FROM DATA IMBALANCE TO OPTIMIZER DIFFERENCES: A HESSIAN PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Prior work has used data imbalance to explain performance differences between optimizers, but frequency statistics alone are insufficient to characterize the optimization difficulties faced by different optimizers. In this work, we connect data imbalance to the blockwise Hessian structure that shapes optimizer behavior. Specifically, we characterize this structure through two properties—*block-scale heterogeneity* and *within-block anisotropy*—and find that optimizers respond differently to them: Adam cancels block-scale factors but does not necessarily alleviate within-block anisotropy, whereas Muon can somewhat mitigate the optimization difficulties arising from both properties.

To understand how these two properties arise, we study a two-layer linear bigram model. Under mean-squared error (MSE), the Hessians depend on the data distribution only through input frequencies: input imbalance induces block-scale heterogeneity in embedding training, while shaping within-block anisotropy in output-head training. Under cross-entropy (CE), output frequencies also play a role in shaping these properties. Together, these results show that data imbalance does not act alone: it interacts with the choice of trainable parameters and loss function to shape these two Hessian properties, thereby influencing the relative performance of different optimizers.

1 INTRODUCTION

Data imbalance has been proposed as an explanation for performance differences between optimizers. Prior studies connect heavy-tailed class frequencies to the advantage of Adam (Kingma & Ba, 2014) over gradient descent (GD) (Kunstner et al., 2024; Kunstner & Bach, 2025), and examine the benefits of Muon (Jordan et al., 2024) under imbalanced data (Wang et al., 2026b;a; Vasudeva et al., 2026). However, frequency statistics alone do not specify which optimization difficulties arise or which optimizer can address them: the effects of data imbalance depend on how it interacts with the network structure and loss function. Understanding this dependence calls for a finer-grained account of *how these interactions shape the optimization problem and how the resulting structure affects different optimizers*.

To develop such an account, we study the *blockwise structure of Hessian* as a link between data imbalance and optimization behavior. Prior work has shown that heterogeneity across Hessian blocks can be important for understanding differences in optimizer behavior (Zhang et al., 2024; 2025). We refine this perspective by distinguishing two properties of the blockwise Hessian structure: *block-scale heterogeneity*, which captures differences in curvature scale across blocks, and *within-block anisotropy*, which captures variation in curvature across directions within each block. We study how data imbalance shapes these properties in a controlled two-layer linear bigram model and how GD, Adam, and Muon respond to them.

Our analysis shows that whether data imbalance gives Adam or Muon an advantage over GD depends on how the imbalance manifests in the optimization geometry. Data imbalance alone does not determine this geometry; rather, its interaction with the trainable parameters and loss function shapes the resulting block-scale heterogeneity and within-block anisotropy. Adam and Muon, in turn, differ in how effectively they address these distinct sources of optimization difficulty.

Our main contributions are threefold:

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.

- **Layer-dependent effects of data imbalance.** In a two-layer bigram model under MSE, we derive exact Hessians for embedding and output-head training. We show that input imbalance induces block-scale heterogeneity in the former and shapes within-block anisotropy in the latter, while neither Hessian depends on output frequencies. These results identify how the choice of trainable parameters changes the optimization difficulties associated with the same input imbalance.
- **Optimizer responses to imbalance-induced curvature.** For the resulting quadratic problems, we derive one-step loss-contraction bounds for GD, Adam, and Muon. Specializing these bounds to the two layers distinguishes the effects of input imbalance that the optimizers can address: Adam cancels explicit block-scale factors but need not alleviate within-block anisotropy, whereas Muon can mitigate either source of ill-conditioning, subject to the current error geometry.
- **Loss-dependent effects.** Using a block-diagonal approximation, we extend the output-head Hessian analysis to cross-entropy, where output imbalance can also shape the blockwise structure.

Due to space constraints, we defer a more comprehensive discussion of related work to Appendix A.

2 A Controlled Two-Layer Linear Bigram Model

Bigram Data. Let $e_i \in \mathbb{R}^C$ be the one-hot vector for token i in a vocabulary of size C. For a bigram $(x, y) = (e_i, e_j) \sim \mathcal{D}$, denote the input and conditional target probabilities by $\boldsymbol{\pi}_x(i)$ and $\boldsymbol{\pi}_{y|x}(j \mid i)$, respectively. The output marginal is $\boldsymbol{\pi}_y(j) = \sum_i \boldsymbol{\pi}_x(i)\boldsymbol{\pi}_{y|x}(j \mid i)$. Input and output imbalance refer to nonuniform $\boldsymbol{\pi}_x$ and $\boldsymbol{\pi}_y$, respectively. We denote $D_{\boldsymbol{\pi}_x} := \mathrm{diag}(\boldsymbol{\pi}_x)$.

Two-Layer Neural Network and Losses. We study a two-layer linear network with hidden dimension d, embedding matrix $W = [w_1, \ldots, w_C] \in \mathbb{R}^{d\times C}$, and output head $V \in \mathbb{R}^{C\times d}$. We consider the population mean-squared error (MSE) and cross-entropy (CE) losses:

$$\mathcal{L}_{\mathrm{MSE}}(V, W) = \frac{1}{2}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\|VWx - y\|_2^2\right], \tag{1}$$

$$\mathcal{L}_{\mathrm{CE}}(V, W) = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[y^\top \log \mathrm{softmax}(VWx)\right]. \tag{2}$$

Training Configuration. To examine how the same data imbalance manifests in different network modules, under MSE we optimize one weight matrix at a time: W with V fixed, or V with W fixed. Freezing one layer lets us study how imbalance affects training of the other. For CE, we focus on training the output head V with W fixed. The resulting MSE subproblems are quadratic, and the CE head subproblem is convex. We optimize the population loss using exact gradients computed from the prescribed data distribution. This isolates the effects of data imbalance from minibatch sampling noise. This choice is motivated by evidence that the Adam–SGD gap can persist even in full-batch Transformer training (Kunstner et al., 2023). Appendix B specifies the shared experimental settings.

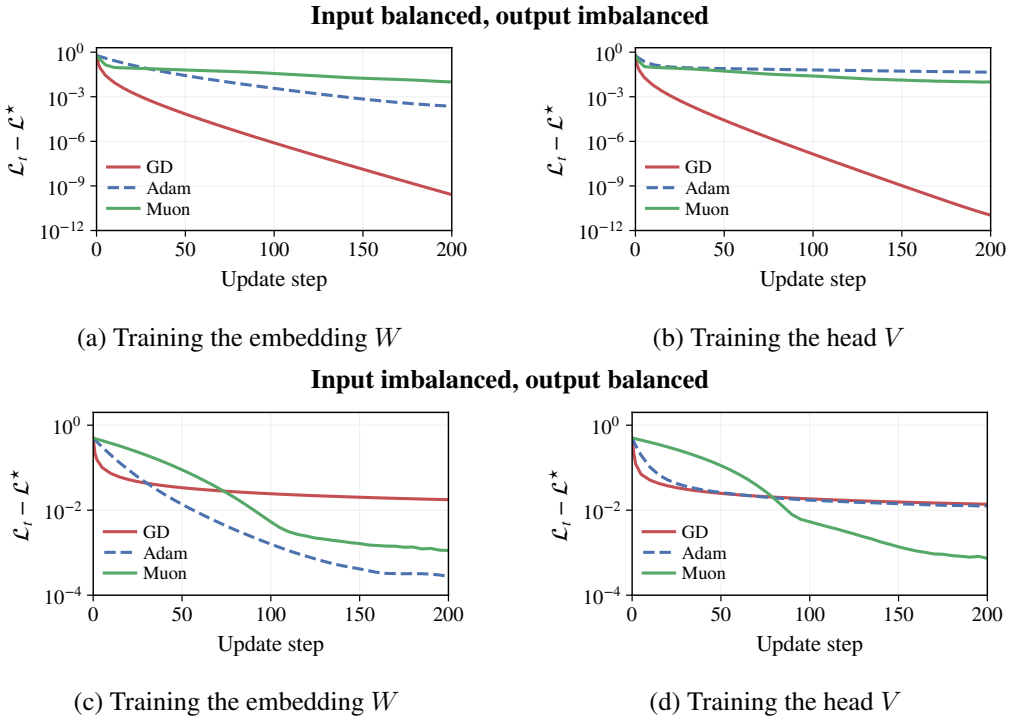
3 Data Imbalance and Optimization under MSE Loss

Explaining optimizer performance through data imbalance requires understanding how the data distribution affects the parameters being optimized. In our bigram model, embedding columns correspond to input tokens, whereas head rows correspond to output classes. This distinction motivates separating input from output imbalance and comparing their effects on the two layers.

We compare GD, Adam, and Muon under MSE on two distributions: balanced inputs with imbalanced outputs, and imbalanced inputs with balanced outputs. We set the vocabulary size to $C = 10000$ and the hidden dimension to $d = 5000$. For each distribution, we train either the embedding W or the output head V, keeping the other matrix fixed. All comparisons start from the same pair of Gaussian-initialized matrices (W_0, V_0). Experimental details are given in Appendix B.2; Appendix C presents all four combinations of balanced and imbalanced marginals.

Figure 1 yields two observations:

- **Input imbalance, rather than output imbalance, reveals an advantage over GD.** With balanced inputs and imbalanced outputs, GD outperforms both Adam and Muon. With imbalanced inputs and balanced outputs, Adam or Muon outperforms GD, depending on the trained layer.



(a) Training the embedding W (b) Training the head V

(c) Training the embedding W (d) Training the head V

Figure 1: MSE comparisons for GD, Adam, and Muon. Rows exchange the input/output marginals; columns change the trained layer. Each panel plots excess loss relative to its frozen-model optimum on a logarithmic scale.

- **The advantage under input imbalance depends on the trained layer.** When training the embedding, Adam attains the lowest final excess loss, followed by Muon and then GD. When training the head, Muon attains the lowest final excess loss, while Adam remains close to GD.

To understand these observations, we examine whether—and how—input and output frequencies enter the Hessian of each trainable matrix.

These observations suggest that understanding the effects of data imbalance requires examining the optimization difficulties it creates for each layer. Motivated by prior work linking Hessian block heterogeneity to optimizer behavior (Zhang et al., 2024), we examine how the data distribution shapes the blockwise Hessian structure of the two weight matrices.

