Too Sharp, Too Sure: When Calibration Follows Curvature

Anonymous Authors[1]

Abstract

Modern neural networks can achieve high accuracy while remaining poorly calibrated, producing confidence estimates that do not match empirical correctness. Yet calibration is often treated as a post-hoc attribute. We take a different perspective: we study calibration as a *training-time* phenomenon on small vision tasks, and ask how it co-evolves with loss-landscape geometry. We identify a tight coupling between calibration, curvature, and margins during training of deep networks under multiple gradient-based methods. Empirically, Expected Calibration Error (ECE) closely tracks curvature-based sharpness throughout optimization. Mathematically, we show that both ECE and Gauss–Newton curvature are controlled, up to problem-specific constants, by the same margin-dependent exponential tail functional along the trajectory. Causal tests via interventions that target sharpness versus directional curvature confirm the mechanism, with directional interventions yielding more reliable in-sample calibration gains.

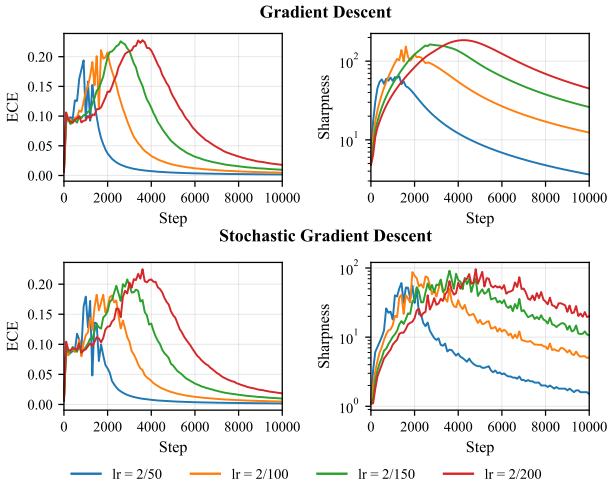


Figure 1. Training dynamics for Gradient Descent and Stochastic Gradient Descent across learning rates on CIFAR-10. **Expected Calibration Error closely tracks sharpness throughout training**: both rise as the model enters the edge of stability regime, peak around the same time, and decay together as training progresses.

1. Introduction

Neural networks are now routinely used in settings where a model's stated uncertainty matters as much as its accuracy, for example, in risk-sensitive domains such as healthcare or autonomous driving. In these contexts, we would like predicted probabilities to reflect empirical correctness: among predictions made with confidence p, approximately a fraction p should be correct. However, modern deep networks are often *miscalibrated*, frequently exhibiting overconfidence even when wrong (Guo et al., 2017).

A widely adopted response to overconfidence is *post-hoc* calibration: models are trained for accuracy and their predicted probabilities are adjusted afterward. While effective in many regimes, this framing treats calibration as a post-training concern, rather than a property that *emerges during training*. Recent work, however, suggests that calibration may be influenced by training dynamics. For example, Sharpness-Aware Minimization (SAM), which biases optimization toward flatter regions of the loss landscape, has been observed to generally overconfidence (Tan et al., 2026). Since sharpness is a geometric property shaped along the optimization trajectory, these results hint at a link between calibration and loss geometry during training. Yet this connection remains poorly understood, motivating the question:

How does calibration evolve throughout optimization, and what aspects of the training govern it?

Importantly, answering this question is non-trivial. Calibration is defined in terms of the model's *predictive confidence distribution*, whereas most training-time analyses characterize optimization through *loss-landscape geometry*, using quantities such as curvature or sharpness to reason about stability and generalization. Bridging these viewpoints is challenging, particularly early in training when predictions

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

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

are still rapidly evolving. While several recent studies have examined relationships between sharpness, flat minima, and calibration at convergence, the resulting picture is mixed: curvature proxies do not reliably predict calibration across architectures, regularization schemes, or optimizers (Mason-Williams et al., 2024). Crucially, these analyses focus on converged solutions. By contrast, we study how calibration and loss geometry co-evolve during training—a perspective that not only clarifies their relationship but also reveals dynamics that can be exploited to improve calibration. After formalizing a training-time connection between calibration and loss geometry, we turn to a prescriptive goal:

Can we intervene on the training procedure to reliably obtain calibrated solutions?

