frequency ranges shape the learning curve.

Building on this decomposition, we study how the bias and variance terms scale with large vocabulary size V. We report the relative quantities $r_{\mathrm{bias}}(T) := \frac{\mathcal{E}_{\mathrm{bias}}(T)}{\mathcal{E}(0)}$ and $r_{\mathrm{var}}(T) := \frac{\mathcal{L}_{\mathrm{var}}(T)}{\mathcal{E}(0)}$, which allow comparisons across different V and α.



(a) $\alpha = 0.5$. (b) $\alpha = 1$.

Figure 1: **Exact bias–variance decomposition.** Online SGD follows the exact theoretical decomposition closely for $V = 1000$, $b = 1024$ at $\alpha = 0.5$ and $\alpha = 1$. Simulations use five independent seeds and $\eta = z$. The dashed vertical line denotes $T_{\text{crit}}(V) = V^{\alpha}$.

Theorem 4 (Bias and variance scaling) *Let the learning-rate rescaling factor be $\kappa = \eta/z \leq 1$, the row-wise effective learning rate be $\delta_i = \eta\pi_i = \kappa i^{-\alpha}$, the batch size be b, and the active-vocabulary cutoff be $i_{\kappa,*}(T) = \min\{V, \lfloor(\kappa T)^{1/\alpha}\rfloor\}$. For $\eta \leq z$, the relative bias satisfies*

$$r_{\text{bias}}(T) \asymp \frac{1}{z}\left(\frac{1}{\kappa^2 T^2}\sum\nolimits_{i \leq i_{\kappa,*}(T)} i^{\alpha} + \sum\nolimits_{i > i_{\kappa,*}(T)} i^{-\alpha}\right). \tag{2}$$

For $\eta \leq \min\{z, b/2\}$, the relative variance satisfies

$$r_{\text{var}}(T) \asymp \frac{1}{b\mathcal{E}(0)}\left(\frac{i_{\kappa,*}(T)}{T} + \kappa^2 T \sum\nolimits_{i > i_{\kappa,*}(T)} i^{-2\alpha}\right). \tag{3}$$

The two terms in (3) refine the activation interpretation. Activated rows contribute the mini-batch variance floor $i_{\kappa,*}(T)/T$, while inactive rows contribute the tail term $\kappa^2 T \sum_{i > i_{\kappa,*}(T)} i^{-2\alpha}$. Thus inactive rows are not only a source of residual bias; in sufficiently heavy-tailed regimes, their stochastic fluctuations can also control the variance scaling.

The active-vocabulary cutoff $i_{\kappa,*}(T)$ can be interpreted as the effective dimension at time T. High-frequency rows with indices below $i_{\kappa,*}(T)$ have been activated by time T and are in their asymptotic regime, while lower-frequency rows remain under-activated and drive the residual bias through the tail sums. This cutoff reaches the full vocabulary at the *critical time* $T_{\text{crit}}^{(\kappa)} = \kappa^{-1}V^{\alpha}$, the smallest time at which all V tokens have been activated.

Corollary 5 (Zipfian total risk) *For $\alpha = 1$, with learning rate $\eta = z$ and batch size $b \geq 2z$, the relative total loss $r_{\text{total}}(T) := r_{\text{bias}}(T) + r_{\text{var}}(T)$ has the following scaling:*

$$r_{\text{total}}(T) \asymp \begin{cases} 1 - \frac{\log T}{\log V} + \frac{(\log V)^2}{b}, & T \lesssim V \\ \frac{V^2}{T^2 \log V} + \frac{V(\log V)^2}{bT}, & T \gg V \end{cases}$$

Corollary 5 recovers the worst-case $\Theta(1 - \log T/\log V)$ scaling [3] in the bias term, while the variance term stays flat in the pre-critical regime. A further consequence is that for heavier-tailed regimes $\alpha < 1$, optimization difficulty can stem from the increasing variance term (Figure 1 panel

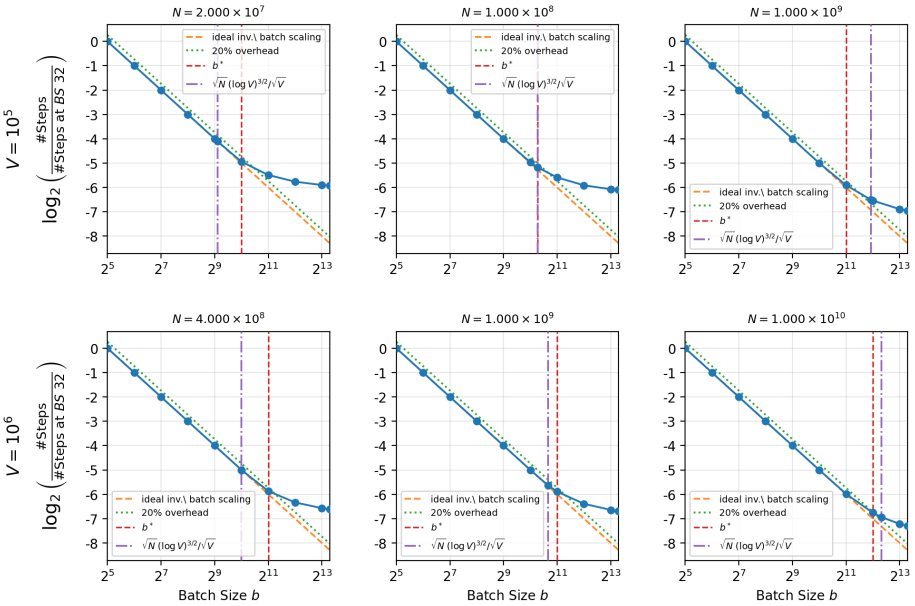


Figure 2: **Critical-batch-size validation for $\alpha = 1$.** The measured critical batch size b^*, defined by a 20% overhead in T_{hit} relative to ideal inverse-batch scaling, aligns with the theoretical value $b_{\text{crit}}(N)$ at the predicted scale.

(a) and Figure 8), even though the critical time is smaller and the bias term scales better than in the Zipfian case. We define the *critical batch size* (CBS) as the batch size that balances the asymptotic rates of the bias and variance terms, and validate it experimentally (Figure 2) against the empirical CBS definition used by [11].

Proposition 6 (Critical batch size) *In the supercritical regime $T \gg V^\alpha$, $b_{\text{crit}}(T) \asymp \frac{z}{\mathcal{E}(0)} \frac{T}{V^\alpha}$. At fixed sample compute $N = bT$, the Zipfian $\alpha = 1$ case satisfies $b_{\text{crit}}(N) \asymp \sqrt{N} \frac{(\log V)^{3/2}}{\sqrt{V}}$.*

The full regime-wise critical-batch scaling, together with the bias and variance scalings before and after the critical time, is summarized in Appendix A.

4. Experiments

We validate the exact decomposition on synthetic power-law bigrams (Figure 1) and test the CBS prediction with a batch-size sweep (Figure 2; details are deferred to Appendix B). We also validate the Zipf power-law assumption on OpenWebText in Appendix B.1.

5. Conclusion

We analyzed mini-batch SGD with Polyak averaging in the linear bigram model under power-law token frequencies. The exact bias–variance decomposition exposes row-wise frequency-mode dynamics that are hidden in standard linear regression analyses. From this decomposition, we derive bias and variance scaling laws across two phases and all power-law exponent regimes. We then use the supercritical scaling to predict the critical batch size and validate its empirical scale.

References

[1] Shane Bergsma, Nolan Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Power Lines: Scaling Laws for Weight Decay and Batch Size in LLM Pre-training, November 2025.

[2] Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, Venkata Krishna Pillutla, and Aaron Sidford. A Markov Chain Theory Approach to Characterizing the Minimax Optimality of Stochastic Gradient Descent (for Least Squares). *LIPIcs, Volume 93, FSTTCS 2017*, 93:2:1–2:10, 2018. ISSN 1868-8969. doi: 10.4230/LIPIcs.FSTTCS.2017.2.

[3] Frederik Kunstner and Francis Bach. Scaling Laws for Gradient Descent and Sign Descent for Linear Bigram Models under Zipf's Law, May 2025.

[4] Binghui Li, Kaifei Wang, Han Zhong, Pinyan Lu, and Liwei Wang. Muon in Associative Memory Learning: Training Dynamics and Scaling Laws, February 2026.

[5] Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An Empirical Model of Large-Batch Training, December 2018.

[6] William Merrill, Shane Arora, Dirk Groeneveld, and Hannaneh Hajishirzi. Critical Batch Size Revisited: A Simple Empirical Approach to Large-Batch Language Model Training, November 2025.

[7] Alexandru Meterez, Depen Morwani, Costin-Andrei Oncescu, Jingfeng Wu, Cengiz Pehlevan, and Sham Kakade. A Simplified Analysis of SGD for Linear Regression with Weight Averaging, June 2025.

[8] Alexandru Meterez, Pranav Ajit Nair, Depen Morwani, Cengiz Pehlevan, and Sham Kakade. Anytime Pretraining: Horizon-Free Learning-Rate Schedules with Weight Averaging, February 2026.

[9] B. T. Polyak and A. B. Juditsky. Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, July 1992. ISSN 0363-0129, 1095-7138. doi: 10.1137/0330046.

[10] Guodong Zhang, Lala Li, Zachary Nado, James Martens, Sushant Sachdeva, George E. Dahl, Christopher J. Shallue, and Roger Grosse. Which Algorithmic Choices Matter at Which Batch Sizes? Insights From a Noisy Quadratic Model, October 2019.

[11] Hanlin Zhang, Depen Morwani, Nikhil Vyas, Jingfeng Wu, Difan Zou, Udaya Ghai, Dean Foster, and Sham Kakade. How Does Critical Batch Size Scale in Pre-training?, April 2025.

[12] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. Benign Overfitting of Constant-Stepsize SGD for Linear Regression, 2021.

AI Assistance Disclosure

The authors used AI assistance for editing, formatting, and submission preparation. All scientific content, claims, proofs, and final text were reviewed and approved by the authors.

Appendix Contents

A Scaling Tables 7

B Additional Experimental Details 8
B.1 Zipf Law Validation . . . 10
B.2 Exact Decomposition Figures . . . 10
B.3 Surrogate Scaling Verification . . . 10

C Proofs for Preliminaries 17
C.1 Proof of Lemma 2 (Population structure) . . . 17
C.2 Asymptotics of the normalizing constant . . . 17
C.3 Asymptotics of $\mathcal{E}(0)$ . . . 17

D Stochastic Gradient Analysis 18
D.1 Stochastic gradient moments . . . 18
D.2 Coordinate recursion . . . 19
D.3 Mean trajectory and bias . . . 19
D.4 Second moments and temporal covariances . . . 20

E Proof of the Bias–Variance Decomposition (Theorem 3) 22

F Proof of the Bias Scaling Law (Theorem 4) 23

G Proof of the Variance Scaling Law (Theorem 4) 25

H Proofs for Critical Batch Size 30
H.1 Proof of the Zipf intermediate-compute scaling . . . 30
H.2 Proof of the Zipf supercritical scaling . . . 32
H.3 Critical-batch scaling for other power-law exponents . . . 33

Appendix A. Scaling Tables

Regime	$b_{\mathrm{crit}}(N)$ at fixed sample compute $N = bT$
$0 < \alpha < 1/2$	$\sqrt{N}\, V^{1-\alpha}$
$\alpha = 1/2$	$\sqrt{NV/\log V}$
$1/2 < \alpha < 1$	$\sqrt{N}\, V^{3/2-2\alpha}$
$\alpha = 1$	$\sqrt{N}\, (\log V)^{3/2}/\sqrt{V}$
$\alpha > 1$	$\sqrt{N}/V^{\alpha/2}$

Table 1: Regime-wise critical-batch-size scaling at fixed sample compute $N = bT$.

Regime	Pre-critical $T \lesssim V^\alpha$		Post-critical $T \gg V^\alpha$	
	$r_{\text{bias}}(T)$	$r_{\text{var}}(T)$	$r_{\text{bias}}(T)$	$r_{\text{var}}(T)$
$0 < \alpha < 1/2$	$1 - \frac{2\alpha}{1+\alpha}\frac{T^{(1-\alpha)/\alpha}}{V^{1-\alpha}}$	$TV^{2-2\alpha}/b$	$V^{2\alpha}/T^2$	$V^2/(bT)$
$\alpha = 1/2$	$1 - \frac{2\alpha}{1+\alpha}\frac{T^{(1-\alpha)/\alpha}}{V^{1-\alpha}}$	TV/b	$V^{2\alpha}/T^2$	$V^2/(bT\log V)$
$1/2 < \alpha < 1$	$1 - \frac{2\alpha}{1+\alpha}\frac{T^{(1-\alpha)/\alpha}}{V^{1-\alpha}}$	$T^{1/\alpha-1}V^{2-2\alpha}/b$	$V^{2\alpha}/T^2$	$V^{3-2\alpha}/(bT)$
$\alpha = 1$	$1 - \log T/\log V$	$(\log V)^2/b$	$V^2/(T^2\log V)$	$V(\log V)^2/(bT)$
$\alpha > 1$	$T^{(1-\alpha)/\alpha}$	$T^{1/\alpha-1}/b$	$V^{\alpha+1}/T^2$	$V/(bT)$

Table 2: Relative bias and variance scalings across power-law regimes, before and after the critical time $T_{\text{crit}}(V) = V^\alpha$ for the canonical step size $\eta = z$. The variance column includes its $1/b$ batch-size dependence, while the bias is batch-size independent.