3.1 Blockwise Hessian Structure of the Two Weight Matrices

We group parameters by their associated tokens: input-token columns for the embedding W and output-class rows for the head V. This grouping allows us to distinguish curvature variation across tokens from variation within each token's parameter block. With the other layer fixed at V_0 or W_0, differentiation of (1) gives the block Hessians

$$\mathcal{H}_{WW} = \begin{bmatrix} \boldsymbol{\pi}_x(1)V_0^\top V_0 & 0 & \cdots & 0 \\ 0 & \boldsymbol{\pi}_x(2)V_0^\top V_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\pi}_x(C)V_0^\top V_0 \end{bmatrix} = D_{\boldsymbol{\pi}_x} \otimes (V_0^\top V_0). \tag{3}$$

For head training, each block is the frequency-weighted second moment of the embeddings:

$$\mathcal{H}_{VV} = \begin{bmatrix} W_0 D_{\boldsymbol{\pi}_x} W_0^\top & 0 & \cdots & 0 \\ 0 & W_0 D_{\boldsymbol{\pi}_x} W_0^\top & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_0 D_{\boldsymbol{\pi}_x} W_0^\top \end{bmatrix} = I_C \otimes (W_0 D_{\boldsymbol{\pi}_x} W_0^\top). \tag{4}$$

Appendix E.1 provides the detailed calculations.

No dependence on output imbalance. Both Hessians depend on the data distribution only through the input frequencies $\boldsymbol{\pi}_x$, neither the output frequencies $\boldsymbol{\pi}_y$ nor the conditional target probabilities $\boldsymbol{\pi}_{y|x}$. Thus, with the input frequencies and frozen weights fixed, varying the output marginal—including its degree of imbalance—leaves both Hessians unchanged.

Different effects of input imbalance across layers. Both Hessians are block diagonal with C blocks of size $d \times d$, but input frequencies enter their blocks differently:

- **Embedding training: heterogeneity across blocks.** Input frequencies rescale the common within-block curvature $V_0^\top V_0$, creating unequal block scales while preserving a shared shape.
- **Head training: anisotropy within blocks.** All output rows share the block $W_0 D_{\boldsymbol{\pi}_x} W_0^\top$. Input frequencies shape the anisotropy of this common matrix through the frozen embeddings, rather than producing differences across blocks.

Table 1 reports the condition numbers of these Hessian factors for the shared Gaussian initialization used in Figure 1, where $\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$ for a positive-definite matrix A.

Table 1: Condition numbers of the MSE Hessian factors, using the shared frozen initialization. The input distribution determines these factors independently of the output distribution.

	Embedding Hessian		Head Hessian
Input distribution $\boldsymbol{\pi}_x$	$\kappa(D_{\boldsymbol{\pi}_x})$	$\kappa(V_0^\top V_0)$	$\kappa(W_0 D_{\boldsymbol{\pi}_x} W_0^\top)$
Balanced	1	33.5	33.7
Imbalanced	10^8	33.5	1.4×10^8

For embedding training, input imbalance increases the block-scale ratio from 1 to 10^8, while the within-block condition number remains 33.5. For head training, the blocks remain identical, but their shared condition number increases from 33.7 to 1.4×10^8.

These Hessian structures distinguish two optimization difficulties: unequal scales across blocks and ill-conditioning within each block. In the next section, we will analyze how GD, Adam, and Muon respond to these factors under its stated update rules.

4 Convergence Analysis on Quadratics with a Block Hessian

We next connect the blockwise Hessian structures in Section 3 to one-step loss-contraction bounds for GD, Adam, and Muon. We study a general quadratic loss model that includes both frozen-layer training settings as special cases. To simplify the analysis, we use exact gradients without first-moment momentum and take the exact polar factor in the Muon update.

Center the trainable matrix at a least-squares minimizer and denote the error by $X \in \mathbb{R}^{C \times d}$, with rows $x_i^\top$. Let $S = \mathrm{diag}(s_1, \ldots, s_C) \succ 0$ contain the block scales and let $H = H^\top \succ 0$ in $\mathbb{R}^{d \times d}$ be the shared within-block curvature. In this section and Appendix F, we denote the loss above the frozen-layer optimum by f:

$$f(X) := \tfrac{1}{2} \operatorname{tr}(SXHX^\top) = \tfrac{1}{2} \sum_{i=1}^{C} s_i x_i^\top H x_i. \tag{5}$$

The gradient is $G = \nabla f(X) = SXH$, and the Hessian is

$$\mathcal{H} = \mathrm{blockdiag}_i(s_i H) = S \otimes H. \tag{6}$$

The two frozen-layer settings correspond to

- **Embedding training:** $X = (W - W^\star)^\top$, $S = D_{\boldsymbol{\pi}_x}$, and $H = V_0^\top V_0$.
- **Head training:** $X = V - V^\star$, $S = I_C$, and $H = W_0 D_{\boldsymbol{\pi}_x} W_0^\top$.

Appendix F gives the detailed proofs in this section.

4.1 GD

For a positive-definite quadratic, the classical GD convergence guarantee is governed by the Hessian condition number. In our setting, $\kappa(\mathcal{H}) = \kappa(S)\kappa(H)$, so both unequal block scales and ill-conditioned within-block curvature can slow convergence. The following theorem specializes this standard result to our block-structured objective.

Theorem 1 (One-step contraction for GD)**.** *Consider GD with the update $X_{t+1} = X_t - \eta S X_t H$ and the learning rate $\eta = [\lambda_{\max}(S)\lambda_{\max}(H)]^{-1}$. Then*

$$f(X_{t+1}) \le \left(1 - \frac{1}{\kappa(S)\kappa(H)}\right)^2 f(X_t). \tag{7}$$

Under input imbalance, the two training settings expose different sources of ill-conditioning. Embedding training has a large block-scale condition number $\kappa(S)$, whereas head training has identical block scales but a large within-block condition number $\kappa(H)$. GD addresses neither source, providing a common explanation for its slow progress in both settings in Figure 1(c-d).

4.2 ADAM

Unlike GD, Adam normalizes its gradients using their second-moment history. For our quadratic, this normalization cancels the explicit block-scale factor s_i. To see this, consider the row gradient $g_{i,t} = s_i H x_{i,t}$ and the second moments initialized by $v_{i,-1} = 0$:

$$v_{i,t} = \beta_2 v_{i,t-1} + (1-\beta_2) g_{i,t}^{\odot 2}, \qquad \widehat{v}_{i,t} = \frac{v_{i,t}}{1-\beta_2^{t+1}}, \tag{8}$$

where vector powers and square roots are coordinatewise. Since $g_{i,k} = s_i H x_{i,k}$ at every past step k, unfolding the recurrence factors s_i^2 out of $\widehat{v}_{i,t}$. The remaining diagonal normalizer is therefore

$$P_{i,t} := \frac{1}{s_i}\operatorname{diag}\left(\sqrt{\widehat{v}_{i,t}}\right) = \operatorname{diag}\left(\sqrt{\frac{1-\beta_2}{1-\beta_2^{t+1}}\sum_{k=0}^{t}\beta_2^{t-k}(Hx_{i,k})^{\odot 2}}\right). \tag{9}$$

Thus $P_{i,t}$ depends on H and the row's error history, with no explicit dependence on s_i. The same factor s_i appears in the gradient and its normalizing denominator, so the row update becomes

$$\begin{aligned} x_{i,t+1} &= x_{i,t} - \eta_t (s_i P_{i,t})^{-1} s_i H x_{i,t} \\ &= x_{i,t} - \eta_t P_{i,t}^{-1} H x_{i,t}. \end{aligned} \tag{10}$$

The remaining curvature is determined by H and the row-specific diagonal normalizers $P_{i,t}$. The following theorem quantifies contraction through the largest and smallest eigenvalues across these preconditioned blocks.

Theorem 2 (One-step contraction for Adam)**.** *Suppose every $P_{i,t} \succ 0$. Define the effective blocks and their spectral extrema by*

$$\begin{aligned} M_{i,t} &:= P_{i,t}^{-1/2} H P_{i,t}^{-1/2}, \\ \Lambda_{M,t} &:= \max_i \lambda_{\max}(M_{i,t}), \qquad \mu_{M,t} := \min_i \lambda_{\min}(M_{i,t}). \end{aligned}$$

Denote $\kappa_{M,t} := \Lambda_{M,t}/\mu_{M,t}$ and choose the learning rate $\eta_t = \Lambda_{M,t}^{-1}$. Then

$$f(X_{t+1}) \le \left(1 - \frac{1}{\kappa_{M,t}}\right)^2 f(X_t).$$

Unlike GD's bound, Adam's bound contains no explicit $\kappa(S)$ factor. This cancellation directly addresses the frequency-induced block-scale disparity in embedding training, where $S = D_{\boldsymbol{\pi}_x}$ and $H = V_0^\top V_0$ is comparatively better conditioned. In head training, however, $S = I_C$, so there is no block-scale disparity to remove. The difficulty instead lies in the shared block $H = W_0 D_{\boldsymbol{\pi}_x} W_0^\top$, whose ill-conditioning coordinatewise normalization is not guaranteed to alleviate. This distinction is consistent with the input-imbalanced MSE experiments: Adam improves over GD in embedding training but performs similarly to GD in head training.

4.3 MUON

We analyze a Muon variant that scales the polar direction by the gradient nuclear norm (Bernstein & Newhouse, 2024; Lau et al., 2025). For a compact SVD $G_t = U_t \Sigma_t R_t^\top$ containing its positive singular values, write $\mathrm{polar}(G_t) = U_t R_t^\top$ and update

$$X_{t+1} = X_t - \eta \|G_t\|_* \,\mathrm{polar}(G_t), \qquad \|G_t\|_* := \mathrm{tr}(\Sigma_t). \tag{11}$$

The base learning rate $\eta > 0$ is fixed, and we take $\mathrm{polar}(0) = 0$. The nuclear-norm factor makes the update vanish with the gradient, as discussed by Lau et al. (2025): since $\|\mathrm{polar}(G_t)\|_2 \le 1$, we have $\|X_{t+1} - X_t\|_2 \le \eta \|G_t\|_* \to 0$ as $G_t \to 0$. Here $\|\cdot\|_2$ denotes the spectral norm.

To describe Muon's contraction, we introduce a measurement on the dispersion of the error spectrum. For $A \succ 0$ in $\mathbb{R}^{n\times n}$ and nonzero $Z \in \mathbb{R}^{m\times n}$, define

$$\mathcal{D}_A(Z) := \frac{\mathrm{tr}(A)\,\mathrm{tr}(AZ^\top Z)}{\left[\mathrm{tr}\big(A(Z^\top Z)^{1/2}\big)\right]^2}. \tag{12}$$

Here A assigns each unit right singular direction r_j of Z the relative curvature weight $r_j^\top A r_j / \mathrm{tr}(A)$. Under these weights, $\mathcal{D}_A(Z)$ roughly measures the relative dispersion of the singular magnitudes of Z. For full-column-rank Z, it is bounded above by $\kappa_2(Z) = \sigma_{\max}(Z)/\sigma_{\min}(Z)$. Appendix F.3 gives more detailed discussions.