Contributions. We study calibration *during* optimization by jointly tracking calibration metrics, such as Expected Calibration Error (ECE), and curvature-based sharpness proxies, such as Gauss–Newton (GN) sharpness, along the training trajectory (Section 3). We conduct this analysis throughout training rather than only at convergence. We observe across multiple optimization methods that **calibration error and curvature-based measures exhibit a strong and consistent temporal correlation throughout training** (Contribution 1). Next, we probe whether the coupling between calibration and curvature is causal. We compare optimizers designed to minimize sharpness (i.e., favoring flatter minima) with methods that instead suppress steep descent directions along the trajectory. Despite both affecting curvature, we find that **directional interventions yield consistently better in-sample calibration than flat-minima methods** (Contribution 2).

We then provide a unifying explanation through the lens of the *margin* in Section 4. Intuitively, both confidence and curvature are shaped by how strongly the model separates the correct class from its nearest competitor. We formalize this connection by showing mathematically that **a single margin-based functional controls both calibration error and Gauss–Newton sharpness, up to problem-dependent constants** (Contribution 3). This perspective also clarifies an often-observed phenomenon: training and test calibration can diverge even when accuracy improves (Carrell et al., 2022; Wu et al., 2025). Once most examples achieve large positive margins, a relatively small set of near-boundary or negative-margin points can dominate the margin functional, making calibration highly sensitive to how optimization shapes this tail.

Our margin-based view also yields concrete diagnostics for optimizer behavior. In particular, we observe that **Muon induces unusually large training margins, leading to near-zero training ECE, but severe test overconfidence**.

2. Related Work

Calibration. A model is calibrated if predicted confidences match empirical correctness frequencies (Niculescu-Mizil & Caruana, 2005; DeGroot & Fienberg, 1983); we measure miscalibration via the Expected Calibration Error (ECE) (Naeini et al., 2015; Guo et al., 2017), whose formal definition we recall in Appendix B.1. Calibration-improvement methods divide into post-hoc rescaling (Platt, 2000; Guo et al., 2017) and intrinsic training-time objectives, including entropy regularization (Pereyra et al., 2017), label smoothing (Müller et al., 2019), focal loss (Mukhoti et al., 2020), and differentiable surrogates of calibration metrics (Kumar et al., 2018; Bohdal et al., 2023). A separate line links calibration to *local margin geometry*: low-robust-margin points are disproportionately miscalibrated, motivating margin-aware label smoothing (Qin et al., 2021), and many calibration losses can be unified as logit-distance penalties (Liu et al., 2022).

Curvature along the trajectory. A complementary literature treats curvature as a dynamical quantity along training rather than a static attribute of the converged solution. Gradient methods generically approach an *edge-of-stability* (EoS) regime in which the top Hessian eigenvalue saturates near $2/\eta$ before often decreasing later in training (Cohen et al., 2021; 2022; Jastrzębski et al., 2018; Andreyev & Beneventano, 2024); sharpness-aware optimization (SAM) that biases toward flatter minima has separately been observed to reduce overconfidence under cross-entropy (Foret et al., 2021; Zhou et al., 2025; Tan et al., 2026). The sharpness–calibration relationship is, however, fragile across architectures and regularizers (Mason-Williams et al., 2024), pointing to the importance of *how* curvature directions are traversed, not only where optimization converges. We extend this trajectory-level view by jointly tracking calibration and curvature throughout training and contrasting convergence-to-flatness with explicit suppression of unstable high-curvature directions; an extended discussion is provided in Appendix A.

3. The Coupling Between Calibration and Sharpness

We track sharpness and ECE throughout training to study how loss landscape geometry relates to calibration. Following Cohen et al. (2021), we train an MLP (2 hidden layers, 200 units, tanh activation) on CIFAR-10 under cross-entropy (CE) loss. This small-scale setup enables frequent computation of GN sharpness, which serves as a proxy for the top Hessian eigenvalue $\lambda_{\max}$. Models are trained using gradient descent (GD), stochastic gradient descent (SGD), AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019), Muon (Jordan et al., 2024), and SAM (Foret et al., 2021).