Appendix B. Additional Experimental Details

Critical-batch sweep. The experiment in Figure 2 uses the exact Polyak bias and variance formulas. For each (V, α, N, b), it evaluates relative risk at the nominal horizon $T = \lfloor N/b \rfloor$. We fix a reference batch size b_{ref} that sits under the linear scaling regime, and defines the reference loss L_{ref} at $T_{\text{ref}} = \lfloor N/b_{\text{ref}} \rfloor$. For every other batch size, the experiment records the smallest T_{hit} for which the closed-form Polyak risk is at most L_{ref}, and compares it with the ideal inverse-batch prediction $T_{\text{ideal}} = T_{\text{ref}} b_{\text{ref}}/b$. The plotted knee is where the overhead $T_{\text{hit}}/T_{\text{ideal}} - 1$ becomes non-negligible. We follow the 20% overhead convention as in [11] to define empirical critical batch size b^*. The vertical theory reference is $b_{\text{crit}}(N)$, e.g. $\sqrt{N}(\log V)^{3/2}/\sqrt{V}$ for $\alpha = 1$.

Learning-rate sweep for Figure 2. The batch-size sweep uses the synthetic Zipf–Markov linear bigram model in the $\alpha = 1$ regime. For each plotted triple (V, N, b), we set $T = \lfloor N/b \rfloor$, and sweep the rescaled learning rate $\kappa = \eta/z$ on a dense log grid over $(0, 1]$, followed by local refinement around the best point. Each candidate η is scored by the $\alpha = 1$ surrogate Polyak risk

$$r_{\text{total}}(\eta) = \frac{1}{z}\left(\frac{i_*(i_*+1)}{2\kappa^2 T^2} + H_V - H_{i_*}\right) + \frac{1}{b\mathcal{E}(0)}\left(\frac{i_*}{T} + \kappa^2 T\left(H_V^{(2)} - H_{i_*}^{(2)}\right)\right), \qquad i_* = \min\{V, \lfloor \kappa T \rfloor\}.$$

The optimum η_* is reported in Table 3. The main pattern is simple: in the lowest-compute panels, η_* grows almost linearly with batch size so that the active cutoff i_* stays nearly fixed; once the budget is comfortably supercritical, the sweep usually saturates at the canonical rate $\eta_* = z$.

Table 3: Learning-rate sweep used for the critical-batch experiment. Here $z = H_V$, $\kappa_* = \eta_*/z$, and $i_* = \lfloor \kappa_* T \rfloor$ after clipping at V. Numeric columns use grouped digits for readability.

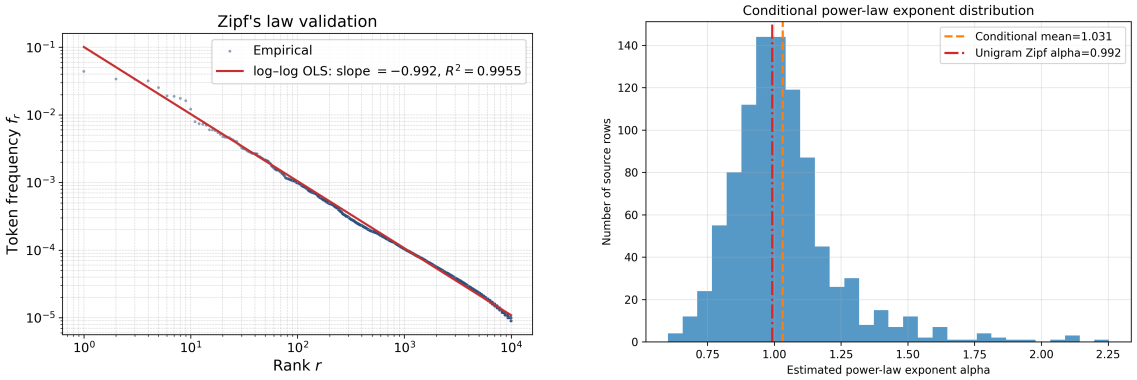
V	N (samples)	b	$T = N/b$	η_*	i_*
$\boldsymbol{V = 10^5,\ z = 12.090146}$					
100,000	20,000,000	32	625,000	0.094585	4,889
100,000	20,000,000	64	312,500	0.189579	4,900
100,000	20,000,000	128	156,250	0.378586	4,892
100,000	20,000,000	256	78,125	0.756032	4,885
100,000	20,000,000	512	39,062	1.515335	4,895
100,000	20,000,000	1,024	19,531	3.031661	4,897
100,000	20,000,000	2,048	9,765	6.061716	4,895
100,000	20,000,000	4,096	4,882	12.090146	4,882
100,000	20,000,000	8,192	2,441	12.090146	2,441
100,000	100,000,000	32	3,125,000	12.090146	100,000
100,000	100,000,000	64	1,562,500	12.090146	100,000
100,000	100,000,000	128	781,250	12.090146	100,000
100,000	100,000,000	256	390,625	12.090146	100,000
100,000	100,000,000	512	195,312	12.090146	100,000
100,000	100,000,000	1,024	97,656	12.090146	97,656
100,000	100,000,000	2,048	48,828	9.115626	36,814
100,000	100,000,000	4,096	24,414	12.090146	24,414
100,000	100,000,000	8,192	12,207	12.090146	12,207
100,000	1,000,000,000	32	31,250,000	12.090146	100,000
100,000	1,000,000,000	64	15,625,000	12.090146	100,000
100,000	1,000,000,000	128	7,812,500	12.090146	100,000
100,000	1,000,000,000	256	3,906,250	12.090146	100,000
100,000	1,000,000,000	512	1,953,125	12.090146	100,000
100,000	1,000,000,000	1,024	976,562	12.090146	100,000
100,000	1,000,000,000	2,048	488,281	12.090146	100,000
100,000	1,000,000,000	4,096	244,140	12.090146	100,000
100,000	1,000,000,000	8,192	122,070	12.090146	100,000
$\boldsymbol{V = 10^6,\ z = 14.392727}$					
1,000,000	400,000,000	32	12,500,000	0.067521	58,641
1,000,000	400,000,000	64	6,250,000	0.135085	58,660
1,000,000	400,000,000	128	3,125,000	0.269764	58,572
1,000,000	400,000,000	256	1,562,500	0.539705	58,591
1,000,000	400,000,000	512	781,250	1.077784	58,503
1,000,000	400,000,000	1,024	390,625	2.160231	58,629
1,000,000	400,000,000	2,048	195,312	4.313955	58,541
1,000,000	400,000,000	4,096	97,656	8.640061	58,623
1,000,000	400,000,000	8,192	48,828	14.392727	48,828

Continued on next page

V	N (samples)	b	$T = N/b$	η_*	i_*
1,000,000	1,000,000,000	32	31,250,000	0.076208	165,466
1,000,000	1,000,000,000	64	15,625,000	0.152187	165,217
1,000,000	1,000,000,000	128	7,812,500	0.304474	165,271
1,000,000	1,000,000,000	256	3,906,250	0.609147	165,325
1,000,000	1,000,000,000	512	1,953,125	1.216459	165,076
1,000,000	1,000,000,000	1,024	976,562	2.438180	165,433
1,000,000	1,000,000,000	2,048	488,281	4.864101	165,017
1,000,000	1,000,000,000	4,096	244,140	9.747727	165,348
1,000,000	1,000,000,000	8,192	122,070	14.392727	122,070
1,000,000	10,000,000,000	32	312,500,000	14.392727	1,000,000
1,000,000	10,000,000,000	64	156,250,000	14.392727	1,000,000
1,000,000	10,000,000,000	128	78,125,000	14.392727	1,000,000
1,000,000	10,000,000,000	256	39,062,500	14.392727	1,000,000
1,000,000	10,000,000,000	512	19,531,250	14.392727	1,000,000
1,000,000	10,000,000,000	1,024	9,765,625	14.392727	1,000,000
1,000,000	10,000,000,000	2,048	4,882,812	14.392727	1,000,000
1,000,000	10,000,000,000	4,096	2,441,406	14.392727	1,000,000
1,000,000	10,000,000,000	8,192	1,220,703	14.392727	1,000,000

B.1. Zipf Law Validation

We estimate bigram statistics from 5×10^6 OpenWebText tokens using the `t5-small` SentencePiece tokenizer, restricted to the top $V = 10{,}000$ types. The empirical marginal is Zipfian, and among 948 well-supported conditional rows the median fitted exponent is $\hat{\alpha} \approx 1.01$.



(a) Marginal Zipf law.

(b) Conditional-row fitted exponents.

Figure 3: OpenWebText validation of the power-law bigram assumption.

B.2. Exact Decomposition Figures

B.3. Surrogate Scaling Verification

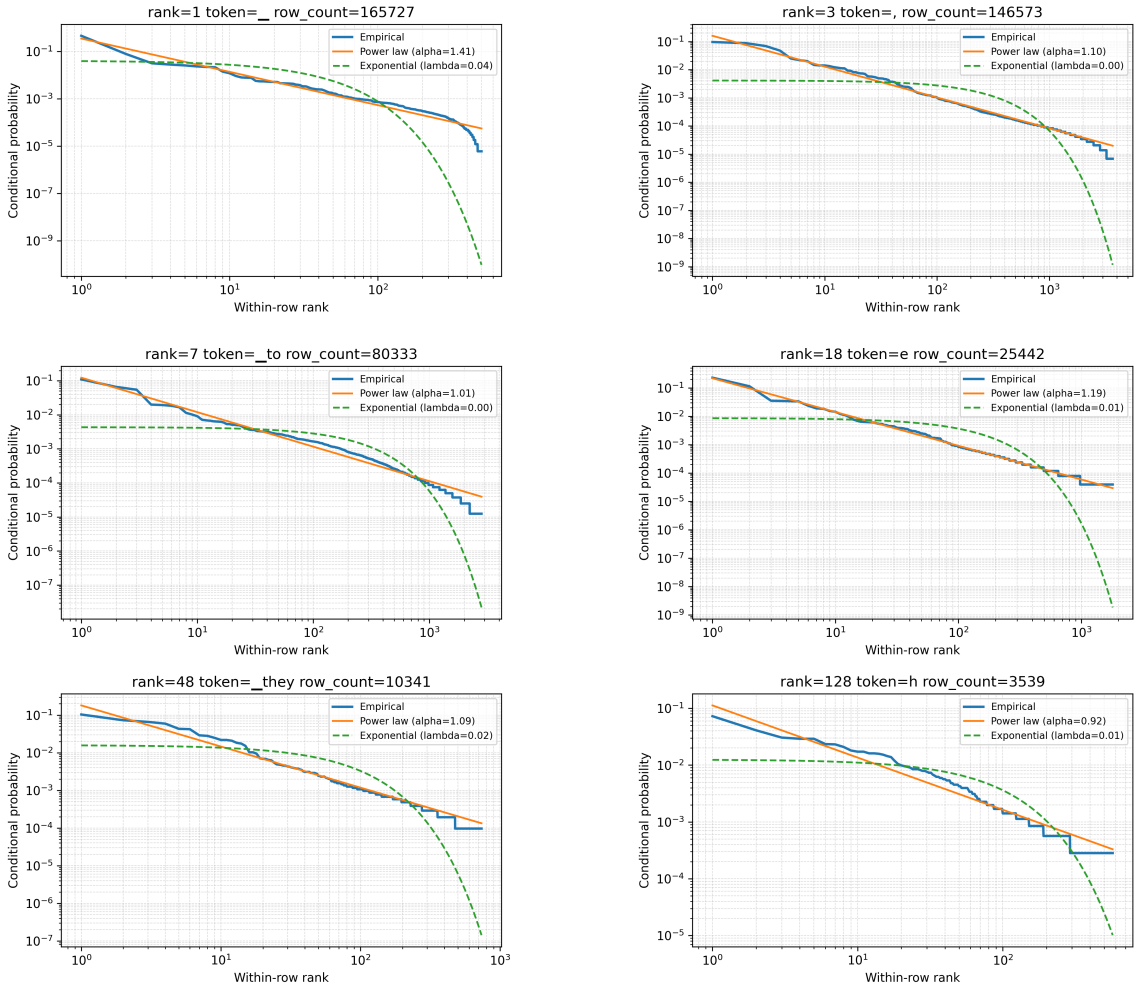
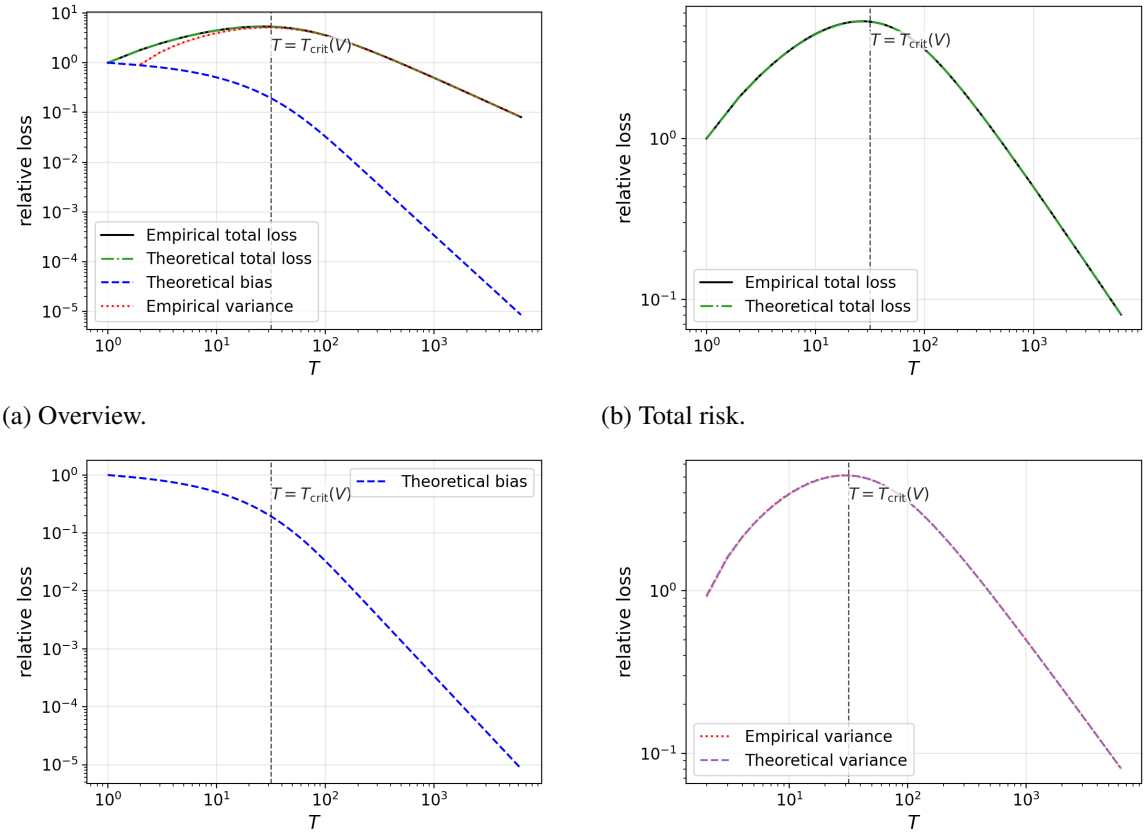