Theorem 3 (One-step contraction for Muon). *Suppose $X_t \in \mathbb{R}^{C\times d}$ is nonzero. Choose the learning rate $\eta = [\min\{\lambda_{\max}(S)\,\mathrm{tr}(H), \lambda_{\max}(H)\,\mathrm{tr}(S)\}]^{-1}$. Then*

$$f(X_{t+1}) \le \min\left\{1 - \frac{1}{\kappa(S)\mathcal{D}_H(S^{1/2}X_t)},\, 1 - \frac{1}{\kappa(H)\mathcal{D}_S(H^{1/2}X_t^\top)}\right\} f(X_t). \tag{13}$$

Unlike Adam's cancellation of block scales, Muon's bound offers a way to mitigate either source of ill-conditioning. In embedding training, the second term avoids the large explicit $\kappa(S)$ factor; in head training, the first term avoids the large explicit $\kappa(H)$ factor. These bounds suggest a possible mechanism for Muon's improvement in both input-imbalanced settings. Moreover, head training admits a particularly simple specialization because $S = I_C$.

Corollary 4 (Head training). *For $S = I_C$ and $d \le C$, suppose X_t has full column rank. With $\eta = 1/\mathrm{tr}(H)$, the update (11) satisfies*

$$f(X_{t+1}) \le \left(1 - \frac{1}{\mathcal{D}_H(X_t)}\right) f(X_t) \le \left(1 - \frac{1}{\kappa_2(X_t)}\right) f(X_t). \tag{14}$$

Thus the weighted-dispersion result refines the head-training condition-number bound.

5 EXTENSION TO CROSS-ENTROPY: PREDICTION-DEPENDENT CURVATURE

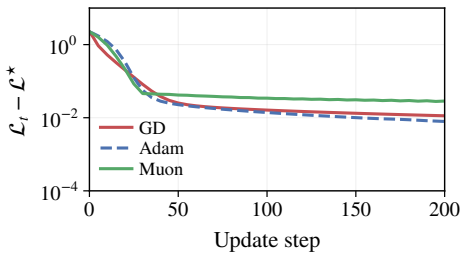
The quadratic analysis distinguishes unequal block scales from ill-conditioned within-block geometry and characterizes how these factors enter optimizer-specific contraction bounds. We now use this distinction to study cross-entropy.

We first characterize its blockwise structure and formulate hypotheses about optimizer-dependent responses to increasing imbalance, then test these hypotheses across a range of Zipf exponents.

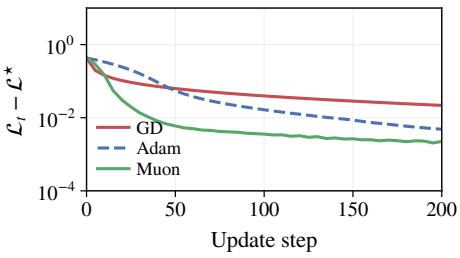
5.1 HESSIAN STRUCTURE OF THE OUTPUT HEAD

The CE head Hessian differs from its MSE counterpart in Section 3.1 in two respects. First, the softmax couples different output rows, introducing nonzero off-diagonal Hessian blocks. Second, the blocks depend on the current predictions and therefore evolve during training.

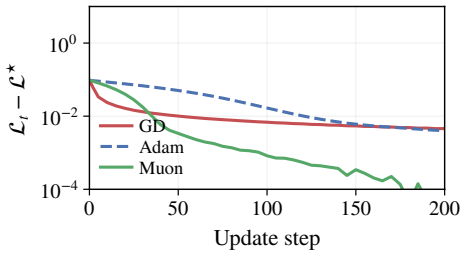
For input e_i, write $w_{0,i} := W_0 e_i$ and denote the predicted conditional probability by $\boldsymbol{p}_{y|x}(j \mid i; V) := [\mathrm{softmax}(V w_{0,i})]_j$. Here $\boldsymbol{p}$ denotes predictions and their induced distributions, whereas $\boldsymbol{\pi}$ denotes data probabilities; we suppress the parameter dependence below.



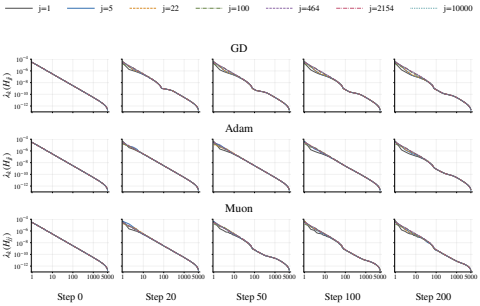
(a) $\zeta = 1$



(b) $\zeta = 1.5$



(c) $\zeta = 2$



(c') Output-head diagonal-block spectra

Figure 2: CE head training under input imbalance and balanced outputs. (a)–(c) excess-loss trajectories at $\zeta = 1, 1.5, 2$; (c') output-head diagonal-block spectra at $\zeta = 2$ (rows: GD, Adam, Muon; columns: training steps).

Near-block-diagonal Hessian structure has been observed and theoretically studied (Zhang et al., 2024; Dong et al., 2025). We therefore approximate the CE head Hessian using its diagonal blocks:

$$\mathcal{H}_{VV}^{\mathrm{CE}} \approx \widetilde{\mathcal{H}}_{VV}^{\mathrm{CE}} := \mathrm{blockdiag}(H_1, \ldots, H_C). \tag{15}$$

Here H_j is the exact diagonal block for output class j; Appendix E.2 gives the derivative calculations.

The current predictions induce the output marginal $\boldsymbol{p}_y(j) := \sum_i \boldsymbol{\pi}_x(i)\boldsymbol{p}_{y|x}(j \mid i)$ and reverse conditional $\boldsymbol{p}_{x|y}(i \mid j) := \boldsymbol{\pi}_x(i)\boldsymbol{p}_{y|x}(j \mid i)/\boldsymbol{p}_y(j)$. The exact blocks can therefore be written as

$$\begin{aligned} H_j &= \sum_{i=1}^{C} \boldsymbol{\pi}_x(i)\boldsymbol{p}_{y|x}(j \mid i)\big[1 - \boldsymbol{p}_{y|x}(j \mid i)\big] w_{0,i} w_{0,i}^\top \\ &= \boldsymbol{p}_y(j) \sum_i \boldsymbol{p}_{x|y}(i \mid j)\big[1 - \boldsymbol{p}_{y|x}(j \mid i)\big] w_{0,i} w_{0,i}^\top. \end{aligned} \tag{16}$$

Thus $\boldsymbol{p}_y(j)$ supplies an explicit block-scale factor. The reverse conditional $\boldsymbol{p}_{x|y}(i \mid j)$ and the factor $1 - \boldsymbol{p}_{y|x}(j \mid i)$ reweight the frozen embeddings and jointly shape the within-block geometry. To interpret this decomposition, we consider the block structure at initialization and during training:

- **Initialization:** At zero or sufficiently small i.i.d. random head initialization, predictions are approximately uniform, so $\boldsymbol{p}_{x|y}(i \mid j) \approx \boldsymbol{\pi}_x(i)$ and $H_j \approx \frac{1-1/C}{C} W_0 D_{\boldsymbol{\pi}_x} W_0^\top$. All blocks share the same input-frequency-weighted embedding geometry.
- **During training:** The predicted output marginal $\boldsymbol{p}_y$ is expected to approach the data marginal $\boldsymbol{\pi}_y$, and the conditional is expected to approach the data conditional $\boldsymbol{\pi}_{\cdot|\cdot}$, so H_j is expected to approach $\boldsymbol{\pi}_y(j) \sum_i \boldsymbol{\pi}_{x|y}(i \mid j)\big[1 - \boldsymbol{\pi}_{y|x}(j \mid i)\big] w_{0,i} w_{0,i}^\top$, whose scale and spectrum are determined by $\boldsymbol{\pi}$ and the embeddings.

Figures 2(c') and 3(c') later confirm this interpretation in the input- and output-imbalanced constructions, respectively.

5.2 Optimizer Responses to Increasing Imbalance

With the embedding fixed, we compare input- and output-imbalanced constructions, varying the Zipf exponent ζ while keeping the other marginal uniform. The block decomposition in Section 5.1 and the quadratic analysis in Section 4 motivate qualitative expectations about how relative optimizer performance changes as imbalance increases.

Increasing input imbalance. For this construction, we expect increasing ζ to strengthen within-block anisotropy while retaining comparable block spectra across output classes. In the quadratic model, Muon's contraction bound can avoid an explicit within-block condition-number factor, although it still depends on the current error geometry. Adam's coordinatewise normalization, in contrast, need not alleviate this anisotropy. This motivates the hypothesis that Muon improves relative to both Adam and GD as input imbalance increases.

Increasing output imbalance. As predictions reflect the imbalanced output marginal, the prefactors $\boldsymbol{p}_y(j)$ in the block decomposition can develop greater classwise disparities. For this construction, we expect stronger output imbalance to accentuate differences in block scale. In the quadratic model, Adam cancels fixed block-scale factors. This motivates the hypothesis that Adam improves relative to GD as output imbalance increases.

5.3 Testing the Hypotheses across Imbalance Levels

We evaluate the input- and output-imbalanced constructions at $\zeta \in \{1, 1.5, 2\}$, keeping the other marginal uniform and the embedding fixed. Figures 2 and 3 compare excess-loss trajectories across imbalance levels. The accompanying spectral plots show the evolution of the exact diagonal Hessian blocks at $\zeta = 2$, with the remaining exponents deferred to Appendix D.

To quantify relative performance at the common final checkpoint $T = 200$, we denote the excess loss by $E_o(\zeta; T) = \mathcal{L}_o(T; \zeta) - \mathcal{L}^\star(\zeta)$, where $\mathcal{L}^\star(\zeta)$ is the corresponding frozen-embedding optimum and $o \in \{G, A, M\}$ denotes GD, Adam, or Muon. For positive excess losses, define

$$R_{o/o'}(\zeta; T) := \log \frac{E_o(\zeta; T)}{E_{o'}(\zeta; T)}.$$

Positive values favor o'; we suppress T below.

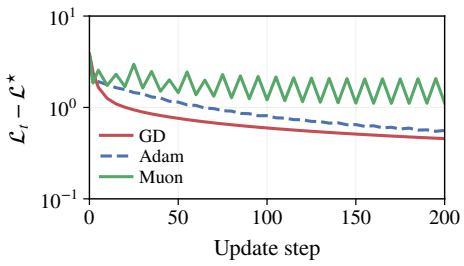
Input imbalance. Muon improves relative to both Adam and GD as input imbalance increases. Across the three exponents, $R_{A/M}$ increases from -1.29 to 0.77 to 6.35, while $R_{G/M}$ increases from -0.93 to 2.27 to 6.49. Thus, Muon moves from trailing both methods at $\zeta = 1$ to attaining the lowest final excess loss at $\zeta = 1.5$, with a substantially larger advantage at $\zeta = 2$. Adam's advantage over GD does not grow consistently, and their final excess losses are close at the strongest imbalance level.