We monitor GN sharpness and batch sharpness (Andreyev & Beneventano, 2024) as proxies for loss-landscape geometry, alongside ECE, KCE, loss, and accuracy.

3.1. Calibration Temporally Correlates with Sharpness

Figure 1 shows training dynamics for GD and SGD. Across all CE experiments[1], training ECE and GN sharpness follow the same trajectory: both quantities are small at initialization, increase as training enters an EoS regime, and decrease again later in training. This holds across optimizers and learning rates (see Appendix C.1 for per-optimizer trajectories). Similar observations extend to CIFAR-100 (Appendix C.1). Table 1 quantifies this effect, showing strong Pearson correlations between ECE and (batch) sharpness throughout training; KCE closely matches ECE across all settings, confirming the coupling is not a binning artifact.

Both calibration and sharpness converge to zero once all training points are correctly classified, so their coupling at the end of training is expected. What is surprising is the strong correlation during training, well before convergence, when the model is far from interpolation and calibration is nontrivial. To our knowledge, this has not been observed or explained before. These results suggest that calibration does not depend on training-metrics at convergence, but on the trajectory itself: models that stay in lower-sharpness regions remain better calibrated throughout training, not just asymptotically. We formalize this connection in Section 4.

3.2. Converging to Flat Minima or Following Flat Directions?

The strong temporal correlation between sharpness and calibration established in the previous section raises a causal question: *does calibration improve because optimization converges to flatter minima, or because training dynamics suppress movement along high-curvature directions?* To disentangle these mechanisms, we formulate two competing hypotheses, and empirically find that suppressing directions of steep descent leads to improved in-sample calibration.

Hypothesis 1 (Flat Minima)**.** *Training procedures that bias optimization toward flat minima lead to lower in-sample calibration error.*

Hypothesis 2 (Directional Flatness)**.** *Training procedures that suppress updates along directions of steep curvature lead to lower in-sample calibration error, even if the final solution is not globally flat.*

We test Hypothesis 1 using SAM (Zhou et al., 2025), which explicitly penalizes worst-case loss perturbations within a local neighborhood, and is known to bias optimization toward flatter minima. To test Hypothesis 2, we use optimizers that directly suppress high-curvature directions during training. Muon (Jordan et al., 2024) rescales gradient components to equalize their magnitudes, effectively clamping updates along sharp directions while amplifying flatter ones. BulkSGD (Song et al., 2025) achieves a more extreme intervention by projecting gradients onto the subspace orthogonal to the top Hessian eigenvectors, thereby removing the steepest descent directions entirely. A more detailed analysis of the optimizers and their benefits in this experimental setting can be found in Appendix C.2.

[1]We observe a similar temporal correlation on models trained with mean-squared error (MSE) loss, and defer a detailed discussion to Appendix D.

		GD	SGD	AdamW	Muon	SAM
Train	ECE	$.83_{\pm.08}$	$.84_{\pm.07}$	$.72_{\pm.07}$	$.63_{\pm.26}$	$.92_{\pm.04}$
	KCE	$.83_{\pm.08}$	$.84_{\pm.07}$	$.70_{\pm.08}$	$.61_{\pm.28}$	$.91_{\pm.04}$
Test	ECE	$.96_{\pm.02}$	$.97_{\pm.01}$	$.15_{\pm.15}$	$-.10_{\pm.28}$	$.98_{\pm.01}$
	KCE	$.97_{\pm.01}$	$.97_{\pm.01}$	$.21_{\pm.22}$	$-.16_{\pm.31}$	$.98_{\pm.01}$

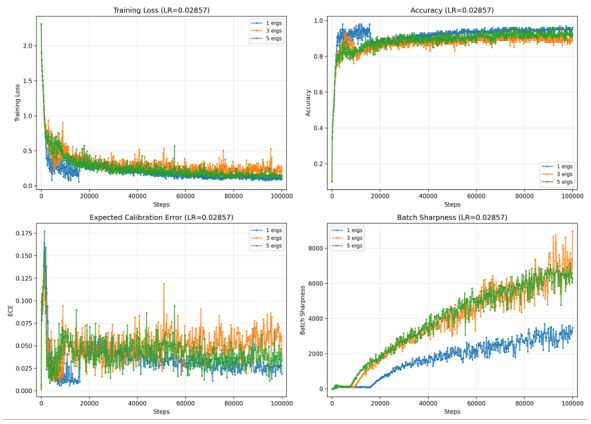
Table 1. Pearson correlation between calibration metrics (ECE and KCE) and GN sharpness, mean $\pm$ std over 4 learning rates.