Figure 4: **Per-row power-law overlays.** Sorted-tail log–log plots of six representative rows of the empirical bigram transition matrix, with fitted finite-support discrete power-law overlays.



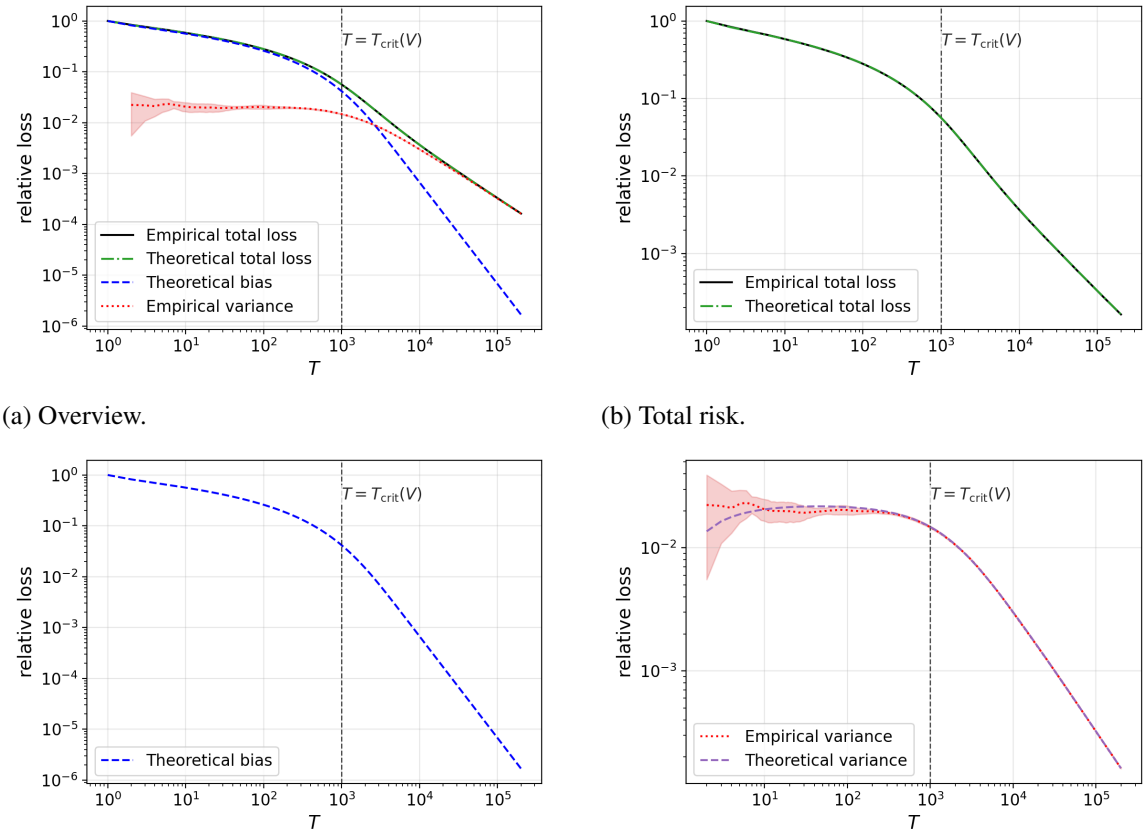
(a) Overview.

(b) Total risk.

(c) Bias.

(d) Variance.

Figure 5: **Exact bias–variance decomposition,** $\alpha = 0.5$**.** Detailed views for $V = 1000$, $b = 1024$, and $K = 5$ seeds.



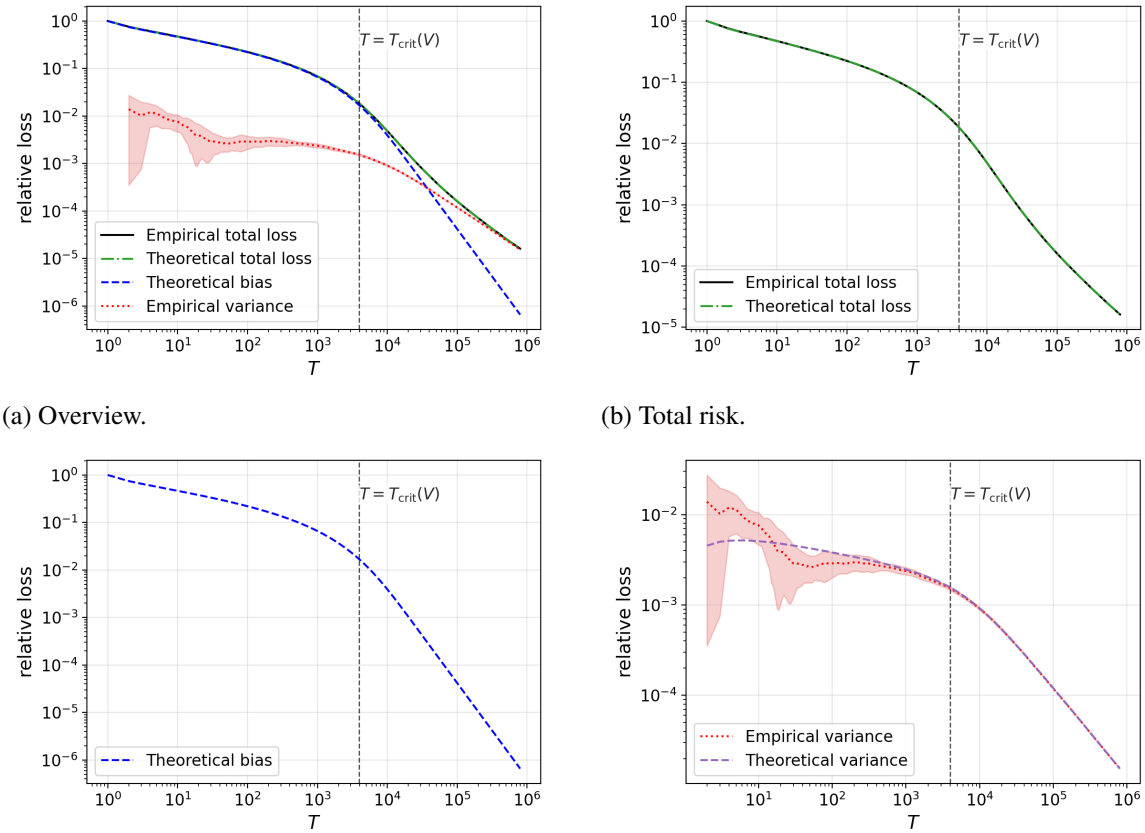
(a) Overview.

(b) Total risk.

(c) Bias.

(d) Variance.

Figure 6: **Exact bias–variance decomposition,** $\alpha = 1$**.** Detailed views for $V = 1000$, $b = 1024$, and $K = 5$ seeds.



(a) Overview.

(b) Total risk.

(c) Bias.

(d) Variance.

Figure 7: **Exact bias–variance decomposition,** $\alpha = 1.2$**.** Detailed views for $V = 1000$, $b = 1024$, and $K = 5$ seeds.

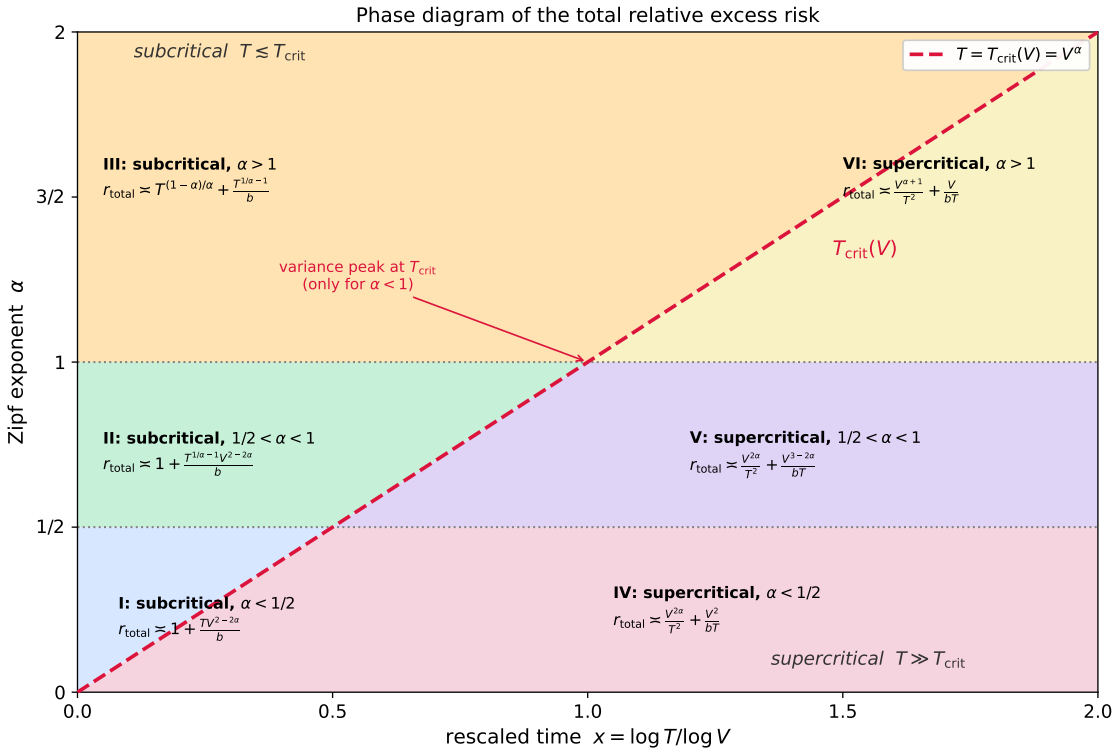


Figure 8: Phase diagram of total relative excess risk over $x = \log T / \log V$ and α. The red diagonal marks $T_{\text{crit}}(V) = V^{\alpha}$. For $\alpha < 1$, the variance term increases during subcritical phrase.