Figure 2(c') shows the curvature encountered in the $\zeta = 2$ setting. The sampled blocks have spectra spanning many orders of magnitude from initialization onward, with substantial overlap across output classes. This agrees with the shared input-frequency-weighted geometry at initialization and within-block anisotropy during training described in Section 5.1

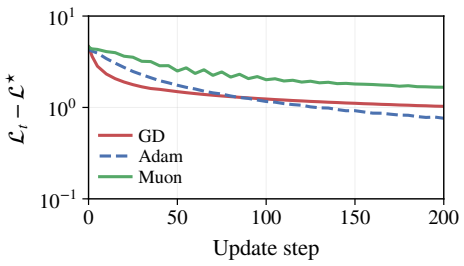
Output imbalance. Adam improves relative to GD as output imbalance increases: $R_{G/A}$ rises from -0.207 to 0.301 to 0.486 across the three exponents. GD has the lowest final excess loss at $\zeta = 1$, whereas Adam attains the lowest final excess loss among all three optimizers at $\zeta = 1.5$ and 2. Muon also improves relative to GD, with $R_{G/M}$ increasing from -0.896 to -0.481 to 0.293. It moves from trailing both methods at $\zeta = 1$ to attaining a lower final excess loss than GD at $\zeta = 2$, although its late-stage trajectory still fluctuates around GD.

Figure 3(c') shows a different spectral pattern at $\zeta = 2$. The sampled classwise spectra largely overlap at initialization and separate in scale during training, while the bulk of each spectrum stays relatively homogeneous This is consistent with the expected emergence of unequal block scales under output imbalance.

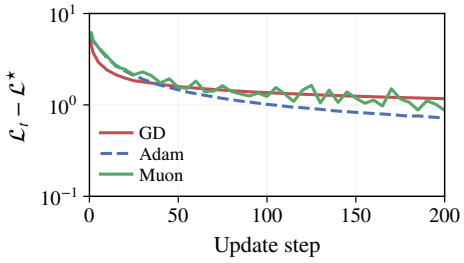
Together, the loss and spectral comparisons support the proposed distinction between within-block anisotropy and differences in block scale, while showing that optimizer rankings depend on imbalance strength.



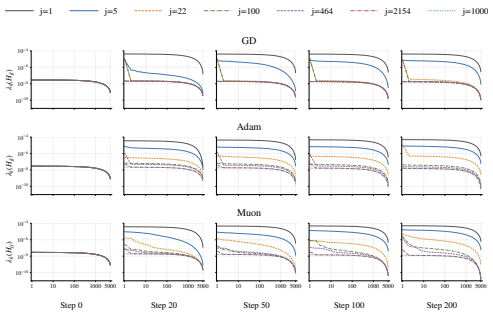
(a) $\zeta = 1$



(b) $\zeta = 1.5$



(c) $\zeta = 2$



(c') Output-head diagonal-block spectra

Figure 3: CE head training under output imbalance and balanced inputs. (a)–(c) excess-loss trajectories at $\zeta = 1, 1.5, 2$; (c') output-head diagonal-block spectra at $\zeta = 2$ (rows: GD, Adam, Muon; columns: training steps).

6 Discussion and Conclusion

We connect data imbalance to optimizer differences through two properties of the blockwise Hessian: block-scale heterogeneity and within-block anisotropy. In a two-layer linear bigram model, we show that the same frequency imbalance can induce different curvature structures depending on the trainable parameters and loss function. Our quadratic analysis characterizes how optimizers respond to these structures: under the analyzed updates, Adam cancels explicit block-scale factors but need not alleviate within-block anisotropy, whereas Muon can mitigate the difficulties associated with either property, with guarantees also depending on the current error geometry. Together, these findings suggest that understanding optimizer differences requires going beyond frequency statistics to examine how data imbalance shapes the curvature encountered during training.

Our study uses frozen-layer linear models, idealized updates for the quadratic analysis, and a block-diagonal approximation for CE. Extending this perspective to jointly trained nonlinear networks remains an important direction.

REFERENCES

Lukas Balles, Fabian Pedregosa, and Nicolas Le Roux. The geometry of sign gradient descent. *arXiv preprint arXiv:2002.08056*, 2020.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

Léon Bottou. Reconnaissance de la parole par reseaux connexionnistes. In *Proceedings of neuro Nimes*, volume 88, pp. 197–218, 1988.

Rudrajit Das, Naman Agarwal, Sujay Sanghavi, and Inderjit S Dhillon. Towards quantifying the preconditioning effect of adam. *arXiv preprint arXiv:2402.07114*, 2024.

Damek Davis and Dmitriy Drusvyatskiy. When do spectral gradient updates help in deep learning? *arXiv preprint arXiv:2512.04299*, 2025.

Zhaorui Dong, Yushun Zhang, Jianfeng Yao, and Ruoyu Sun. Towards quantifying the hessian structure of neural networks. *arXiv preprint arXiv:2505.02809*, 2025.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Frederik Kunstner and Francis Bach. Scaling laws for gradient descent and sign descent for linear bigram models under zipf's law. *Advances in Neural Information Processing Systems*, 38:144435–144467, 2025.

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. *arXiv preprint arXiv:2304.13960*, 2023.

Frederik Kunstner, Alan Milligan, Robin Yadav, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why adam outperforms gradient descent on language models. *Advances in Neural Information Processing Systems*, 37:30106–30148, 2024.

Tim Tsz-Kit Lau, Qi Long, and Weijie Su. Polargrad: A class of matrix-gradient optimizers from a unifying preconditioning perspective. *arXiv preprint arXiv:2505.21799*, 2025.

Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pp. 9–50. Springer, 2002.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in muon. *arXiv preprint arXiv:2601.13474*, 2026.

Martin Marek, Sanae Lotfi, Aditya Somasundaram, Andrew Wilson, and Micah Goldblum. Small batch size training for language models: When vanilla sgd works, and why gradient accumulation is wasteful. *Advances in Neural Information Processing Systems*, 38:148837–148862, 2025.

Yan Pan and Yuanzhi Li. Toward understanding why adam converges faster than sgd for transformers. *arXiv preprint arXiv:2306.00204*, 2023.

Egor Shulgin, Sam Laing, Antonio Orvieto, and Peter Richtárik. A quadratic lens on muon: Orthogonalization, invariance, and implicit preconditioning. In *High-dimensional Learning Dynamics*, 2026.

Jasraj Singh, Enea Monzio Compagnoni, and Antonio Orvieto. How the hessian-spectrum of neural networks depends on data. *arXiv preprint arXiv:2607.13631*, 2026.

Teodora Srećković, Jonas Geiping, and Antonio Orvieto. Is your batch size the problem? revisiting the adam-sgd gap in language modeling. *arXiv preprint arXiv:2506.12543*, 2025.

Akiyoshi Tomihari and Issei Sato. Gradient heterogeneity complements hessian heterogeneity in transformer optimization. *Transactions on Machine Learning Research*, 2026.

Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How muon's spectral design benefits generalization: A study on imbalanced data. In *International Conference on Learning Representations*, volume 2026, pp. 142276–142311, 2026.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Dirk Bergemann, and Zhuoran Yang. Why muon outperforms adam: A curvature perspective. *arXiv preprint arXiv:2606.04662*, 2026a.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Tan. Muon outperforms adam in tail-end associative memory learning. In *International Conference on Learning Representations*, volume 2026, pp. 37382–37419, 2026b.

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*, volume 2025, pp. 7937–7965, 2025.

Chenxiang Zhang, Rustem Islamov, Enea Monzio Compagnoni, Jun Pang, Aurelien Lucchi, and Antonio Orvieto. Beyond a single explanation of the adam–sgd gap. *arXiv preprint arXiv:2606.14259*, 2026a.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

Tianyue Zhang, Lucas Maes, Alan Milligan, Alexia Jolicoeur-Martineau, Ioannis Mitliagkas, Damien Scieur, Simon Lacoste-Julien, and Charles Guille-Escuret. Understanding adam requires better rotation dependent assumptions. *Advances in Neural Information Processing Systems*, 38:115111–115152, 2026b.

Tianyue H Zhang, Francis Bach, and Frederik Kunstner. Signal frequency imbalance and ill-conditioning. In *High-dimensional Learning Dynamics*, 2026c.

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, 2024.

Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik Durk Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. In *International Conference on Learning Representations*, volume 2025, pp. 28033–28063, 2025.

APPENDIX CONTENTS

A Related Work **13**

B Experimental Settings **14**

B.1 Experimental Design and Shared Configuration 14

B.2 Data Distributions 14

B.3 Optimizer Implementations 15

B.4 Learning-Rate Selection and Evaluation 15

C MSE across All Marginal Combinations **16**

D Additional CE Comparisons across Data Constructions **18**

E Gradients and Hessians for MSE and CE Losses **21**

E.1 MSE Loss 21

E.2 Cross-Entropy Loss 21

F Convergence Analysis and Proofs **23**

F.1 Quadratic Geometry and Gradient Descent 23

F.2 Adam: History Cancellation and Spectral Contraction 23

F.3 Muon: Nuclear-Norm Bounds and One-Step Contraction 24

G AI Use Statement **26**

A Related Work

Data imbalance and optimization. One line of work investigates data imbalance as an explanation for performance gaps between optimizers. Kunstner et al. (2024) connect heavy-tailed class frequencies to the slow learning of rare classes under gradient descent, while Kunstner & Bach (2025) quantify the separation between gradient descent and sign descent in linear bigram models under power-law token frequencies. Zhang et al. (2026c) relate ill-conditioning to signal-frequency imbalance that can arise, in part, from data imbalance. Studies of Muon further examine its advantages under heavy-tailed data imbalance, including balanced tail-end associative-memory learning (Wang et al., 2026b) and an amplified directional-curvature advantage (Wang et al., 2026a). We distinguish *input- and output-frequency imbalance* and analyze how their effects depend on the network's weight matrices, representations, and loss function. Through explicit Hessian characterizations and convergence analysis, we connect these interactions to the curvature structures encountered by different optimizers. This provides a geometric account of why data imbalance need not lead to the same optimization difficulties or optimizer gaps.

Hessian structure and optimization. Zhang et al. (2024) relate the Adam–SGD gap to Hessian block heterogeneity, explaining how a single learning rate can be restrictive when different blocks have different curvature spectra. Related work also examines gradient heterogeneity (Tomihari & Sato, 2026). Theoretically, Das et al. (2024) quantify the extent to which an Adam variant can mitigate ill-conditioning on quadratic objectives, with different guarantees for diagonal and diagonally dominant Hessians. The origin of Hessian structure has been studied from both architectural and data-dependent perspectives: Dong et al. (2025) analyze near-block-diagonality at random initialization, while Singh et al. (2026) relate the Hessian spectra of linear networks to data statistics. Our focus is on separating two sources of ill-conditioning: unequal scales *across* Hessian blocks and anisotropy *within* individual blocks. We derive how frequency imbalance shapes these structures and compare their effects on GD, Adam, and Muon.