Figure 2 (and Figure A15 in Appendix C.2) show the training dynamics. Although SAM consistently maintains lower sharpness than GD and SGD, its calibration trajectory closely mirrors that of standard training, with a comparable peak ECE and slower convergence. In contrast, both Muon and BulkSGD achieve substantially lower peak calibration error and faster ECE decay, despite exhibiting markedly different sharpness profiles.

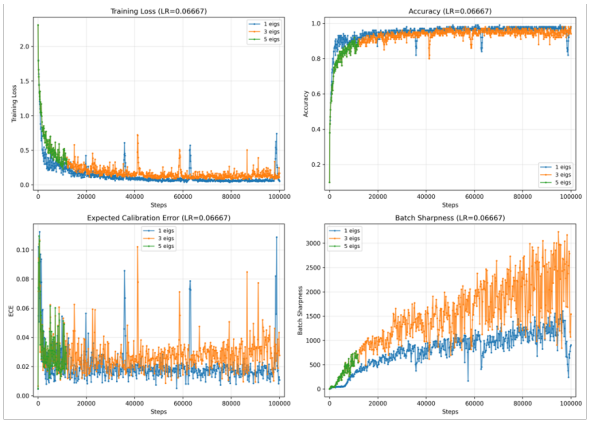
Notably, Muon maintains low calibration error while operating in regimes that are not globally flat, and BulkSGD improves calibration even in the presence of pronounced instability. This suggests that calibration is sensitive to how optimization traverses sharp directions, rather than to the absolute flatness of the loss landscape. However, BulkSGD induces oscillatory dynamics and a sharpness divergence when too many dominant directions are projected out, making it impractical as a standalone optimizer. Together, these results support Hypothesis 2 over Hypothesis 1: suppressing updates along high-curvature directions during training leads to improved in-sample calibration, whereas convergence to flat minima alone does not.

3.3. Out-of-Sample Behavior

The sharpness–calibration coupling is less consistent out of sample (Table 1). Across optimizers, test ECE does not consistently decrease alongside training ECE—in some cases it worsens as training progresses, even after sharpness and training calibration improve (Appendix C.1). Muon is an extreme example: training ECE drops to near zero while test ECE remains high. This reflects the calibration generalization gap in overparameterized models (Carrell et al., 2022; Berta et al., 2025; Wu et al., 2025): a model that fits training data well can become overconfident on



(a) BulkSGD with learning rate $\frac{2}{70}$



(b) BulkSGD with learning rate $\frac{2}{30}$

Figure 2. Training dynamics for BulkSGD across different learning rates and number of projected-out gradients on CIFAR10.

misclassified test examples, causing test ECE to increase and decouple from sharpness. In Muon's case, the near-zero training ECE is consistent with the large training margins it induces (Section 4): once these are extreme, the model becomes overconfident on test examples near the decision boundary, precisely where the margin functional is most sensitive.

Together with the findings from Section 3.2, these results point to an important distinction: on the one hand, directional interventions yield better in-sample calibration than flat-minima methods, suggesting that *how* optimization traverses curvature matters more than where it converges; on the other hand, in-sample improvements do not automatically transfer to test data, pointing to a fundamental train–test gap. In the following section, we formalize this train–test gap and use it to design a training-time intervention that improves out-of-sample calibration.

4. Curvature and Calibration in the Separable and Non-separable Regimes

We explain the alignment between calibration error and curvature observed in Section 3 through a common underlying quantity, the (robust) *true logit margin*: across training, both ECE and Gauss–Newton sharpness respond to the evolution of the same margin-dependent tail functional.

The analysis naturally separates into two regimes. In the *overlap-dominated* regime (early training, and test data whenever accuracy is below 1), a nontrivial fraction of examples have small or negative true margins, so neither ECE nor curvature is forced to be small (Rosenfeld & Risteski, 2024). In the *interpolating* regime, late in training on the training set, all true margins become strictly positive and the coupling becomes two-sided: once the margin tail contracts, both quantities are forced to decrease together.