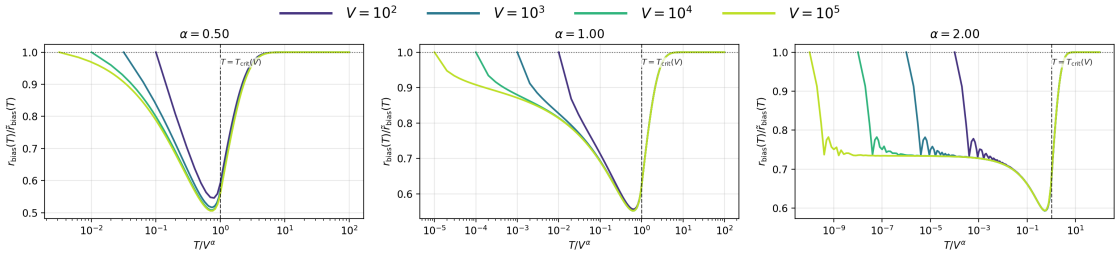


Figure 9: **Bias surrogate verification across vocabulary sizes and power-law exponents.** Ratio $r_{\text{bias}}(T)/\tilde{r}_{\text{bias}}(T)$ versus T/V^{α} for $V \in \{10^2, 10^3, 10^4, 10^5\}$ and $\alpha \in \{0.5, 1, 2\}$, using the surrogate in (2). The ratio stays bounded across all V,α regime.

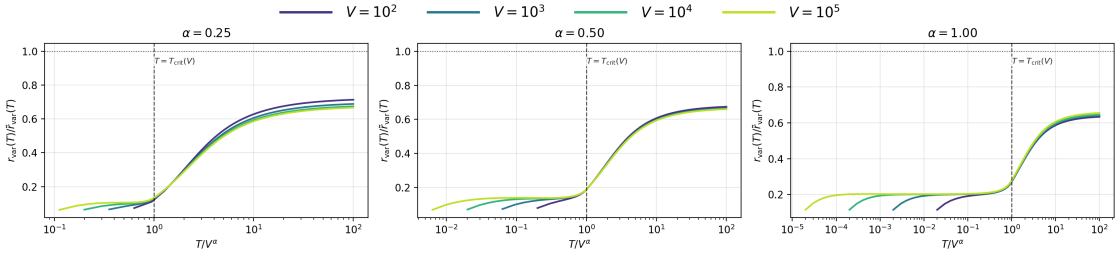


Figure 10: **Variance surrogate verification across vocabulary sizes and power-law exponents.** For $V \in \{10^2, 10^3, 10^4, 10^5\}$ and $\alpha \in \{0.25, 0.5, 1\}$, using the surrogate in 3. The ratio stays bounded across all V,α regime.

Appendix C. Proofs for Preliminaries

C.1. Proof of Lemma 2 (Population structure)

Proof Since $x = e_I$, we have $xx^\top = e_I e_I^\top$, and therefore

$$D = \mathbb{E}[xx^\top] = \sum_{i=1}^{V} \mathbb{P}(I = i)\, e_i e_i^\top = \mathrm{diag}(\pi_1, \ldots, \pi_V).$$

Next, conditioning on I gives

$$\mathcal{L}(W) = \frac{1}{2} \sum_{i=1}^{V} \pi_i\, \mathbb{E}\Big[\|W_{i,:}^\top - y\|_2^2 \,\Big|\, I = i\Big].$$

Therefore the minimization over W decouples row-by-row. For fixed i, the unique minimizer is $\mathbb{E}[y \mid I = i] = \sum_{j=1}^{V} \pi_{j|i} e_j$, hence $[W^\star]_{ij} = \pi_{j|i}$.

For the quadratic expansion, we apply the conditional bias–variance identity:

$$\mathbb{E}\big[\|a - y\|_2^2 \,\big|\, I = i\big] = \|a - \mathbb{E}[y \mid I = i]\|_2^2 + \mathbb{E}\big[\|y - \mathbb{E}[y \mid I = i]\|_2^2 \,\big|\, I = i\big].$$

Substituting $a = W_{i,:}^\top$ and summing gives $\mathcal{E}(W) = \frac{1}{2}\sum_{i=1}^{V} \pi_i \|W_{i,:} - W_{i,:}^\star\|_2^2 = \frac{1}{2}\,\mathrm{Tr}\big((W - W^\star)^\top D (W - W^\star)\big)$. ■

C.2. Asymptotics of the normalizing constant

Let $\zeta(z) = \sum_{i=1}^{\infty} i^{-z}$ be the Riemann zeta function.

Lemma 7 (Asymptotics of normalization factor z) *As $V \to \infty$,*

$$z = H_V^{(\alpha)} = \begin{cases} \frac{V^{1-\alpha}}{1-\alpha} + \mathrm{O}(1), & 0 < \alpha < 1, \\ \log V + \mathrm{O}(1), & \alpha = 1, \\ \zeta(\alpha) + \mathrm{O}(V^{1-\alpha}), & \alpha > 1. \end{cases}$$

C.3. Asymptotics of $\mathcal{E}(0)$

Lemma 8 (Asymptotics of $\mathcal{E}(0)$) *Under Assumption 1, as $V \to \infty$,*

$$\mathcal{E}(0) = \frac{1}{2}\frac{H_V^{(2\alpha)}}{(H_V^{(\alpha)})^2} = \begin{cases} \dfrac{(1-\alpha)^2}{2(1-2\alpha)} V^{-1} + O(V^{-2(1-\alpha)}), & 0 < \alpha < \frac{1}{2}, \\ \dfrac{\log V}{8V} + O(V^{-1}), & \alpha = \frac{1}{2}, \\ \dfrac{(1-\alpha)^2 \zeta(2\alpha)}{2} V^{-2(1-\alpha)}(1 + o(1)), & \frac{1}{2} < \alpha < 1, \\ \dfrac{\zeta(2)}{2(\log V)^2} + O((\log V)^{-3}), & \alpha = 1, \\ \dfrac{\zeta(2\alpha)}{2\zeta(\alpha)^2} + O(V^{1-\alpha}), & \alpha > 1. \end{cases}$$

Appendix D. Stochastic Gradient Analysis

D.1. Stochastic gradient moments

Proposition 9 (Expectation and covariance of the stochastic gradient) *Let $v \in \mathbb{R}^{V^2}$ denote the row-vectorized stochastic gradient associated with a single sampled pair (I, J), partitioned into V consecutive blocks $v^{(i)} \in \mathbb{R}^V$. Define $q_i := \mathbb{E}[e_J \mid I = i] = W^\star_{i,:}$ and its jth entry be q_{ij}. Then*

$$\mathbb{E}[v]^{(i)} = \pi_i(W_{i,:} - q_i).$$

The second-moment matrix $\mathbb{E}[vv^\top]$ is block diagonal with

$$\mathbb{E}[vv^\top]_{i,i} = \sum_{j=1}^{V} \pi_{ij}(W_{i,:} - e_j)(W_{i,:} - e_j)^\top, \qquad \mathbb{E}[vv^\top]_{i,j} = 0 \;\; (i \neq j).$$

The diagonal entries of the covariance diagonal blocks satisfy

$$[\Sigma_{i,i}]_{jj} = \pi_i q_{ij}(1 - q_{ij}) + \pi_i(1 - \pi_i)(W_{ij} - q_{ij})^2, \tag{4}$$

separating the irreducible sampling noise $\pi_i q_{ij}(1 - q_{ij})$ from the state-dependent contribution $\pi_i(1 - \pi_i)(W_{ij} - q_{ij})^2$.

Proof The row-vectorized gradient for one-hot inputs $x = e_I$, $y = e_J$ is

$$v(W; I, J) = \overline{\mathrm{vec}}\big(e_I(W_{I,:} - e_J)^\top\big) = e_I \otimes (W_{I,:} - e_J) \in \mathbb{R}^{V^2}.$$

The i-th block (V entries corresponding to row i) is

$$v^{(i)} = \mathbf{1}_{\{I=i\}}(W_{i,:} - e_J).$$

Taking expectation we get:

$$\mathbb{E}[v^{(i)}] = \mathbb{P}(I = i)\,\mathbb{E}[W_{i,:} - e_J \mid I = i] = \pi_i(W_{i,:} - q_i).$$

For the (i, i) diagonal block:

$$\begin{aligned}\mathbb{E}[v^{(i)}(v^{(i)})^\top] &= \mathbb{E}\big[\mathbf{1}_{\{I=i\}}(W_{i,:} - e_J)(W_{i,:} - e_J)^\top\big] \\ &= \pi_i \sum_{j=1}^{V} \pi_{j|i}(W_{i,:} - e_j)(W_{i,:} - e_j)^\top = \sum_{j=1}^{V} \pi_{ij}(W_{i,:} - e_j)(W_{i,:} - e_j)^\top.\end{aligned}$$

For the (i, k) off-diagonal block with $i \neq k$:

$$\mathbb{E}[v^{(i)}(v^{(k)})^\top] = \mathbb{E}\big[\mathbf{1}_{\{I=i\}}\mathbf{1}_{\{I=k\}}(\cdots)\big] = 0,$$

since I cannot equal both i and k simultaneously.

Next we calculate the covariance diagonal entries which will be used in 10. The (j, j)-th diagonal entry of $\Sigma_{i,i} = \mathbb{E}[v^{(i)}(v^{(i)})^\top] - \mathbb{E}[v^{(i)}]\mathbb{E}[v^{(i)}]^\top$ is

$$\begin{aligned}[\Sigma_{i,i}]_{jj} &= \sum_{k=1}^{V} \pi_{ik}(W_{ij} - \mathbf{1}_{\{k=j\}})^2 - \pi_i^2(W_{ij} - q_{ij})^2 \\ &= \pi_i\big[q_{ij}(W_{ij} - 1)^2 + (1 - q_{ij})W_{ij}^2\big] - \pi_i^2(W_{ij} - q_{ij})^2.\end{aligned}$$

Expanding and simplifying with $W_{ij} = q_{ij} + \Delta_{ij}$:

$$\begin{aligned} q_{ij}(W_{ij}-1)^2 + (1-q_{ij})W_{ij}^2 &= q_{ij}(1-q_{ij}) + (W_{ij}-q_{ij})^2 \\ &= q_{ij}(1-q_{ij}) + \Delta_{ij}^2. \end{aligned}$$

Therefore

$$[\Sigma_{i,i}]_{jj} = \pi_i q_{ij}(1-q_{ij}) + \pi_i \Delta_{ij}^2 - \pi_i^2 \Delta_{ij}^2 = \pi_i q_{ij}(1-q_{ij}) + \pi_i(1-\pi_i)\Delta_{ij}^2. \square$$

■

D.2. Coordinate recursion

Lemma 10 (Coordinate recursion and conditional second moment) *For every $t \geq 0$ and every pair (i,j),*

$$\Delta_{ij,t+1} = \phi_i \Delta_{ij,t} - \eta \epsilon_{ij,t},$$

where $\epsilon_{ij,t} := [G_t]_{ij} - \pi_i \Delta_{ij,t}$ satisfies $\mathbb{E}[\epsilon_{ij,t} \mid \mathcal{F}_t] = 0$ and

$$\mathbb{E}[\epsilon_{ij,t}^2 \mid \mathcal{F}_t] = \frac{1}{b}\big[\pi_i q_{ij}(1-q_{ij}) + \pi_i(1-\pi_i)\Delta_{ij,t}^2\big].$$

Proof For one sample (I,J), the (i,j) coordinate of the sample gradient is $g_{ij}(W;I,J) = \mathbf{1}_{\{I=i\}}(W_{ij} - \mathbf{1}_{\{J=j\}})$. Conditioning on W yields

$$\mathbb{E}[g_{ij}(W;I,J) \mid W] = \pi_i(W_{ij} - q_{ij}).$$

Hence $\mathbb{E}[[G_t]_{ij} \mid \mathcal{F}_t] = \pi_i \Delta_{ij,t}$, and $\mathbb{E}[\epsilon_{ij,t} \mid \mathcal{F}_t] = 0$. The SGD update becomes $\Delta_{ij,t+1} = \phi_i \Delta_{ij,t} - \eta \epsilon_{ij,t}$.