Other perspectives on optimizer differences. Beyond data imbalance and Hessian structure, existing explanations of the Adam–SGD gap consider stochasticity in gradient estimates (Zhang et al., 2020; Marek et al., 2025; Srećković et al., 2025) and coordinate-dependent geometry (Balles et al., 2020; Pan & Li, 2023; Xie et al., 2025; Zhang et al., 2026b). More broadly, Zhang et al. (2026a) examine the interplay among data, architecture, and batch size, with empirical results suggesting that no single factor consistently explains the gap. For Muon and related spectral optimizers, recent analyses investigate the optimization benefits of spectral updates through their geometry and preconditioning effects (Davis & Drusvyatskiy, 2025; Ma et al., 2026; Shulgin et al., 2026).

B EXPERIMENTAL SETTINGS

This appendix specifies the shared settings for the MSE and CE experiments.

B.1 EXPERIMENTAL DESIGN AND SHARED CONFIGURATION

We use the two-layer linear model of Section 2, with $C = 10000$, $d = 5000$, embedding $W \in \mathbb{R}^{d\times C}$, and head $V \in \mathbb{R}^{C\times d}$. The model has no bias or hidden nonlinearity. Table 2 lists the three training settings, each compared across the five distributions using GD, Adam, and Muon. Freezing one layer isolates how the same data distribution affects optimization through the other.

Table 2: The three training settings; W_0, V_0 denote the initial matrices.

Loss	Trained matrix	Frozen matrix	Output
MSE	Embedding W	Head V_0	$V_0 W e_i$
MSE	Head V	Embedding W_0	$V W_0 e_i$
CE	Head V	Embedding W_0	$\mathrm{softmax}(V W_0 e_i)$

We use Bottou–LeCun weight initialization (Bottou, 1988; LeCun et al., 2002), with independent entries

$$(W_0)_{ij} \sim \mathcal{N}(0, 1/C), \qquad (V_0)_{ij} \sim \mathcal{N}(0, 1/d).$$

All comparisons share the same pair (W_0, V_0); each learning-rate candidate starts from these parameters and a fresh optimizer state.

Training uses the full population losses in (1) and (2), without sampling. MSE sums over output coordinates with factor $1/2$; CE uses natural logarithms.

B.2 DATA DISTRIBUTIONS

We construct all four combinations of balanced and imbalanced input/output marginals. Balanced marginals follow the uniform distribution u, and imbalanced marginals follow the Zipf distribution q with exponent two in the baseline comparisons:

$$u_i = \frac{1}{C}, \qquad q_i = \frac{i^{-2}}{\sum_{k=1}^{C} k^{-2}}, \qquad i = 1, \dots, C.$$

The CE comparisons in Section 5.3 vary the exponent over $\zeta \in \{1, 1.5, 2\}$, using $q_i = i^{-\zeta} / \sum_{k=1}^{C} k^{-\zeta}$ for the imbalanced marginal while keeping the other marginal uniform. Table 3 lists the four choices of input and output probability vectors $\boldsymbol{\pi}_x$ and $\boldsymbol{\pi}_y$.

Table 3: Four marginal combinations under the common coupling construction.

Input	Output	$\boldsymbol{\pi}_x$	$\boldsymbol{\pi}_y$
Balanced	Balanced	u	u
Balanced	Imbalanced	u	q
Imbalanced	Balanced	q	u
Imbalanced	Imbalanced	q	q

Joint and conditional distributions. For each pair $(\boldsymbol{\pi}_x, \boldsymbol{\pi}_y)$, let $m_i = \min\{\boldsymbol{\pi}_x(i), \boldsymbol{\pi}_y(i)\}$ and $\tau = 1 - \sum_i m_i$. Define a coupling that places their shared probability mass on the diagonal:

$$K(\boldsymbol{\pi}_x, \boldsymbol{\pi}_y) = \begin{cases} \mathrm{diag}(m) + \dfrac{(\boldsymbol{\pi}_y - m)(\boldsymbol{\pi}_x - m)^\top}{\tau}, & \tau > 0, \\ \mathrm{diag}(\boldsymbol{\pi}_x), & \tau = 0. \end{cases}$$

With output rows and input columns, the joint distribution $J_{ji} = \Pr(x = e_i, y = e_j)$ and conditional target matrix $\Pi_{ji} = \boldsymbol{\pi}_{y|x}(j \mid i)$ are

$$J = (1 - \rho)\boldsymbol{\pi}_y \boldsymbol{\pi}_x^\top + \rho K(\boldsymbol{\pi}_x, \boldsymbol{\pi}_y), \qquad \Pi = J\,\mathrm{diag}(\boldsymbol{\pi}_x)^{-1}. \tag{17}$$

This gives $J^\top \mathbf{1} = \boldsymbol{\pi}_x$, $J\mathbf{1} = \boldsymbol{\pi}_y$, and $\Pi\boldsymbol{\pi}_x = \boldsymbol{\pi}_y$; each column of Π sums to one.

The endpoints of (17) explain the construction:

- $\rho = 0$: x and y are independent, with $\Pi = \boldsymbol{\pi}_y \mathbf{1}^\top$.
- $\rho = 1$: the coupling maximizes $\Pr(x = y)$ subject to the prescribed marginals, attaining $\sum_i \min\{\boldsymbol{\pi}_x(i), \boldsymbol{\pi}_y(i)\}$. When $\boldsymbol{\pi}_x = \boldsymbol{\pi}_y$, it reduces to the deterministic mapping $y = x$; otherwise, it couples the remaining input and output mass.
- $0 < \rho < 1$: mixing these constructions preserves both marginals. All four distributions use $\rho = 0.8$ and have strictly positive joint probabilities.

Additional balanced data distribution: shifted conditionals. We also consider an additional data distribution with balanced input and output marginals but heterogeneous conditional probabilities across inputs:

$$\boldsymbol{\pi}_x = \boldsymbol{\pi}_y = u, \qquad \Pi_{ji} = \tilde{q}_{1+((i-j) \bmod C)},$$

where $\tilde{q}_i = i^{-1/2} / \sum_{k=1}^{C} k^{-1/2}$ defines a Zipf probability vector with exponent 0.5.

B.3 Optimizer Implementations

For the trainable matrix $\theta \in \{W, V\}$, write $G_t = \nabla_\theta \mathcal{L}(\theta_t)$. Learning rates are constant in the experiments, with no first-moment momentum, weight decay, or gradient clipping.

Adam. We set $\beta_2 = 0.999$, initialize $v_{-1} = 0$, and set the numerical stability constant to $\epsilon = 10^{-8}$. The bias-corrected update is

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) G_t^{\odot 2}, \qquad \widehat{v}_t = \frac{v_t}{1 - \beta_2^{t+1}},$$

$$\theta_{t+1} = \theta_t - \eta_t \frac{G_t}{\sqrt{\widehat{v}_t} + \epsilon},$$

where the square, square root, and division act entrywise.

Muon. We set the numerical stability constant to $\epsilon = 10^{-7}$. We normalize the gradient and apply five iterations of the Newton–Schulz map:

$$U_t = \frac{G_t}{\|G_t\|_F + \epsilon}, \qquad \phi(U) = 3.4445U - 4.7750(UU^\top)U + 2.0315(UU^\top)^2 U.$$

The parameter update uses the RMS alignment trick of Liu et al. (2025):

$$\theta_{t+1} = \theta_t - 0.2\eta\sqrt{\max(C, d)}\, \phi^{(5)}(U_t),$$

where $\phi^{(5)}$ denotes five successive applications of ϕ.

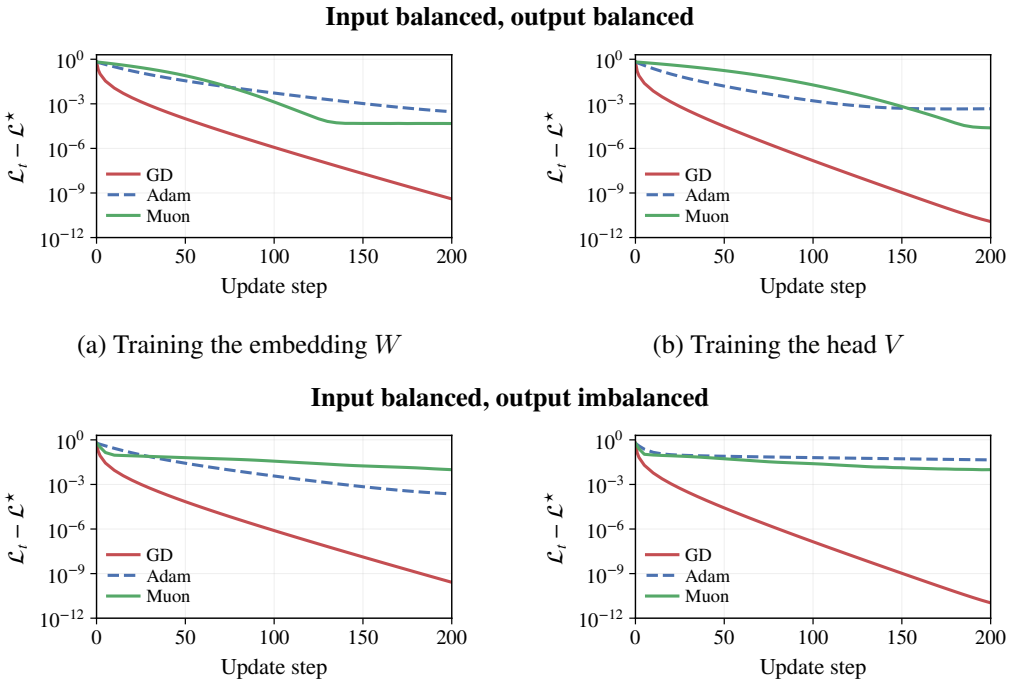
B.4 Learning-Rate Selection and Evaluation

We train for 200 steps. For each distribution, training setting, and optimizer, we search the grid $\{\gamma\, 10^k : \gamma \in \{1, 2, 5\},\ k \in \mathbb{Z}\}$ and select the learning rate that minimizes the loss at step 200. Search boundaries are extended by one decade until the selected rates are interior.

We report population loss $\mathcal{L}_t$ and excess loss $\mathcal{L}_t - \mathcal{L}^\star$, where $\mathcal{L}^\star$ is the optimum with the other layer frozen. MSE optima follow from the normal equations; numerical CE references are checked with primal–dual certificates.

C MSE ACROSS ALL MARGINAL COMBINATIONS

We complete the comparison in Section 3 with all four combinations of balanced and imbalanced input/output marginals under the coupling construction of Appendix B.2. Columns select the trained layer; rows vary the output distribution at a fixed input distribution.