Together, these results provide a mechanism-level explanation for the observed co-evolution of calibration and curvature during training, and clarify why training and test calibration can diverge even as accuracy improves.

Setup and Notation. Let $(X, Y) \sim \pi$ with $Y \in \{1, \dots, K\}$. A model $\theta \in \mathbb{R}^d$ produces logits $z_\theta(x) \in \mathbb{R}^K$ and probabilities $p_\theta(x) = \mathrm{softmax}(z_\theta(x))$. Let $\hat{y}(x) = \arg\max_k z_\theta(x)_k$ (deterministic tie-break) and confidence $\hat{P}(x) = \max_k p_\theta(x)_k$. Define the *true (logit) margin*

$$m_\theta(x, y) := z_\theta(x)_y - \max_{j \neq y} z_\theta(x)_j,$$

and the *robust true margin* at radius $\varepsilon > 0$,

$$m_{\varepsilon,\theta}(x, y) := \inf_{\|\delta\| \leq \varepsilon} m_\theta(x + \delta, y).$$

Let ECE_M denote the population π/sample $\mathcal{D}$ binned ECE computed by binning $\hat{P}(X)$ into M bins. Let $J_\theta(x) := \partial z_\theta(x)/\partial\theta \in \mathbb{R}^{K \times d}$ and, for cross-entropy, $H_z(p) := \mathrm{diag}(p) - pp^\top$. Define the population Gauss–Newton matrix and its curvature proxy

$$H_{\mathrm{GN}}(\theta; \pi) := \mathbb{E}_\pi\big[J_\theta(X)^\top H_z(p_\theta(X)) J_\theta(X)\big],$$

$$\lambda_{\max} := \lambda_{\max}\big(H_{\mathrm{GN}}(\theta; \pi)\big).$$

4.1. Regime I: overlap-dominated (non-separable) behavior

In this subsection all the quantities (ECE, GN matrix, robust margin, robust margin moment) are considered at a population level. See details in Appendix B. Define the robust exponential margin moment $Q(\theta) := \mathbb{E}_{(X,Y)\sim\pi}\big[e^{-m_{\varepsilon,\theta}(X,Y)}\big]$.

Theorem 4.1 (Overlap regime: robust-margin upper bounds)**.** *For any θ and any distribution π,*

$$\mathrm{ECE}_M \;\leq\; (K - 1)\, Q(\theta).$$

If additionally $\|J_\theta(X)\|_{\mathrm{op}} \le C_J$ *holds* π*-a.s., then*

$$\lambda_{\max} \;\le\; 2C_J^2\,(K-1)\,Q(\theta).$$

Proof is in Appendix B.4.

Interpretation (two bottlenecks). Theorem 4.1 exposes two multiplicative controls:

- a *probability bottleneck* $Q(\theta)$, dominated by the tail of small/negative robust margins,
- a *geometry bottleneck* C_J^2 (how parameter perturbations move logits).

In overlap-dominated regimes, a persistent set of small robust margins can keep $Q(\theta)$ bounded away from 0, so these bounds need not certify vanishing calibration error or curvature even if loss continues to decrease.

4.2. Regime II: Interpolating (separable) behavior on the training set

In this subsection all the quantities (ECE, GN matrix, robust margin, robust margin moment) are considered at a finite-sample level. See details in Appendix B. Let $\gamma(\theta;\mathcal{D}) := \min_{1\le i\le n} m_\theta(x_i,y_i)$ and the empirical exponential margin moment $Q_{\mathcal{D}}(\theta) := \frac{1}{n}\sum_{i=1}^n e^{-m_{\theta_t}(x_i,y_i)}$.