For the conditional second moment, the b terms in the mini-batch average are conditionally i.i.d., so the conditional variance of the batch average is the one-sample conditional variance divided by b. For one sample,

$$\mathbb{E}[g_{ij}(W;I,J)^2 \mid W] = \pi_i[q_{ij}(W_{ij}-1)^2 + (1-q_{ij})W_{ij}^2].$$

Writing $W_{ij} = q_{ij} + \Delta_{ij}$ gives $q_{ij}(W_{ij}-1)^2 + (1-q_{ij})W_{ij}^2 = q_{ij}(1-q_{ij}) + \Delta_{ij}^2$. Therefore,

$$\mathrm{Var}(g_{ij}(W;I,J) \mid W) = \pi_i q_{ij}(1-q_{ij}) + \pi_i(1-\pi_i)\Delta_{ij}^2.$$

Dividing by b gives the stated formula. ■

D.3. Mean trajectory and bias

Proposition 11 (Mean trajectory is deterministic GD) *Let W_t^{gd} be the full-gradient descent iterates defined by $W_{t+1}^{\mathrm{gd}} = W_t^{\mathrm{gd}} - \eta D(W_t^{\mathrm{gd}} - W^\star)$ with $W_0^{\mathrm{gd}} = W_0$. Then for every $t \geq 0$, $\mathbb{E}[W_t] = W_t^{\mathrm{gd}}$. Moreover, for every (i,j),*

$$\mathbb{E}[\Delta_{ij,t}] = \phi_i^t \Delta_{ij,0}.$$

Let $\bar{\Delta}_{ij,T} = \frac{1}{T}\sum_{t=0}^{T-1} \Delta_{ij,t}$, then

$$\mathbb{E}[\bar{\Delta}_{ij,T}] = \frac{1}{T}\sum_{t=0}^{T-1} \phi_i^t \Delta_{ij,0} = \frac{1-\phi_i^T}{T(1-\phi_i)}\Delta_{ij,0}.$$

Proof Taking conditional expectation in Lemma 10 gives

$$\mathbb{E}[\Delta_{ij,t+1} \mid \mathcal{F}_t] = \phi_i \Delta_{ij,t}.$$

Taking expectation once more yields

$$\mathbb{E}[\Delta_{ij,t+1}] = \phi_i \mathbb{E}[\Delta_{ij,t}],$$

and induction gives $\mathbb{E}[\Delta_{ij,t}] = \phi_i^t \Delta_{ij,0}$. This is exactly the coordinate recursion of deterministic gradient descent. Averaging over $t = 0, \ldots, T-1$ and using the geometric series formula:

$$\mathbb{E}[\bar{\Delta}_{ij,T}] = \frac{1}{T} \sum_{t=0}^{T-1} \phi_i^t \Delta_{ij,0} = \frac{1-\phi_i^T}{T(1-\phi_i)} \Delta_{ij,0}. \square$$

■

Proposition 12 (Bias term) *Define* $\Gamma_i(T) := \left(\frac{1-\phi_i^T}{1-\phi_i}\right)^2$. *Then*

$$\mathcal{E}_{\text{bias}}(T) := \mathcal{L}(\mathbb{E}[\bar{W}_T]) - \mathcal{L}(W^\star) = \frac{1}{2T^2} \sum_{i=1}^{V} \pi_i \Gamma_i(T) \sum_{j=1}^{V} \Delta_{ij,0}^2.$$

Proof By Proposition 11,

$$\mathbb{E}[\bar{\Delta}_{ij,T}] = \frac{1-\phi_i^T}{T(1-\phi_i)} \Delta_{ij,0}.$$

Squaring and summing with weights $\pi_i/2$:

$$\begin{aligned}\mathcal{E}_{\text{bias}}(T) &= \frac{1}{2} \sum_{i=1}^{V} \pi_i \sum_{j=1}^{V} (\mathbb{E}[\bar{\Delta}_{ij,T}])^2 = \frac{1}{2} \sum_{i=1}^{V} \pi_i \sum_{j=1}^{V} \left(\frac{1-\phi_i^T}{T(1-\phi_i)}\right)^2 \Delta_{ij,0}^2 \\ &= \frac{1}{2T^2} \sum_{i=1}^{V} \pi_i \Gamma_i(T) \sum_{j=1}^{V} \Delta_{ij,0}^2. \square\end{aligned}$$

■

D.4. Second moments and temporal covariances

Proposition 13 (Second moments) *Define* $a_i := \phi_i^2 + \frac{\eta^2}{b} \pi_i (1-\pi_i)$ *and* $c_{ij} := \frac{\eta^2}{b} \pi_i q_{ij} (1-q_{ij})$. *If* $a_i < 1$, *then*

$$\mathbb{E}[\Delta_{ij,t}^2] = a_i^t \Delta_{ij,0}^2 + \frac{c_{ij}}{1-a_i}(1-a_i^t).$$

Proof Squaring the recursion of Lemma 10: $\Delta_{ij,t+1}^2 = \phi_i^2 \Delta_{ij,t}^2 - 2\eta\phi_i \Delta_{ij,t} \epsilon_{ij,t} + \eta^2 \epsilon_{ij,t}^2$. Taking conditional expectation and using $\mathbb{E}[\epsilon_{ij,t} \mid \mathcal{F}_t] = 0$ to get:

$$\mathbb{E}[\Delta_{ij,t+1}^2 \mid \mathcal{F}_t] = a_i \Delta_{ij,t}^2 + c_{ij}.$$

Taking expectation gives the linear recursion $\mathbb{E}[\Delta_{ij,t+1}^2] = a_i \mathbb{E}[\Delta_{ij,t}^2] + c_{ij}$. Since $a_i < 1$, solving yields the stated formula. ■

Lemma 14 (Cross-time covariance) *For* $0 \le s < t$, $\mathrm{Cov}(\Delta_{ij,t}, \Delta_{ij,s}) = \phi_i^{t-s}\,\mathrm{Var}(\Delta_{ij,s})$. *Consequently,*

$$\mathrm{Var}(\bar{\Delta}_{ij,T}) = \frac{1}{T^2}\sum_{s=0}^{T-1} w_{i,s}\,\mathrm{Var}(\Delta_{ij,s}),$$

where $w_{i,s} := 1 + 2\sum_{k=1}^{T-1-s}\phi_i^k = \frac{1+\phi_i}{1-\phi_i} - \frac{2\phi_i^{T-s}}{1-\phi_i}$.

Proof Unrolling the recursion from time s to time t gives

$$\Delta_{ij,t} = \phi_i^{t-s}\Delta_{ij,s} - \eta\sum_{k=s}^{t-1}\phi_i^{t-1-k}\epsilon_{ij,k}.$$

Multiplying by $\Delta_{ij,s}$ and taking expectations yields

$$\mathbb{E}[\Delta_{ij,t}\Delta_{ij,s}] = \phi_i^{t-s}\mathbb{E}[\Delta_{ij,s}^2] - \eta\sum_{k=s}^{t-1}\phi_i^{t-1-k}\mathbb{E}[\epsilon_{ij,k}\Delta_{ij,s}].$$

For every $k \ge s$, the variable $\Delta_{ij,s}$ is $\mathcal{F}_k$-measurable, hence

$$\mathbb{E}[\epsilon_{ij,k}\Delta_{ij,s}] = \mathbb{E}\big[\mathbb{E}[\epsilon_{ij,k} \mid \mathcal{F}_k]\Delta_{ij,s}\big] = 0.$$

Therefore

$$\mathbb{E}[\Delta_{ij,t}\Delta_{ij,s}] = \phi_i^{t-s}\mathbb{E}[\Delta_{ij,s}^2].$$

Since $\mathbb{E}[\Delta_{ij,t}]\mathbb{E}[\Delta_{ij,s}] = \phi_i^{t-s}(\mathbb{E}[\Delta_{ij,s}])^2$ by Proposition 11, subtraction gives

$$\mathrm{Cov}(\Delta_{ij,t}, \Delta_{ij,s}) = \phi_i^{t-s}\,\mathrm{Var}(\Delta_{ij,s}).$$

The variance of the average follows from

$$\begin{aligned}\mathrm{Var}\Big(\frac{1}{T}\sum_{t=0}^{T-1}\Delta_{ij,t}\Big) &= \frac{1}{T^2}\Big(\sum_{s=0}^{T-1}\mathrm{Var}(\Delta_{ij,s}) + 2\sum_{0\le s<t\le T-1}\mathrm{Cov}(\Delta_{ij,t}, \Delta_{ij,s})\Big)\\ &= \frac{1}{T^2}\sum_{s=0}^{T-1}\Big(1 + 2\sum_{k=1}^{T-1-s}\phi_i^k\Big)\,\mathrm{Var}(\Delta_{ij,s}).\end{aligned}$$

For the closed form of $w_{i,s}$, note that

$$w_{i,s} = 1 + 2\sum_{k=1}^{T-1-s}\phi_i^k = \frac{1+\phi_i}{1-\phi_i} - \frac{2\phi_i^{T-s}}{1-\phi_i}.\square$$

■

Corollary 15 (Exact variance of Polyak-averaged coordinate) *Define* $A_i := \frac{1+\phi_i}{1-\phi_i}$ *and* $B_i := \frac{2\phi_i}{1-\phi_i}$. *Then for every* (i, j),

$$\mathrm{Var}(\bar{\Delta}_{ij,T}) = \frac{\Delta_{ij,0}^2}{T^2}\big(S_i^{(a)}(T) - \Gamma_i(T)\big) + \frac{c_{ij}}{(1-a_i)T^2} S_i^{(1-a)}(T),$$

where $w_{i,s} = A_i - B_i\phi_i^{T-1-s}$, *and*

$$S_i^{(1-a)}(T) = \Big(A_i T - B_i \frac{1-\phi_i^T}{1-\phi_i}\Big) - S_i^{(a)}(T).$$

Proof Substituting the variance formula from Proposition 13 into Lemma 14 gives

$$\mathrm{Var}(\bar{\Delta}_{ij,T}) = \frac{1}{T^2}\sum_{s=0}^{T-1} w_{i,s}\Big[(a_i^s - \phi_i^{2s})\Delta_{ij,0}^2 + \frac{c_{ij}}{1-a_i}(1-a_i^s)\Big].$$

It therefore remains to evaluate $\sum_{s=0}^{T-1} w_{i,s}\phi_i^{2s}$. Using the definition of $w_{i,s}$,

$$\begin{aligned}\sum_{s=0}^{T-1} w_{i,s}\phi_i^{2s} &= \sum_{s=0}^{T-1}\phi_i^{2s} + 2\sum_{0\le s<t\le T-1}\phi_i^{t-s}\phi_i^{2s}\\ &= \sum_{s=0}^{T-1}\phi_i^{2s} + 2\sum_{0\le s<t\le T-1}\phi_i^{s+t} = \Big(\sum_{t=0}^{T-1}\phi_i^t\Big)^2 = \Gamma_i(T).\end{aligned}$$

This proves the variance formula.

For the closed forms, note that $w_{i,s} = A_i - B_i\phi_i^{T-1-s}$. Therefore,

$$S_i^{(a)}(T) = \sum_{s=0}^{T-1}(A_i - B_i\phi_i^{T-1-s})a_i^s = A_i\frac{1-a_i^T}{1-a_i} - B_i\sum_{s=0}^{T-1}\phi_i^{T-1-s}a_i^s.$$

Also,

$$\sum_{s=0}^{T-1} w_{i,s} = A_i T - B_i\sum_{s=0}^{T-1}\phi_i^{T-1-s} = A_i T - B_i\frac{1-\phi_i^T}{1-\phi_i},$$

whence

$$S_i^{(1-a)}(T) = \sum_{s=0}^{T-1} w_{i,s} - S_i^{(a)}(T) = \Big(A_i T - B_i\frac{1-\phi_i^T}{1-\phi_i}\Big) - S_i^{(a)}(T).\square$$

■

Appendix E. Proof of the Bias–Variance Decomposition (Theorem 3)

Proof By Lemma 2, $\mathcal{L}(W) - \mathcal{L}(W^\star) = \frac{1}{2}\|D^{1/2}(W - W^\star)\|_F^2$. Therefore $\mathbb{E}[\mathcal{L}(\bar{W}_T)] - \mathcal{L}(W^\star) = \frac{1}{2}\mathbb{E}[\|D^{1/2}\bar{\Delta}_T\|_F^2]$. Decompose $\bar{\Delta}_T = \mathbb{E}[\bar{\Delta}_T] + (\bar{\Delta}_T - \mathbb{E}[\bar{\Delta}_T])$. Expanding the squared Frobenius

norm and using $\mathbb{E}[\bar{\Delta}_T - \mathbb{E}[\bar{\Delta}_T]] = 0$, the cross term vanishes:

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\bar{W}_T)] - \mathcal{L}(W^\star) &= \frac{1}{2}\|D^{1/2}\mathbb{E}[\bar{\Delta}_T]\|_F^2 + \frac{1}{2}\mathbb{E}[\|D^{1/2}(\bar{\Delta}_T - \mathbb{E}[\bar{\Delta}_T])\|_F^2] \\ &= \frac{1}{2}\sum_{i=1}^{V}\pi_i\sum_{j=1}^{V}(\mathbb{E}[\bar{\Delta}_{ij,T}])^2 + \frac{1}{2}\sum_{i=1}^{V}\pi_i\sum_{j=1}^{V}\mathrm{Var}(\bar{\Delta}_{ij,T}).\end{aligned}$$