(a) Training the embedding W

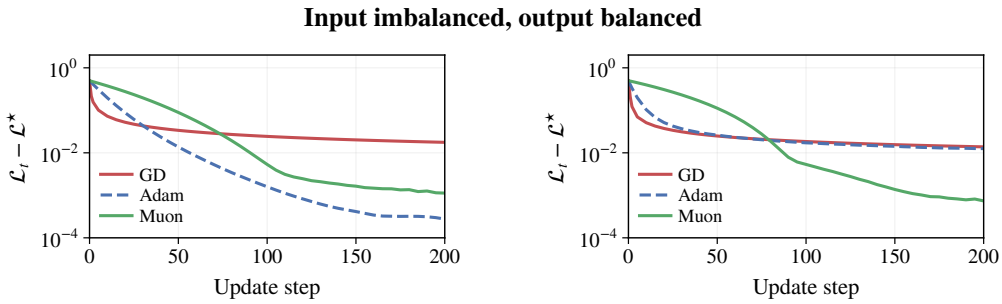
(b) Training the head V

(c) Training the embedding W

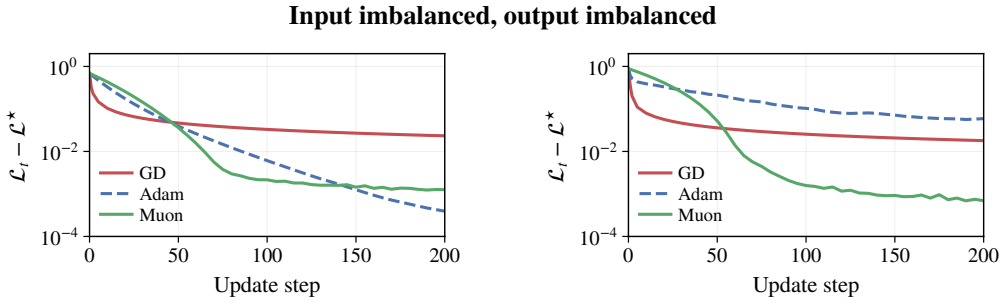
(d) Training the head V

Figure 4: MSE with balanced inputs and balanced or imbalanced outputs. Curves show excess loss relative to each frozen-model optimum.

Balanced inputs. GD attains a lower final excess loss than Adam and Muon in both layers, whether the outputs are balanced or imbalanced. Changing the output marginal changes the trajectories but preserves this ordering.



(a) Training the embedding W (b) Training the head V

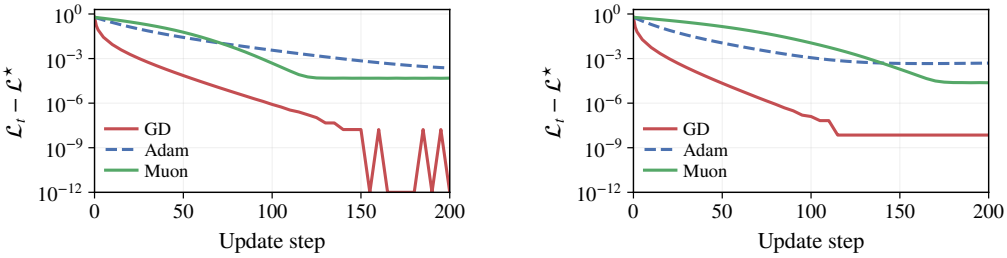


(c) Training the embedding W (d) Training the head V

Figure 5: MSE with imbalanced inputs and balanced or imbalanced outputs. Columns again select the trained layer, with the same excess-loss definition as Figure 4.

Imbalanced inputs. Both output distributions exhibit the layer-dependent pattern highlighted in Section 3: Adam and Muon attain much lower excess loss than GD when training the embedding, whereas Muon provides the largest improvement when training the head and Adam remains close to GD.

By (3)–(4), both Hessians depend only on the input frequencies and frozen weights, so Table 1 covers both output distributions in each input group. Since the conditional targets still set the optimum and initial error, equal Hessians do not imply identical trajectories: curvature, set by the input imbalance and trained layer, is not the only factor.



(a) Training the embedding W (b) Training the head V

Figure 6: MSE on the additional distribution with balanced input/output marginals and shifted conditional distributions, defined in Appendix B.2. Curves show excess loss relative to each frozen-model optimum.

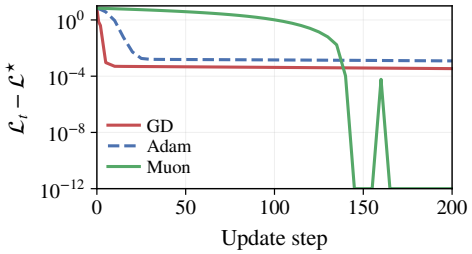
D ADDITIONAL CE COMPARISONS ACROSS DATA CONSTRUCTIONS

We extend the CE analysis in Section 5 to all four combinations of balanced and imbalanced marginals. All experiments train the output head V with the embedding fixed at W_0.

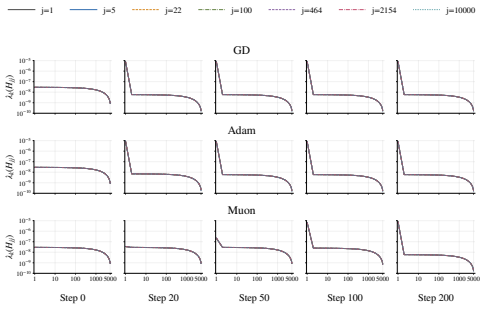
D.1 BLOCK STRUCTURE AND OPTIMIZER COMPARISONS

The comparisons below use $\zeta = 2$ for the Zipf distributions.

Balanced inputs and outputs. The uniform output marginal gives a common prefactor $\pi_y(j) = 1/C$, but the coupling still produces nonuniform reverse conditionals. In the target-implied blocks, this contributes a class-specific rank-one term to a shared embedding second moment. The spectral spikes in Figure 7 are qualitatively consistent with this structure, while the block scales remain comparable across classes.



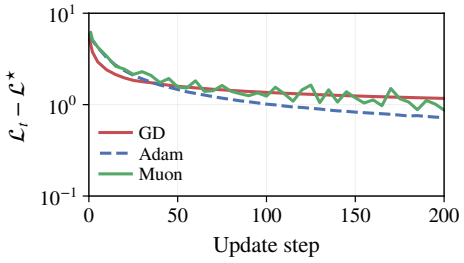
(a) Input balanced, output balanced



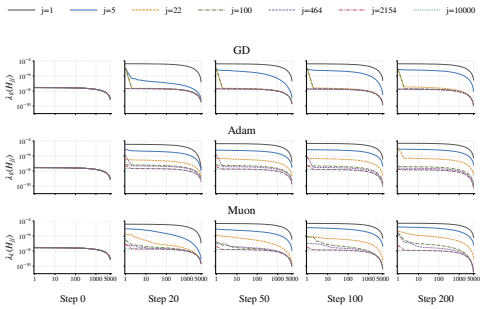
(a') Output-head diagonal-block spectra

Figure 7: CE head training with balanced inputs and outputs (left) and the corresponding output-head diagonal-block spectra (right; rows: GD, Adam, Muon; columns: training steps).

Balanced inputs, imbalanced outputs. The target-implied blocks carry unequal prefactors $\pi_y(j)$, suggesting scale separation across output classes. The spectra in Figure 8 exhibit this separation.



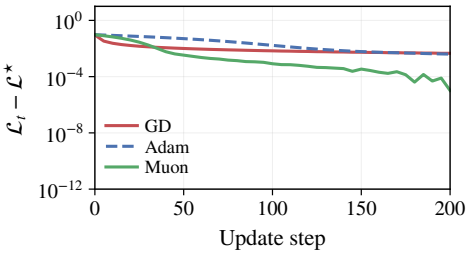
(b) Input balanced, output imbalanced



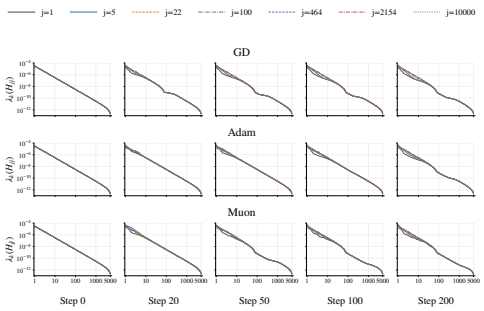
(b') Output-head diagonal-block spectra

Figure 8: CE head training with balanced inputs and imbalanced outputs (left) and the corresponding output-head diagonal-block spectra (right; rows: GD, Adam, Muon; columns: training steps).

Imbalanced inputs, balanced outputs. The uniform output marginal gives equal prefactors, while the reverse conditionals reweight the frozen embeddings. Pronounced anisotropy is already present at initialization, where nearly uniform predictions give $H_j(0) \approx W_0 D_{\pi_x} W_0^\top / C$.



(c) Input imbalanced, output balanced

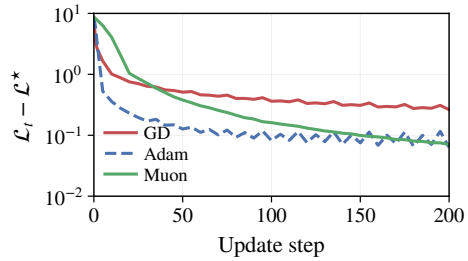


(c') Output-head diagonal-block spectra

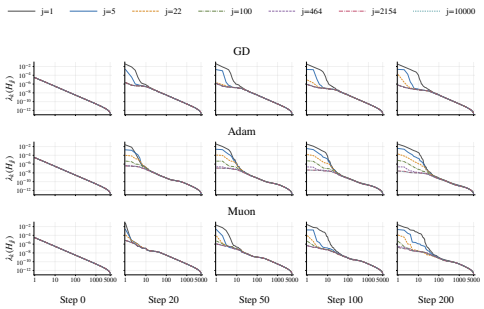
Figure 9: CE head training with imbalanced inputs and balanced outputs (left) and the corresponding output-head diagonal-block spectra (right; rows: GD, Adam, Muon; columns: training steps).

Imbalanced inputs and outputs. Both nonuniform output prefactors and input-frequency-weighted curvature enter the target-implied blocks.

The scale differences and broad within-block spectra in Figure 10 are qualitatively consistent with these two contributions. The accompanying loss trajectories examine their combined effect, rather than isolating either block-scale variation or within-block anisotropy.



(d) Input imbalanced, output imbalanced



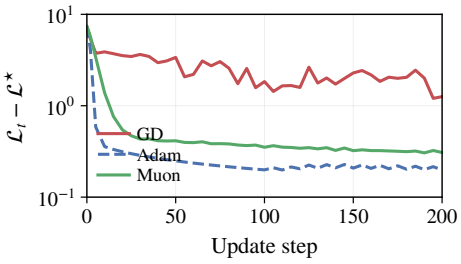
(d') Output-head diagonal-block spectra

Figure 10: CE head training with imbalanced inputs and imbalanced outputs (left) and the corresponding Output-head diagonal-block spectra (right; rows: GD, Adam, Muon; columns: training steps).