Theorem 4.2 (Interpolating regime: two-sided ECE–margin control and coupling to $\lambda_{\max}$)**.** *Assume* $\gamma(\theta;\mathcal{D}) > 0$ *(all training points correctly classified with strictly positive true margin). Then*

$$\frac{1}{K}\,Q_{\mathcal{D}}(\theta) \;\le\; \mathrm{ECE}_M \;\le\; (K-1)\,Q_{\mathcal{D}}(\theta) \le\; (K-1)\,e^{-\gamma(\theta;\mathcal{D})}.$$

If additionally $\max_{i\in[n]}\|J_\theta(x_i)\|_{\mathrm{op}} \le C_J$*, then*

$$\lambda_{\max} \;\le\; 2C_J^2\,(K-1)\,Q_{\mathcal{D}}(\theta) \;\le\; 2C_J^2\,K(K-1)\,\mathrm{ECE}_M,$$

equivalently $\mathrm{ECE}_M \ge \lambda_{\max}/(2C_J^2K(K-1))$.

Proof is in Appendix B.5.

Interpretation. In the interpolating regime, ECE_M is equivalent up to constants to the exponential margin moment $Q_{\mathcal{D}}(\theta)$, and $\lambda_{\max}$ is controlled by the *same* moment (under bounded Jacobians). This implies that once the training set is correctly classified, GN sharpness cannot be large without in-sample ECE also being large. Moreover, in this regime, empirical binning becomes immaterial: ECE_M reduces to the mean misconfidence (formalized in Appendix B).

Connection to the observed train/test split. Theorems 4.1–4.2 jointly explain the training-time co-evolution of ECE and sharpness and their decoupling on held-out data: the bounds depend on *true* margins, so test ECE can worsen even as predicted margins grow. Under a local label-preserving shift, the empirical robust-margin moment $Q^0_{\varepsilon,\mathcal{D}}(\theta)$ further provides an actionable training-side certificate, $\mathrm{ECE}_M(\theta;P_{\mathrm{test}}) \le (K-1)\,Q^0_{\varepsilon,\mathcal{D}}(\theta)$ (Proposition B.7 and Corollary B.8; see Appendix B).

5. Conclusion

We studied calibration as a *training-time phenomenon* rather than a static property of a converged model, and showed that calibration and sharpness are tightly coupled along the optimization trajectory across multiple optimizers. This coupling arises from a shared dependence on margin growth, explaining both the temporal co-evolution of Expected Calibration Error and curvature during training, as well as the frequent divergence between train and test calibration in overlap-dominated regimes.

Building on this perspective, we distinguished between two competing mechanisms for improving calibration: convergence to flat minima and suppression of updates along high-curvature directions. Empirically, optimizers that implement directional control, such as Muon and BulkSGD, yielded consistently better in-sample calibration than methods targeting flat minima alone, providing causal support for the margin-tail mechanism.

Limitations. Our study relies on explicit curvature diagnostics (Gauss–Newton / Hessian-based measurements), which are computationally expensive and constrain the scale of architectures and datasets we can probe. Our theoretical results identify a margin-tail mediator that upper-bounds both calibration error and curvature under a Jacobian-control assumption; we do not claim that these certificates are tight or that the Jacobian bounds hold uniformly in all deep networks. A natural next step is to develop scalable, distributionally robust proxies for the mediator (for example, low-rank spectral estimators, mini-batch surrogates, or input-space stability measurements) and to characterize when they preserve the qualitative regime predictions we derive.

Our experiments also cover small-scale image classification only; extending the trajectory-level analysis to language-model training, where calibration is a central open question, is a natural follow-up. More broadly, we see the trajectory perspective itself—tracking how calibration, curvature, and margins co-evolve rather than inspecting them only at convergence—as a lens applicable to other training-time phenomena in deep networks.

References

Andreyev, A. and Beneventano, P. Edge of stochastic stability: Revisiting the edge of stability for SGD.

arXiv preprint arXiv:2412.20553, 2024. doi: 10.48550/arXiv.2412.20553. URL https://arxiv.org/abs/2412.20553.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6240–6249, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/b22b257ad0519d4500539da3c8bcf4dd-Abstract.html.

Berta, E., Holzmüller, D., Jordan, M. I., and Bach, F. Rethinking early stopping: Refine, then calibrate. *arXiv preprint arXiv:2501.19195*, 2025. doi: 10.48550/arXiv.2501.19195. URL https://arxiv.org/abs/2501.19195.

Bohdal, O., Yang, Y., and Hospedales, T. Meta-calibration: Learning of model calibration using differentiable expected calibration error. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=R2hUure381. Accepted by TMLR.