The first term is $\mathcal{E}_{\text{bias}}(T)$ by Proposition 12 and the second term is expressed explicitly by Corollary 15. Adding them yields the total-risk formula. Under initialization $W_0 = 0$ and assumption 1 we get $\sum_{j=1}^{V}\Delta_{ij,0}^2 = 2\mathcal{E}(0)$ and $\sum_{j=1}^{V} c_{ij} = \frac{\eta^2}{b}\pi_i(1 - 2\mathcal{E}(0))$. ■

Appendix F. Proof of the Bias Scaling Law (Theorem 4)

Statement of Theorem 4 for the bias term. Fix $\varepsilon \in (0,1)$ and assume

$$0 < \eta \le (1-\varepsilon)\frac{2bz}{b+z-1}.$$

Let

$$\kappa := \eta/z, \qquad i_*(\eta, T) := \min\left\{V, \left\lfloor (\kappa T)^{1/\alpha}\right\rfloor\right\}, \qquad T_{\text{crit}}(\eta, V) := \kappa^{-1}V^{\alpha}.$$

Then, with constants depending only on α and ε,

$$r_{\text{bias}}(T,\eta) \asymp \frac{1}{z}\left(\frac{1}{\kappa^2 T^2}\sum_{i \le i_*(\eta,T)} i^{\alpha} + \sum_{i > i_*(\eta,T)} i^{-\alpha}\right).$$

Proof Denote $\kappa := \eta/z$ and $\delta_i := \eta\pi_i = \kappa i^{-\alpha}$ the effective learning rate of the i-th row. Define

$$b_T(x) := \frac{1-(1-x)^T}{Tx}, \qquad 0 < x < 2. \tag{5}$$

Then the relative bias term from Proposition 12 can be written as

$$r_{\text{bias}}(T) = \sum_{i=1}^{V}\pi_i\, b_T(\delta_i)^2.$$

We organize the proof into three steps.

Step 1: Two-sided bound on $b_T(x)^2$. We prove that for every fixed $\varepsilon \in (0,1)$, uniformly over $0 < x \le 2(1-\varepsilon)$,

$$b_T(x)^2 \asymp \min\left\{1, \frac{1}{T^2x^2}\right\}. \tag{6}$$

Upper bound. Since $|1-x| \le 1$ for $0 < x \le 2$, we have $|1-(1-x)^T| \le 2$. If $0 < x \le 1$, the elementary inequality $1-(1-x)^T \le Tx$ also gives $|1-(1-x)^T| \le Tx$. If $x > 1$, then $Tx \ge 1$, so $2/(Tx) \lesssim 1$ with an absolute constant. Hence

$$b_T(x)^2 \le \min\left\{1, \frac{1}{T^2x^2}\right\}.$$

Lower bound. There are three cases. If $Tx \le 1$, then necessarily $x \le 1$, and

$$1-(1-x)^T \ge Tx - \binom{T}{2}x^2 \ge Tx - \frac{T^2x^2}{2} \ge \frac{Tx}{2},$$

so $b_T(x)^2 \ge 1/4$. If $x \le 1$ and $Tx \ge 1$, then $(1-x)^T \le e^{-Tx} \le e^{-1}$, hence

$$1-(1-x)^T \ge 1-e^{-1}, \qquad b_T(x)^2 \ge \frac{(1-e^{-1})^2}{T^2x^2}.$$

Finally, if $1 < x \le 2(1-\varepsilon)$, then $1-x \in [-1+2\varepsilon, 0)$ if $\varepsilon \le 1/2$ and $1-x \in (-1,0)$ otherwise, but in all cases $|1-x| \le 1-c_\varepsilon$ for some $c_\varepsilon > 0$. Thus $|1-(1-x)^T| \ge c_\varepsilon$ for every $T \ge 1$, and $b_T(x)^2 \gtrsim_\varepsilon 1/(T^2x^2)$. Since $Tx \ge 1$ in this case, this is the desired lower bound. This proves (6).

Step 2: Splitting into active and inactive rows. Applying (6) to $x = \delta_i = \kappa i^{-\alpha}$ yields

$$r_{\text{bias}}(T) \asymp \frac{1}{z}\sum_{i=1}^{V} i^{-\alpha} \min\left\{1, \frac{i^{2\alpha}}{\kappa^2T^2}\right\}.$$

If $i \le i_*(\eta,T)$, then $T\delta_i = \kappa T i^{-\alpha} \ge 1$, so the summand is comparable to $i^\alpha/(\kappa^2T^2)$. If $i > i_*(\eta,T)$, then $T\delta_i < 1$, so the summand is comparable to $i^{-\alpha}$. Therefore

$$r_{\text{bias}}(T) \asymp \frac{1}{z}\left(\sum_{i>i_*(\eta,T)} i^{-\alpha} + \frac{1}{\kappa^2T^2}\sum_{i\le i_*(\eta,T)} i^\alpha\right). \tag{7}$$

Step 3: Regime-wise asymptotics via integral test. Write $q := \kappa T = \eta T/z$. In the subcritical regime $T \le T_{\text{crit}}(\eta,V)$, equivalently $q \le V^\alpha$, we have $i_*(\eta,T) \asymp q^{1/\alpha}$. In the supercritical regime $T \ge T_{\text{crit}}(\eta,V)$, equivalently $q \ge V^\alpha$, the cutoff is $i_*(\eta,T) = V$. We use the standard integral-test estimates: for $\beta > -1$,

$$\sum_{i\le m} i^\beta \asymp m^{\beta+1};$$

for $\beta \ne 1$,

$$\sum_{i>m}^{V} i^{-\beta} \asymp \begin{cases} V^{1-\beta} - m^{1-\beta}, & 0<\beta<1, \\ m^{1-\beta}, & \beta > 1, \end{cases}$$

and for $\beta = 1$, $\sum_{i>m}^{V} i^{-1} \asymp \log(V/m)$.

Case 1: $0 < \alpha < 1$**.** If $T \le T_{\text{crit}}(\eta,V)$, then $i_*(\eta,T) \asymp q^{1/\alpha}$ and

$$\sum_{i>i_*}^{V} i^{-\alpha} = \frac{1}{1-\alpha}(V^{1-\alpha} - q^{(1-\alpha)/\alpha}), \qquad \frac{1}{\kappa^2T^2}\sum_{i\le i_*} i^\alpha = \frac{1}{\alpha+1}\frac{q^{(\alpha+1)/\alpha}}{q^2} = \frac{q^{(1-\alpha)/\alpha}}{\alpha+1}.$$

Adding the two contributions:

$$r_{\text{bias}}(T,\eta) \asymp \frac{1}{z}\left(\frac{V^{1-\alpha}}{1-\alpha} - \frac{2\alpha}{1-\alpha^2}q^{(1-\alpha)/\alpha}\right) \asymp 1 - \frac{2\alpha}{1+\alpha}\frac{q^{(1-\alpha)/\alpha}}{V^{1-\alpha}}.$$

If $T \geq T_{\text{crit}}(\eta, V)$, then $i_*(\eta, T) = V$ and

$$r_{\text{bias}}(T, \eta) \asymp \frac{1}{zq^2} \sum_{i=1}^{V} i^{\alpha} \asymp \frac{V^{\alpha+1}}{V^{1-\alpha} q^2} = \frac{V^{2\alpha}}{q^2}.$$

Case 2: $\alpha = 1$**.** If $T \leq T_{\text{crit}}(\eta, V)$, then $i_*(\eta, T) \asymp q$ and

$$r_{\text{bias}}(T, \eta) \asymp \frac{1}{\log V} \left(\log \frac{V}{q} + \frac{1}{2} \right) = 1 - \frac{\log q}{\log V} + \frac{1}{2 \log V} \asymp 1 - \frac{\log q}{\log V}.$$

If $T \geq T_{\text{crit}}(\eta, V)$, then $i_*(\eta, T) = V$ and

$$r_{\text{bias}}(T, \eta) \asymp \frac{1}{zq^2} \sum_{i=1}^{V} i \asymp \frac{V^2}{q^2 \log V}.$$

Case 3: $\alpha > 1$**.** If $T \leq T_{\text{crit}}(\eta, V)$, then $i_*(\eta, T) \asymp q^{1/\alpha}$ and, since $z \asymp \zeta(\alpha) = \Theta(1)$,

$$r_{\text{bias}}(T, \eta) \asymp q^{(1-\alpha)/\alpha}.$$

If $T \geq T_{\text{crit}}(\eta, V)$, then $i_*(\eta, T) = V$ and

$$r_{\text{bias}}(T, \eta) \asymp \frac{1}{q^2} \sum_{i=1}^{V} i^{\alpha} \asymp \frac{V^{\alpha+1}}{q^2}.$$

This completes the proof. ■

Appendix G. Proof of the Variance Scaling Law (Theorem 4)

Statement of Theorem 4 for the variance term. Fix $\varepsilon \in (0, 1)$ and assume

$$0 < \eta \leq (1 - \varepsilon) \frac{2bz}{b + z - 1}.$$

Let

$$\kappa := \eta / z, \qquad i_*(\eta, T) := \min \left\{ V, \left\lfloor (\kappa T)^{1/\alpha} \right\rfloor \right\}, \qquad T_{\text{crit}}(\eta, V) := \kappa^{-1} V^{\alpha}.$$

Then, with constants depending only on α and ε,

$$r_{\text{var}}(T, b, \eta) \asymp \frac{1}{b\mathcal{E}(0)} \left(\frac{i_*(\eta, T)}{T} + \kappa^2 T \sum_{i > i_*(\eta, T)} i^{-2\alpha} \right).$$

Proof Write $\mathcal{L}_{\text{var}}(T) = F_V(T) + G_V(T)$, where

$$F_V(T) := \frac{\mathcal{E}(0)}{T^2} \sum_{i=1}^{V} \pi_i \big(S_i^{(a)}(T) - \Gamma_i(T) \big)$$

and

$$G_V(T) := \frac{\eta^2 (1 - 2\mathcal{E}(0))}{2bT^2} \sum_{i=1}^{V} \frac{\pi_i^2}{1 - a_i} S_i^{(1-a)}(T).$$

Recall the row-wise effective learning rate $\delta_i = \eta \pi_i = \kappa i^{-\alpha}$, where $\kappa := \eta / z$, and the second-moment contraction factor $a_i = \phi_i^2 + \frac{\eta^2}{b} \pi_i (1 - \pi_i)$.

We organize the proof into six steps.

Step 1: Reduction to a one-row quantity. Define

$$u_i(T) := \frac{\delta_i^2}{(1-a_i)T^2} S_i^{(1-a)}(T) \qquad (1 \le i \le V). \tag{8}$$

Since $\delta_i = \eta\pi_i$, the second term is exactly

$$G_V(T) = \frac{1-2\mathcal{E}(0)}{2b} \sum_{i=1}^{V} u_i(T). \tag{9}$$

We next show that $F_V(T)$ is controlled by the same sum. Fix i. Because $a_i \ge \phi_i^2$ and all weights $w_{i,s}$ are nonnegative,

$$0 \le a_i^s - \phi_i^{2s} = (a_i - \phi_i^2) \sum_{r=0}^{s-1} a_i^r \phi_i^{2(s-1-r)} \le (a_i - \phi_i^2) \sum_{r=0}^{s-1} a_i^r = \frac{a_i - \phi_i^2}{1-a_i}(1-a_i^s).$$

After multiplying by $w_{i,s}$ and summing over s, we obtain

$$0 \le S_i^{(a)}(T) - \Gamma_i(T) \le \frac{a_i - \phi_i^2}{1-a_i} S_i^{(1-a)}(T). \tag{10}$$

Using $a_i - \phi_i^2 = \eta^2 b^{-1}\pi_i(1-\pi_i) \le \eta^2 b^{-1}\pi_i$, we get

$$0 \le \frac{\pi_i}{T^2}\big(S_i^{(a)}(T) - \Gamma_i(T)\big) \le \frac{\eta^2\pi_i^2}{b(1-a_i)T^2} S_i^{(1-a)}(T) = \frac{1}{b} u_i(T).$$

Therefore

$$0 \le F_V(T) \le \frac{\mathcal{E}(0)}{b} \sum_{i=1}^{V} u_i(T). \tag{11}$$

Combining (9) and (11) gives

$$\frac{1-2\mathcal{E}(0)}{2b} \sum_{i=1}^{V} u_i(T) \le \mathcal{L}_{\text{var}}(T) \le \frac{1}{2b} \sum_{i=1}^{V} u_i(T). \tag{12}$$

By Lemma 8, $2\mathcal{E}(0) = H_V^{(2\alpha)}/(H_V^{(\alpha)})^2 \to 0$ for $\alpha \le 1$, while for $\alpha > 1$, $2\mathcal{E}(0) \to \zeta(2\alpha)/\zeta(\alpha)^2 < 1$. Therefore (12) yields

$$\mathcal{L}_{\text{var}}(T) \asymp \frac{1}{b} \sum_{i=1}^{V} u_i(T). \tag{13}$$

It remains to estimate $\sum_i u_i(T)$.