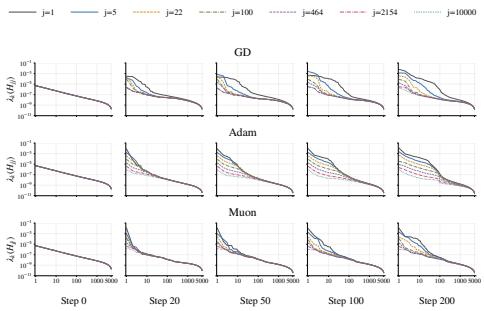
D.2 Dependence on the Zipf exponent

We supplement the detailed comparisons at $\zeta = 2$ with $\zeta \in \{1, 1.5\}$. Section 5.3 reports the two one-sided-imbalance families; the additional comparisons here cover imbalanced inputs and outputs. The ordinary both-balanced coupling is independent of ζ and is therefore shown only once.

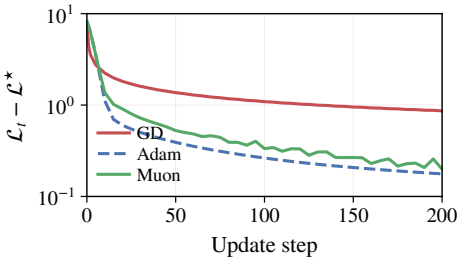
Imbalanced inputs and outputs.



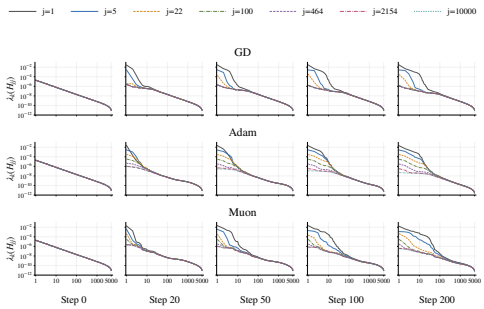
(f) $\zeta = 1$



(f') Diagonal-block spectra, $\zeta = 1$



(g) $\zeta = 1.5$



(g') Diagonal-block spectra, $\zeta = 1.5$

Figure 11: CE head training with imbalanced inputs and outputs at $\zeta = 1$ (top) and $\zeta = 1.5$ (bottom). Left: excess loss; right: output-head diagonal-block spectra (rows: GD, Adam, Muon; columns: training steps).

E GRADIENTS AND HESSIANS FOR MSE AND CE LOSSES

We differentiate the losses in Section 2 with respect to each trainable layer. Embedding-training expressions are evaluated at (V_0, W), and head-training expressions at (V, W_0). Write $w_i = We_i$, $z_i = Vw_i$, $w_{0,i} = W_0 e_i$, and $v_a^\top$ for row a of V; Hessian blocks are grouped by the columns of W and the rows of V.

E.1 MSE LOSS

Conditional decomposition. Since y is one-hot, $\mathbb{E}[\|y\|_2^2 \mid x = e_i] = 1$ and $\mathbb{E}[y \mid x = e_i] = \boldsymbol{\pi}_{y|x}(\cdot \mid i)$. Expanding the square yields

$$\mathbb{E}[\|z_i - y\|_2^2 \mid x = e_i] = \|z_i - \boldsymbol{\pi}_{y|x}(\cdot \mid i)\|_2^2 + 1 - \|\boldsymbol{\pi}_{y|x}(\cdot \mid i)\|_2^2.$$

Averaging over inputs rewrites the expectation in (1) as

$$\begin{aligned}\mathcal{L}_{\mathrm{MSE}}(V, W) = \tfrac{1}{2} \sum_i \boldsymbol{\pi}_x(i) \|z_i - \boldsymbol{\pi}_{y|x}(\cdot \mid i)\|_2^2 \\ + \tfrac{1}{2} \sum_i \boldsymbol{\pi}_x(i) \big(1 - \|\boldsymbol{\pi}_{y|x}(\cdot \mid i)\|_2^2\big).\end{aligned} \tag{18}$$

The second sum is independent of V and W.

Gradients. Collect the conditional target vectors into $\Pi := [\boldsymbol{\pi}_{y|x}(\cdot \mid 1), \ldots, \boldsymbol{\pi}_{y|x}(\cdot \mid C)] \in \mathbb{R}^{C \times C}$. Let $R_W = V_0 W - \Pi$ and $R_V = VW_0 - \Pi$ be the residual matrices for the two frozen-layer problems. Differentiating the weighted squared residuals in (18) gives

$$\begin{aligned}\nabla_W \mathcal{L}_{\mathrm{MSE}}(V_0, W) &= V_0^\top R_W D_{\boldsymbol{\pi}_x}, \\ \nabla_V \mathcal{L}_{\mathrm{MSE}}(V, W_0) &= R_V D_{\boldsymbol{\pi}_x} W_0^\top.\end{aligned} \tag{19}$$

Hessians. Differentiating (19) with the other layer fixed gives the (i, k) embedding block and (a, b) head block:

$$H^{\mathrm{MSE}}_{w_i w_k} = \delta_{ik} \boldsymbol{\pi}_x(i) V_0^\top V_0, \qquad H^{\mathrm{MSE}}_{v_a v_b} = \delta_{ab} S_{W_0},$$

Here δ is the Kronecker delta and $S_{W_0} = W_0 D_{\boldsymbol{\pi}_x} W_0^\top$. Stacking these blocks gives the block matrices and Kronecker forms in (3) and (4).

E.2 CROSS-ENTROPY LOSS

Gradients. For input e_i, the predicted conditional distribution is

$$\boldsymbol{p}_{y|x}(\cdot \mid i; V, W) = \mathrm{softmax}(z_i).$$

In each expression, this distribution is evaluated at the corresponding frozen-layer pair (V_0, W) or (V, W_0). It changes with the trained parameters; we suppress this dependence below and write $\boldsymbol{p}_i = \boldsymbol{p}_{y|x}(\cdot \mid i)$, whose jth entry is $\boldsymbol{p}_{y|x}(j \mid i)$. Write $P = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_C]$, let $\boldsymbol{\pi}_{y|x}(\cdot \mid i)$ be the fixed conditional target distribution given input e_i, and set $\Pi = [\boldsymbol{\pi}_{y|x}(\cdot \mid 1), \ldots, \boldsymbol{\pi}_{y|x}(\cdot \mid C)]$ and $D_{\boldsymbol{\pi}_x} = \mathrm{diag}(\boldsymbol{\pi}_x(1), \ldots, \boldsymbol{\pi}_x(C))$, where $\boldsymbol{\pi}_x(i)$ is the input frequency. Conditioning the expectation in (2) on the input gives

$$\mathcal{L}_{\mathrm{CE}}(V, W) = -\sum_i \boldsymbol{\pi}_x(i) \boldsymbol{\pi}_{y|x}(\cdot \mid i)^\top \log \boldsymbol{p}_{y|x}(\cdot \mid i). \tag{20}$$

Differentiating the input-i term of (20) gives $\nabla_{z_i} \mathcal{L}_{\mathrm{CE}} = \boldsymbol{\pi}_x(i)(\boldsymbol{p}_i - \boldsymbol{\pi}_{y|x}(\cdot \mid i))$. The chain rule then yields

$$\begin{aligned}\nabla_W \mathcal{L}_{\mathrm{CE}}(V_0, W) &= \sum_i \boldsymbol{\pi}_x(i) V_0^\top (\boldsymbol{p}_i - \boldsymbol{\pi}_{y|x}(\cdot \mid i)) e_i^\top = V_0^\top (P - \Pi) D_{\boldsymbol{\pi}_x}, \\ \nabla_V \mathcal{L}_{\mathrm{CE}}(V, W_0) &= \sum_i \boldsymbol{\pi}_x(i) (\boldsymbol{p}_i - \boldsymbol{\pi}_{y|x}(\cdot \mid i)) w_{0,i}^\top = (P - \Pi) D_{\boldsymbol{\pi}_x} W_0^\top.\end{aligned} \tag{21}$$

Hessians. The Hessian of the conditional cross-entropy with respect to the logits z_i is

$$B_i = \frac{\partial \boldsymbol{p}_i}{\partial z_i} = \operatorname{diag}(\boldsymbol{p}_i) - \boldsymbol{p}_i \boldsymbol{p}_i^\top,$$
$$(B_i)_{jk} = \boldsymbol{p}_{y|x}(j \mid i)\big(\delta_{jk} - \boldsymbol{p}_{y|x}(k \mid i)\big),$$

where δ_{jk} is the Kronecker delta. Writing $v_j^\top$ for row j of V, differentiating (21) with the other layer fixed gives

$$H^{\mathrm{CE}}_{w_i w_k} = \delta_{ik} \boldsymbol{\pi}_x(i) V_0^\top B_i V_0,$$
$$H^{\mathrm{CE}}_{v_j v_k} = \sum_i \boldsymbol{\pi}_x(i) \boldsymbol{p}_{y|x}(j \mid i)\big(\delta_{jk} - \boldsymbol{p}_{y|x}(k \mid i)\big) w_{0,i} w_{0,i}^\top.$$

Stacking these blocks with column-wise vectorization for W and row-wise vectorization for V gives the exact Hessians

$$\mathcal{H}^{\mathrm{CE}}_{WW} = \operatorname{blockdiag}_i\big(\boldsymbol{\pi}_x(i) V_0^\top B_i V_0\big),$$
$$\mathcal{H}^{\mathrm{CE}}_{VV} = \sum_i \boldsymbol{\pi}_x(i) B_i \otimes (w_{0,i} w_{0,i}^\top). \tag{22}$$

The block-diagonal approximation in (15) retains the exact diagonal output blocks of (22):

$$H_j := H^{\mathrm{CE}}_{v_j v_j} = \sum_i \boldsymbol{\pi}_x(i) \boldsymbol{p}_{y|x}(j \mid i)\big(1 - \boldsymbol{p}_{y|x}(j \mid i)\big) w_{0,i} w_{0,i}^\top,$$
$$\widetilde{\mathcal{H}}^{\mathrm{CE}}_{VV} = \operatorname{blockdiag}_j(H_j).$$

F Convergence Analysis and Proofs

This appendix proves the one-step contraction guarantees in Section 4. The argument proceeds from the quadratic GD baseline to the history-dependent Adam update, followed by the rectangular Muon refinement and its spectral interpretation.

F.1 Quadratic Geometry and Gradient Descent

At a frozen-layer minimizer, the gradients in (19) give the normal equations

$$V_0^\top (V_0 W^\star - \Pi) D_{\boldsymbol{\pi}_x} = 0, \qquad (V^\star W_0 - \Pi) D_{\boldsymbol{\pi}_x} W_0^\top = 0.$$

They eliminate the linear term in the quadratic expansion around either minimizer:

$$\mathcal{L}(W) - \mathcal{L}_W^\star = \tfrac{1}{2} \operatorname{tr}\big(D_{\boldsymbol{\pi}_x} (W - W^\star)^\top V_0^\top V_0 (W - W^\star)\big),$$
$$\mathcal{L}(V) - \mathcal{L}_V^\star = \tfrac{1}{2} \operatorname{tr}\big((V - V^\star) S_{W_0} (V - V^\star)^\top\big).$$

Here $S_{W_0} = W_0 D_{\boldsymbol{\pi}_x} W_0^\top$, and $\mathcal{L}_W^\star$ and $\mathcal{L}_V^\star$ denote the minima at fixed V_0 and fixed W_0, respectively. The identifications of X, S, H in Section 4 turn these identities into (5).

Let $x_i^\top$ denote row i of the centered error $X \in \mathbb{R}^{C \times d}$. With $S = \operatorname{diag}(s_1, \dots, s_C) \succ 0$ and $H \succ 0$ in $\mathbb{R}^{d \times d}$, equation (5) gives

$$f(X) = \tfrac{1}{2} \sum_i s_i x_i^\top H x_i, \qquad g_i = s_i H x_i.$$

Row-wise vectorization stacks these blocks, so $x = \operatorname{vec}_r(X)$ satisfies $f(X) = \frac{1}{2} x^\top \mathcal{H} x$, where $\mathcal{H} = S \otimes H$ as in (6). Its eigenvalues are $s_i \lambda_j(H)$. Consequently, with $\mu_{\mathrm{GD}} = \lambda_{\min}(S)\lambda_{\min}(H)$ and $\Lambda_{\mathrm{GD}} = \lambda_{\max}(S)\lambda_{\max}(H)$, its condition number is $\kappa_{\mathrm{GD}} = \kappa(S)\kappa(H)$. The function f measures the gap above the frozen-layer minimum, so $f(0) = 0$ even when the original minimum loss is positive.

Proof of Theorem 1. For GD, set $z = \mathcal{H}^{1/2} x$. Then $f(X) = \frac{1}{2}\|z\|_2^2$ and $z^+ = (I - \eta\mathcal{H})z$. Diagonalizing the SPD matrix $\mathcal{H}$ shows that

$$\sup_{X \neq 0} \frac{f(X^+)}{f(X)} = \max_{\lambda \in \operatorname{spec}(\mathcal{H})} |1 - \eta\lambda|^2.$$

At $\eta = 1/\Lambda_{\mathrm{GD}}$ the maximum is $(1 - 1/\kappa_{\mathrm{GD}})^2$, proving (7) and Theorem 1. For comparison, the optimal constant scalar step $2/(\Lambda_{\mathrm{GD}} + \mu_{\mathrm{GD}})$ balances the magnitudes at the two spectral endpoints and yields $((\kappa_{\mathrm{GD}} - 1)/(\kappa_{\mathrm{GD}} + 1))^2$. Both one-step factors retain the product $\kappa(S)\kappa(H)$.

F.2 Adam: History Cancellation and Spectral Contraction

Unfolding the recurrence in (8) with $v_{i,-1} = 0$, applying bias correction, and substituting $g_{i,k} = s_i H x_{i,k}$ gives

$$\widehat{v}_{i,t} = \frac{1 - \beta_2}{1 - \beta_2^{t+1}} \sum_{k=0}^{t} \beta_2^{t-k} g_{i,k}^{\odot 2}$$
$$= s_i^2 \frac{1 - \beta_2}{1 - \beta_2^{t+1}} \sum_{k=0}^{t} \beta_2^{t-k} (H x_{i,k})^{\odot 2}.$$

Since $s_i > 0$, taking coordinatewise square roots yields $\operatorname{diag}(\sqrt{\widehat{v}_{i,t}}) = s_i P_{i,t}$, where $P_{i,t}$ is defined in (9). The denominator and gradient therefore contain the same factor s_i, which cancels in (10) under the assumptions of Section 4.

Proof of Theorem 2. At the analyzed step, suppose $P_{i,t} \succ 0$ for every row. Define the loss coordinates $z_{i,t} = H^{1/2} x_{i,t}$ and the SPD matrix

$$A_{i,t} = H^{1/2} P_{i,t}^{-1} H^{1/2}.$$

Equation (10) becomes $z_{i,t+1} = (I - \eta_t A_{i,t}) z_{i,t}$. The matrices $A_{i,t}$ and $M_{i,t}$ have the same positive eigenvalues: for the invertible matrix $F_{i,t} = H^{1/2} P_{i,t}^{-1/2}$, they are $F_{i,t} F_{i,t}^\top$ and $F_{i,t}^\top F_{i,t}$, respectively. With $\eta_t = 1/\Lambda_{M,t}$, all eigenvalues of $I - \eta_t A_{i,t}$ lie in $[0, 1 - \mu_{M,t}/\Lambda_{M,t}]$. Hence

$$\begin{aligned} \|z_{i,t+1}\|_2^2 &\le \left(1 - \frac{1}{\kappa_{M,t}}\right)^2 \|z_{i,t}\|_2^2, \\ f(X_{t+1}) &= \tfrac{1}{2} \sum_i s_i \|z_{i,t+1}\|_2^2 \\ &\le \left(1 - \frac{1}{\kappa_{M,t}}\right)^2 f(X_t). \end{aligned}$$

This proves Theorem 2.

F.3 Muon: Nuclear-Norm Bounds and One-Step Contraction

We use the rectangular quadratic in (5) and the functional defined in (12). All positive semidefinite square roots are principal.

Proof of Theorem 3. Suppress the iteration index and set $G = SXH$. The compact polar factor satisfies

$$\operatorname{polar}(G)^\top \operatorname{polar}(G) \preceq I_d, \qquad \operatorname{polar}(G) \operatorname{polar}(G)^\top \preceq I_C, \qquad \langle G, \operatorname{polar}(G) \rangle_F = \|G\|_*.$$

Consequently,

$$\operatorname{tr}\big(S \operatorname{polar}(G) H \operatorname{polar}(G)^\top\big) \le \min\{\lambda_{\max}(S) \operatorname{tr}(H), \lambda_{\max}(H) \operatorname{tr}(S)\} = \eta^{-1},$$

and expanding the quadratic gives

$$f(X^+) \le f(X) - \tfrac{1}{2} \eta \|G\|_*^2. \tag{23}$$

For $Z = S^{1/2} X$, nuclear-norm submultiplicativity and spectral–nuclear duality give

$$\|G\|_* \ge \sqrt{\lambda_{\min}(S)}\, \|ZH\|_* \ge \sqrt{\lambda_{\min}(S)}\, \operatorname{tr}\big(H (Z^\top Z)^{1/2}\big).$$

The second inequality uses the dual witness $\operatorname{polar}(Z)$, for which $\|\operatorname{polar}(Z)\|_2 \le 1$ and $\operatorname{polar}(Z)^\top Z = (Z^\top Z)^{1/2}$. Squaring, substituting (12), and applying the same argument to $G^\top = H X^\top S$ yield

$$\begin{aligned} \|G\|_*^2 &\ge \frac{2 \lambda_{\min}(S) \operatorname{tr}(H)}{\mathcal{D}_H(S^{1/2} X)} f(X), \\ \|G\|_*^2 &\ge \frac{2 \lambda_{\min}(H) \operatorname{tr}(S)}{\mathcal{D}_S(H^{1/2} X^\top)} f(X). \end{aligned}$$

Since $\eta \ge [\lambda_{\max}(S) \operatorname{tr}(H)]^{-1}$ and $\eta \ge [\lambda_{\max}(H) \operatorname{tr}(S)]^{-1}$, combining these two bounds with (23) proves (13).

Weighted spectral dispersion.

Proposition 5. *Let $A \succ 0$ and $Z \neq 0$ be as in (12). Write $(Z^\top Z)^{1/2} = V \operatorname{diag}(\sigma_1, \ldots, \sigma_n) V^\top$, including zero eigenvalues, and set $w_j = v_j^\top A v_j / \operatorname{tr}(A)$, where v_j is column j of V. These weights are positive and sum to one. Then*

$$\mathcal{D}_A(Z) = \frac{\mathbb{E}_w[\sigma^2]}{\mathbb{E}_w[\sigma]^2} = 1 + \frac{\operatorname{Var}_w(\sigma)}{\mathbb{E}_w[\sigma]^2}, \qquad \mathcal{D}_A(Z) \leq \frac{\operatorname{tr}(A)}{\lambda_{\min}(A)}.$$

If Z has full column rank, then

$$\mathcal{D}_A(Z) \leq \kappa_2(Z). \tag{24}$$

Proof. Expanding both traces in the basis V gives the moment ratio. Since $w_j \geq \lambda_{\min}(A)/\operatorname{tr}(A)$ and $\sigma_j \geq 0$,

$$\left(\sum_j w_j \sigma_j \right)^2 \geq \sum_j w_j^2 \sigma_j^2 \geq \frac{\lambda_{\min}(A)}{\operatorname{tr}(A)} \sum_j w_j \sigma_j^2,$$

which proves the rank-independent upper bound. For full column rank, $\mathbb{E}_w[\sigma^2] \leq \sigma_{\max}(Z)\mathbb{E}_w[\sigma]$ and $\mathbb{E}_w[\sigma] \geq \sigma_{\min}(Z) > 0$, giving (24). □

Thus $\mathcal{D}_A(Z)$ measures relative dispersion across the full singular spectrum, with weights determined by A. The condition number controls its upper bound but does not determine it, since the intermediate singular values and their weights also matter.

Proof of Corollary 4. Since $\operatorname{tr}(H) \leq d\lambda_{\max}(H) \leq C\lambda_{\max}(H)$, the learning rate in Theorem 3 reduces to $1/\operatorname{tr}(H)$. For $S = I_C$, the first term in (13) is $1 - 1/\mathcal{D}_H(X_t)$, and (24) gives the second inequality in (14).

An example with a nearly flat spectrum. Suppose X has full column rank, $d \leq C$, and the positive singular values of $H^{1/2}X^\top$ are nearly equal. Let P_X project onto $\operatorname{col}(X)$. If this subspace also spreads approximately evenly across rows, so $(P_X)_{ii} \approx d/C$, then (12) gives

$$\mathcal{D}_S(H^{1/2}X^\top) \approx \frac{\operatorname{tr}(S)}{\operatorname{tr}(SP_X)} \approx \frac{C}{d}.$$

For $C = 10000, d = 5000$, the factor is approximately 2, even with nonuniform row scales. Both approximations become equalities when the positive singular values are equal and the projector has uniform diagonal entries.

G AI USE STATEMENT

Generative AI tools were used for language polishing and improving the clarity and readability of the manuscript, as well as for literature search and discovery of potentially relevant prior work. All suggested references were verified against the original sources, and all AI-assisted revisions were carefully reviewed by the authors. The authors are fully responsible for the accuracy and content of the final manuscript.