Step 2: Comparison of $1-a_i$ with δ_i. Since

$$a_i = (1-\delta_i)^2 + \frac{\eta\delta_i}{b}(1-\pi_i),$$

we have

$$1 - a_i = 2\delta_i - \delta_i^2 - \frac{\eta\delta_i}{b}(1-\pi_i).$$

The theorem assumes a fixed margin inside the exact stability region:

$$\eta \le (1-\varepsilon)\frac{2bz}{b+z-1}.$$

Equivalently,

$$\eta\left(\pi_1 + \frac{1-\pi_1}{b}\right) \le 2(1-\varepsilon).$$

Since $\pi_i \le \pi_1$ and the function $p \mapsto p + (1-p)/b$ is increasing for $b \ge 1$, we have

$$\eta\left(\pi_i + \frac{1-\pi_i}{b}\right) \le 2(1-\varepsilon) \qquad \text{for every } i.$$

Therefore

$$2 - \eta\pi_i - \frac{\eta}{b}(1-\pi_i) \ge 2\varepsilon,$$

while the same factor is at most 2. Hence

$$2\varepsilon\,\delta_i \le 1 - a_i \le 2\delta_i. \tag{14}$$

Substituting into (8):

$$u_i(T) \asymp \frac{\delta_i}{T^2} S_i^{(1-a)}(T). \tag{15}$$

Step 3: Active rows ($\delta_i T \ge 1$). *Upper bound.* For every s,

$$w_{i,s} = 1 + 2\sum_{k=1}^{T-1-s} \phi_i^k \le 1 + 2\sum_{k=1}^{\infty}(1-\delta_i)^k \le \frac{2}{\delta_i},$$

so

$$S_i^{(1-a)}(T) \le \sum_{s=0}^{T-1} \frac{2}{\delta_i} = \frac{2T}{\delta_i}. \tag{16}$$

Lower bound. Fix $s \in [T/4, T/2]$. Then $T-1-s \ge T/2 - 1$, hence

$$w_{i,s} \ge \sum_{k=0}^{\lfloor T/2\rfloor - 1} (1-\delta_i)^k = \frac{1-(1-\delta_i)^{\lfloor T/2\rfloor}}{\delta_i} \ge \frac{1-e^{-\delta_i\lfloor T/2\rfloor}}{\delta_i} \ge \frac{c}{\delta_i},$$

where $c > 0$ is an absolute constant because $\delta_i T \ge 1$. Also, by (14),

$$s(1-a_i) \ge \frac{T}{4}\cdot\frac{\delta_i}{2} \ge \frac{1}{8},$$

so

$$1 - a_i^s \ge 1 - e^{-s(1-a_i)} \ge 1 - e^{-1/8} =: c' > 0.$$

Since there are $\asymp T$ indices s in $[T/4, T/2]$, we conclude

$$S_i^{(1-a)}(T) \gtrsim \frac{T}{\delta_i}. \tag{17}$$

Combining (16), (17), and (15):

$$\delta_i T \ge 1 \quad \Longrightarrow \quad u_i(T) \asymp \frac{1}{T}. \tag{18}$$

Step 4: Inactive rows ($\delta_i T \le 1/2$). *Upper bound.* For all s,

$$w_{i,s} \le 1 + 2(T-1-s) \le 2T.$$

Using $1 - a_i^s \le s(1-a_i) \lesssim s\delta_i$, we get

$$S_i^{(1-a)}(T) \le \sum_{s=0}^{T-1} 2T \cdot Cs\delta_i \lesssim \delta_i T^3. \tag{19}$$

Lower bound. Let again $s \in [T/4, T/2]$. If $0 \le k \le \lfloor T/2 \rfloor - 1$, then $k\delta_i \le T\delta_i/2 \le 1/4$, hence

$$(1-\delta_i)^k \ge 1 - k\delta_i \ge \frac{3}{4}.$$

Therefore

$$w_{i,s} \ge \sum_{k=0}^{\lfloor T/2 \rfloor - 1} (1-\delta_i)^k \gtrsim T.$$

Moreover, still for $s \in [T/4, T/2]$, we have by (14)

$$s(1-a_i) \le \frac{T}{2} \cdot 2\delta_i \le \frac{1}{2}.$$

Hence the quadratic remainder estimate gives

$$1 - a_i^s \ge s(1-a_i) - \binom{s}{2}(1-a_i)^2 \ge \frac{1}{2}s(1-a_i) \gtrsim s\delta_i \gtrsim T\delta_i.$$

Summing over $s \in [T/4, T/2]$, of which there are $\asymp T$:

$$S_i^{(1-a)}(T) \gtrsim \delta_i T^3. \tag{20}$$

Combining (19), (20), and (15):

$$\delta_i T \le \frac{1}{2} \quad \Longrightarrow \quad u_i(T) \asymp \delta_i^2 T = \kappa^2 T i^{-2\alpha}. \tag{21}$$

Step 5: Summation over i. Write $m = i_*(\eta, T)$. Equivalently, $m \asymp (\kappa T)^{1/\alpha}$ before it reaches V. Let

$$I_1 := \{i : \delta_i T \ge 1\} = \{i : i \le m\}, \qquad I_0 := \{i : \delta_i T \le 1/2\} = \{i : i \ge 2^{1/\alpha} m\} \cap [V].$$

By (18) and (21),

$$\sum_{i \in I_1} u_i(T) \asymp \frac{|I_1|}{T} \asymp \frac{m}{T}, \qquad \sum_{i \in I_0} u_i(T) \asymp \kappa^2 T \sum_{i \in I_0} i^{-2\alpha}. \tag{22}$$

Also, the upper bounds from Steps 3 and 4 imply that for every i,

$$u_i(T) \lesssim \min\left\{\frac{1}{T},\, \kappa^2 T i^{-2\alpha}\right\}.$$

Therefore,

$$\sum_{i=1}^{V} u_i(T) \lesssim \frac{m}{T} + \kappa^2 T \sum_{i>m} i^{-2\alpha}. \tag{23}$$

For the lower bound, from (22):

$$\sum_{i=1}^{V} u_i(T) \gtrsim \frac{m}{T} + \kappa^2 T \sum_{i \geq 2^{1/\alpha} m}^{V} i^{-2\alpha}.$$

The middle block between m and $2^{1/\alpha}m$ satisfies

$$\sum_{i=m+1}^{\lfloor 2^{1/\alpha} m \rfloor} i^{-2\alpha} \leq (\lfloor 2^{1/\alpha} m \rfloor - m) \cdot m^{-2\alpha} \lesssim m^{1-2\alpha}.$$

Multiplying by $\kappa^2 T$ and using $m^\alpha \asymp \kappa T$:

$$\kappa^2 T \cdot m^{1-2\alpha} = \kappa^2 T \frac{m}{m^{2\alpha}} \asymp \kappa^2 T \frac{m}{\kappa^2 T^2} = \frac{m}{T}.$$

Now decompose the tail:

$$\kappa^2 T \sum_{i>m}^{V} i^{-2\alpha} = \kappa^2 T \sum_{i=m+1}^{\lfloor 2^{1/\alpha} m \rfloor} i^{-2\alpha} + \kappa^2 T \sum_{i > \lfloor 2^{1/\alpha} m \rfloor}^{V} i^{-2\alpha} \leq C \frac{m}{T} + \kappa^2 T \sum_{i \geq 2^{1/\alpha} m}^{V} i^{-2\alpha}.$$

Therefore

$$\sum_{i=1}^{V} u_i(T) \gtrsim \frac{m}{T} + \kappa^2 T \sum_{i \geq 2^{1/\alpha} m}^{V} i^{-2\alpha} \gtrsim \frac{m}{T} + \kappa^2 T \sum_{i>m}^{V} i^{-2\alpha}, \tag{24}$$

because the only extra part in $\sum_{i>m}$ is exactly the middle block whose contribution is $\lesssim m/T$, already present.

Combining (23) and (24) with (13) proves

$$\mathcal{L}_{\text{var}}(T) \asymp \frac{1}{b} \left(\frac{i_*(\eta, T)}{T} + \kappa^2 T \sum_{i > i_*(\eta, T)} i^{-2\alpha} \right).$$

Step 6: Regime-wise asymptotics. We convert the active-block plus inactive-tail expression into closed-form scalings without normalizing κ. Write $q := \kappa T = \eta T / z$. In the subcritical regime $T \leq T_{\text{crit}}(\eta, V)$, equivalently $q \leq V^\alpha$, we have $i_*(\eta, T) \asymp q^{1/\alpha}$.

Subcritical regime $T \lesssim T_{\text{crit}}(\eta, V)$. Here the active block contributes

$$\frac{i_*(\eta, T)}{T} \asymp \frac{q^{1/\alpha}}{T} = \kappa q^{1/\alpha - 1}.$$

For the tail, the integral test gives

$$\kappa^2 T \sum_{i > i_*}^{V} i^{-2\alpha} \asymp \begin{cases} \dfrac{\kappa q V^{1-2\alpha}}{1-2\alpha} - \dfrac{\kappa q\, i_*^{1-2\alpha}}{1-2\alpha} = \dfrac{\kappa q V^{1-2\alpha}}{1-2\alpha} - \dfrac{\kappa q^{1/\alpha - 1}}{1-2\alpha}, & 0 < \alpha < \frac{1}{2}, \\ \kappa q (\log V - \log i_*) = \kappa q \big(\log V - 2 \log q\big), & \alpha = \frac{1}{2}, \\ \dfrac{\kappa q\, i_*^{1-2\alpha}}{2\alpha - 1} - \dfrac{\kappa q V^{1-2\alpha}}{2\alpha - 1} = \dfrac{\kappa q^{1/\alpha - 1}}{2\alpha - 1} - \dfrac{\kappa q V^{1-2\alpha}}{2\alpha - 1}, & \alpha > \frac{1}{2}, \end{cases} \tag{25}$$

Adding the active block $\kappa q^{1/\alpha-1}$ (which absorbs into the leading constant) gives

$$\frac{i_*}{T} + \kappa^2 T \sum_{i>i_*}^{V} i^{-2\alpha} \asymp \begin{cases} \kappa\left(qV^{1-2\alpha} - c_\alpha\, q^{1/\alpha-1}\right), & 0<\alpha<\frac{1}{2}, \\ \kappa q\big(1+\log V - 2\log q\big), & \alpha=\frac{1}{2}, \\ \kappa\left(q^{1/\alpha-1} - c_\alpha\, qV^{1-2\alpha}\right), & \alpha>\frac{1}{2}, \end{cases} \tag{26}$$

with $c_\alpha > 0$ an explicit α-dependent constant.

We then use ratio comparison to select the dominant term.

For the $\alpha \neq 1/2$ cases, (26) is a difference $A_\alpha(T,V) - B_\alpha(T,V)$ of two positive terms. A direct computation gives

$$\frac{B_\alpha}{A_\alpha} = \begin{cases} \dfrac{c_\alpha\, q^{1/\alpha-1}}{qV^{1-2\alpha}} = c_\alpha\big(q/V^\alpha\big)^{1/\alpha-2}, & 0<\alpha<\frac{1}{2}, \\ \dfrac{c_\alpha\, qV^{1-2\alpha}}{q^{1/\alpha-1}} = c_\alpha\big(q/V^\alpha\big)^{2-1/\alpha}, & \alpha>\frac{1}{2}. \end{cases}$$

In both cases the exponent $p_\alpha > 0$, so under the subcritical assumption $q \lesssim V^\alpha$ we have $B_\alpha/A_\alpha \asymp (q/V^\alpha)^{p_\alpha} = O(1)$. Hence

$$A_\alpha - B_\alpha = A_\alpha\big(1 - B_\alpha/A_\alpha\big) \asymp A_\alpha,$$

which identifies the dominant term as A_α.

For $\alpha = 1/2$, the bracket $1 + \log V - 2\log q$ is dominated by $\log V$ since $\log q \lesssim \log V$, so the same principle applies with $A = \kappa q \log V$.

Combining these with the prefactor $1/b$ in the bound for $\mathcal{L}_{\text{var}}(T)$ yields the subcritical scalings

$$\mathcal{L}_{\text{var}}(T) \asymp \frac{\kappa}{b} \cdot \begin{cases} qV^{1-2\alpha} & 0<\alpha<1/2, \\ q\log V & \alpha=1/2, \\ q^{1/\alpha-1} & \alpha>1/2, \end{cases} \qquad T \lesssim T_{\text{crit}}(\eta, V). \tag{27}$$

Supercritical regime $T \gg T_{\text{crit}}(\eta, V)$. Here $i_*(\eta, T) = V$, so the inactive tail in the surrogate is empty and only the active block remains:

$$\mathcal{L}_{\text{var}}(T) \asymp \frac{1}{b} \cdot \frac{V}{T}.$$

Finally, dividing the active-block plus inactive-tail estimate by $\mathcal{E}(0)$ yields the relative-variance formula stated in Theorem 4. Dividing the displayed subcritical and supercritical forms above by $\mathcal{E}(0)$, and then applying Lemma 8, gives the corresponding regime-wise relative-variance asymptotics. ■

Appendix H. Proofs for Critical Batch Size

H.1. Proof of the Zipf intermediate-compute scaling

We specialize Theorem **??** to $\alpha = 1$. Then $z = H_V \asymp \log V$ and $\mathcal{E}(0) \asymp (\log V)^{-2}$. In the subcritical regime

$$\frac{\eta T}{z} \ll V,$$

the active cutoff satisfies $i_*(\eta, T) \asymp \eta T/z$. The bias surrogate is

$$r_{\text{bias}}(T, \eta) \asymp \frac{1}{z}\left(\sum_{i>i_*(\eta,T)}^{V} i^{-1} + \frac{z^2}{\eta^2 T^2}\sum_{i\le i_*(\eta,T)} i\right).$$

Using

$$\sum_{i>m}^{V} i^{-1} \asymp \log(V/m), \qquad \sum_{i\le m} i \asymp m^2,$$

and $m = i_*(\eta, T) \asymp \eta T/z$, the active contribution is order $1/z$, while the tail contribution is $\log(Vz/(\eta T))/z$. Hence

$$r_{\text{bias}}(T, \eta) \asymp 1 - \frac{\log(\eta T/z)}{\log V}.$$

For the variance,

$$r_{\text{var}}(T, b, \eta) \asymp \frac{1}{b\mathcal{E}(0)}\left(\frac{i_*(\eta, T)}{T} + \frac{\eta^2 T}{z^2}\sum_{i>i_*(\eta,T)} i^{-2}\right).$$

The first term equals

$$\frac{i_*(\eta, T)}{T} \asymp \frac{\eta}{z}.$$

The second term satisfies

$$\frac{\eta^2 T}{z^2}\sum_{i>i_*(\eta,T)} i^{-2} \asymp \frac{\eta^2 T}{z^2}\cdot\frac{1}{i_*(\eta, T)} \asymp \frac{\eta}{z}.$$

Thus

$$r_{\text{var}}(T, b, \eta) \asymp \frac{1}{b\mathcal{E}(0)}\frac{\eta}{z} \asymp \frac{\eta}{bz}(\log V)^2.$$

Substituting $T = N/b$ gives the tuned fixed-compute surrogate

$$R(N/b, b, \eta) \asymp 1 - \frac{\log(\eta N/(bz))}{\log V} + \frac{\eta}{bz}(\log V)^2.$$

Optimizing over η before imposing the stability boundary gives

$$\frac{\partial}{\partial\eta}\left[-\frac{\log\eta}{\log V} + \frac{\eta}{bz}(\log V)^2\right] = -\frac{1}{\eta\log V} + \frac{(\log V)^2}{bz}.$$

Therefore

$$\eta_*(b) \asymp \frac{bz}{(\log V)^3}.$$

Equivalently,

$$\frac{\eta_*(b)N}{bz} \asymp \frac{N}{(\log V)^3}.$$

Thus the time-regime condition $1 \ll \eta T/z \ll V$ transforms, along the tuned fixed-compute curve, into the compute-regime condition

$$(\log V)^3 \ll N \ll V(\log V)^3.$$

The critical batch size is the point where the linear-scaling learning rate $\eta_*(b) \asymp bz/(\log V)^3$ reaches the stability boundary

$$\eta_{\max}(b) = \frac{2bz}{b+z-1}.$$

Equating these two scales and canceling bz gives

$$\frac{1}{(\log V)^3} \asymp \frac{1}{b+z-1}.$$

Since $z \asymp \log V$, this yields

$$b_{\text{crit}} \asymp (\log V)^3.$$

H.2. Proof of the Zipf supercritical scaling

When

$$N \gtrsim V(\log V)^3,$$

the intermediate optimum would make $\eta T/z$ at least order V, so the whole vocabulary is active. In the supercritical regime, Theorem **??** gives

$$r_{\text{bias}}(T,\eta) \asymp \frac{1}{z}\frac{z^2}{\eta^2 T^2}\sum_{i=1}^{V} i \asymp \frac{V^2 z}{\eta^2 T^2},$$

and with $z \asymp \log V$ this is equivalently

$$r_{\text{bias}}(T,\eta) \asymp \frac{V^2 z^2}{\eta^2 T^2 \log V}.$$

The variance has no inactive tail and satisfies

$$r_{\text{var}}(T,b,\eta) \asymp \frac{1}{b\mathcal{E}(0)}\frac{V}{T} \asymp \frac{V(\log V)^2}{bT}.$$

Along the fixed-compute curve $N = bT$, the variance is

$$r_{\text{var}} \asymp \frac{V(\log V)^2}{N}.$$

Once the learning rate has saturated at order z, further increases in b cannot be compensated by increasing η. Taking $\eta \asymp z$ and $T = N/b$, the bias becomes

$$r_{\text{bias}} \asymp \frac{V^2 b^2}{N^2 \log V}.$$

The critical batch is the point where this lost-steps bias reaches the fixed-compute variance:

$$\frac{V^2 b^2}{N^2 \log V} \asymp \frac{V(\log V)^2}{N}.$$

Solving for b gives

$$b_{\text{crit}} \asymp \sqrt{\frac{N}{V}}(\log V)^{3/2}.$$

This explains why bias–variance balancing reappears only after learning-rate saturation. Before saturation, increasing η preserves the effective optimization clock $\eta T/z$. After saturation, the variance remains fixed along $N = bT$, while the bias grows as the number of serial steps decreases.

H.3. Critical-batch scaling for other power-law exponents

We now record the same calculation for all $\alpha > 0$. The clean way to do the calculation is to use the effective clock

$$s := \frac{\eta T}{z} = \frac{\eta N}{bz}.$$

Let $s_*(N, V)$ denotes the effective clock selected by the loss surrogate, after optimizing over η for a given b. Along a fixed-compute curve, preserving s is exactly the linear learning-rate rule $\eta \propto b$. If the tuned subcritical optimizer selects a clock $s_*(N, V)$, then the linear rule reaches the stability boundary when

$$\frac{s_* z b}{N} \asymp \frac{bz}{b+z}, \qquad \text{hence} \qquad b_{\text{crit}} \asymp \frac{N}{s_*},$$

up to the lower-order subtraction by z. Thus the subcritical problem reduces to finding the clock s_* selected by the bias–variance surrogate.

Subcritical clocks. Assume first that $1 \ll s \ll V^\alpha$. Theorem **??** gives the following fixed-compute surrogates. For $0 < \alpha < 1$, the bias has the form

$$r_{\text{bias}}(s) \asymp 1 - c_\alpha \frac{s^{(1-\alpha)/\alpha}}{V^{1-\alpha}},$$

with the usual logarithmic replacement at $\alpha = 1$. The variance depends on whether the inactive tail $\sum_{i > s^{1/\alpha}} i^{-2\alpha}$ is controlled by the far end of the vocabulary or by the active boundary:

$$r_{\text{var}}(s) \asymp \begin{cases} \dfrac{s^2 V^{2(1-\alpha)}}{N}, & 0 < \alpha < \frac{1}{2}, \\ \dfrac{V s^2 \log(V/s^2)}{N \log V}, & \alpha = \frac{1}{2}, \\ \dfrac{V^{2(1-\alpha)} s^{1/\alpha}}{N}, & \frac{1}{2} < \alpha < 1, \\ \dfrac{s}{N}(\log V)^2, & \alpha = 1, \\ \dfrac{s^{1/\alpha}}{N}, & \alpha > 1. \end{cases}$$

Optimizing these one-dimensional surrogates over s yields the subcritical clocks in Table 4. The cases $\alpha \le 1/3$ are different: in the far-tail-dominated regime the bias improvement is no more linear than the variance penalty, so the subcritical surrogate has no useful interior minimizer. In that range the optimum is governed by an endpoint of the clock interval, and the clean scaling statement starts once the computation is large enough to enter the supercritical regime.

At the boundary $\alpha = 1/2$, the inactive-tail logarithm $\log(V/s^2)$ degenerates as s approaches the full-activation clock $V^{1/2}$. Thus the transition is only located up to this logarithmic boundary layer by the subcritical surrogate; Table 4 records the implicit relation and the corresponding $\widetilde{\Theta}(V^2)$ compute scale.

The cutoff $\alpha = 1/3$ comes from this exponent comparison. For $0 < \alpha < 1/2$, the subcritical surrogate has the schematic form

$$R(s) \asymp 1 - c_\alpha \frac{s^{(1-\alpha)/\alpha}}{V^{1-\alpha}} + C_\alpha \frac{s^2 V^{2(1-\alpha)}}{N}.$$

The bias gain grows like $s^{(1-\alpha)/\alpha}$, while the variance cost grows like s^2. An interior minimum exists only when $(1-\alpha)/\alpha < 2$, equivalently $\alpha > 1/3$. At $\alpha = 1/3$ the two powers are both s^2, so the surrogate is decided by the coefficient scale rather than by an interior first-order condition.

Supercritical balance. When $s \gtrsim V^\alpha$, all rows are active. After the learning rate has saturated at order z, Theorem **??** gives

$$r_{\text{bias}} \asymp \frac{b^2}{N^2} \frac{1}{z} \sum_{i=1}^{V} i^\alpha, \qquad r_{\text{var}} \asymp \frac{V}{\mathcal{E}(0) N}.$$

Balancing these two terms gives the general supercritical formula

$$b_{\text{crit}} \asymp \left(\frac{zVN}{\mathcal{E}(0) \sum_{i=1}^{V} i^\alpha} \right)^{1/2} \asymp \left(\frac{zN}{\mathcal{E}(0) V^\alpha} \right)^{1/2}. \tag{28}$$

Substituting the asymptotics of z and $\mathcal{E}(0)$ from Lemma 7 and Lemma 8 gives the last column of Table 4.

The table should be read as a scaling summary, not as a claim about sharp constant-factor crossovers. Near the boundaries between regimes, and in the endpoint-dominated cases $\alpha \le 1/3$, the exact closed-form risk from Theorem 3 is the right object for locating the transition numerically. The purpose of the table is to identify which mechanism controls the critical batch size: linear learning-rate scaling in the subcritical interior, and bias–variance balancing after the learning rate has saturated in the supercritical regime.

Table 4: Learning-rate-tuned critical-batch scaling for the linear bigram model. Here $s_* = \eta_*(b)N/(bz)$ is the optimized effective clock in the subcritical regime. The table suppresses constants depending only on α. The column N_{trans} is the compute scale at which the subcritical clock reaches V^α and the supercritical surrogate becomes the relevant description. Boundary logarithms are displayed for $\alpha = 1/2$ and $\alpha = 1$; at $\alpha = 1/2$, N_{trans} is an implicit logarithmic-boundary scale rather than a sharp power law.

Exponent	Subcritical clock s_*	N_{trans}	Subcritical CBS	Supercritical CBS
$0 < \alpha < \frac{1}{3}$	No interior subcritical optimum	V^2	No interior subcritical scaling	$V^{1-\alpha}N^{1/2}$
$\alpha = \frac{1}{3}$	Endpoint transition	V^2	Endpoint transition	$V^{2/3}N^{1/2}$
$\frac{1}{3} < \alpha < \frac{1}{2}$	$\left(N/V^{3(1-\alpha)}\right)^{\alpha/(3\alpha-1)}$	V^2	$N^{(2\alpha-1)/(3\alpha-1)}V^{3\alpha(1-\alpha)/(3\alpha-1)}$	$V^{1-\alpha}N^{1/2}$
$\alpha = \frac{1}{2}$	$\dfrac{N\log V}{V^{3/2}\log(V/s_*^2)}$	$\widetilde{\Theta}(V^2)$	$V^{3/2}\dfrac{\log(V/s_*^2)}{\log V}$	$(NV/\log V)^{1/2}$
$\frac{1}{2} < \alpha < 1$	$N/V^{3(1-\alpha)}$	$V^{3-2\alpha}$	$V^{3(1-\alpha)}$	$\left(NV^{3-4\alpha}\right)^{1/2}$
$\alpha = 1$	$N/(\log V)^3$	$V(\log V)^3$	$(\log V)^3$	$\left(N(\log V)^3/V\right)^{1/2}$
$\alpha > 1$	N	V^α	1	$(N/V^\alpha)^{1/2}$