Carrell, A. M., Mallinar, N., Lucas, J., and Nakkiran, P. The calibration generalization gap. *arXiv preprint arXiv:2210.01964*, 2022. URL https://arxiv.org/abs/2210.01964.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-SGD: Biasing gradient descent into wide valleys. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1YfAfcgl.

Cohen, J. M., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM. ICLR 2021 Poster.

Cohen, J. M., Ghorbani, B., Krishnan, S., Agarwal, N., Medapati, S., Badura, M., Suo, D., Cardoze, D., Nado, Z., Dahl, G. E., and Gilmer, J. Adaptive gradient methods at the edge of stability, 2022. URL https://arxiv.org/abs/2207.14484.

DeGroot, M. H. and Fienberg, S. E. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983. doi: 10.2307/2987588.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028. PMLR, 2017. URL https://proceedings.mlr.press/v70/dinh17b.html.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM. ICLR 2021 Spotlight.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural Computation*, 9(1):1–42, 1997. doi: 10.1162/neco.1997.9.1.1.

Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better: Closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1731–1741, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/a5e0ff62be0b08456fc7f1e88812af3d-Abstract.html.

Jastrzębski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three Factors Influencing Minima in SGD. *arXiv:1711.04623 [cs, stat]*, September 2018. URL http://arxiv.org/abs/1711.04623. arXiv:1711.04623.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

Jordan, K., Jin, Y., Boza, V., You, J., Cesista, F., Newhouse, L., and Bernstein, J. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Kull, M., Perello-Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., and Flach, P. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems*, volume 32, pp. 12316–12326, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8ca01ea920679a0fe3728441494041b9-Abstract.html.

Kumar, A., Sarawagi, S., and Jain, U. Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2805–2814. PMLR, 2018. URL https://proceedings.mlr.press/v80/kumar18a.html.

Lengyel, D., Jennings, N., Parpas, P., and Kantas, N. On flat minima, large margins and generalizability. OpenReview (ICLR 2021 submission), 2021. URL https://openreview.net/forum?id=Ki5Mv0iY8C.

Li, Y. and Sur, P. Optimal and provable calibration in high-dimensional binary classification: Angular calibration and Platt scaling. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=SgQAleMecy. NeurIPS 2025 Spotlight.

Liang, T., Poggio, T., Rakhlin, A., and Stokes, J. Fisher-Rao metric, geometry, and complexity of neural networks. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 888–896. PMLR, 2019.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Liu, B., Ben Ayed, I., Galdran, A., and Dolz, J. The devil is in the margin: Margin-based label smoothing for network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 80–88, 2022.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Maddox, W. J., Benton, G., and Wilson, A. G. Rethinking parameter counting in deep models: Effective dimensionality revisited. *arXiv preprint arXiv:2003.02139*, 2020. URL https://arxiv.org/abs/2003.02139.

Mason-Williams, I., Ekholm, F., and Huszár, F. Explicit regularisation, sharpness and calibration. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning (SciForDL)*. OpenReview.net, October 2024. URL https://openreview.net/forum?id=ZQTiGcykl6.

Möllenhoff, T. and Khan, M. E. SAM as an optimal relaxation of Bayes. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=k4fevFqSQcX.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P. H. S., and Dokania, P. K. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15288–15299, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/aeb7b30ef1d024a76f21a1d40e30c302-Abstract.html.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32, pp. 4696–4705, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f1748d6b0fd9d439f71450117eba2725-Abstract.html.

Naeini, M. P., Cooper, G. F., and Hauskrecht, M. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015. doi: 10.1609/aaai.v29i1.9602. URL https://ojs.aaai.org/index.php/AAAI/article/view/9602.

Nagarajan, V. and Kolter, J. Z. Deterministic PAC-bayesian generalization bounds for deep networks via generalizing noise-resilience. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hygn2o0qKX.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5947–5956, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/10ce03a1ed01077e3e289f3e53c72813-Abstract.html.

Niculescu-Mizil, A. and Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632, 2005. doi: 10.1145/1102351.1102430.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13991–14002, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., and Hinton, G. Regularizing neural networks by penalizing confident output distributions. In *ICLR 2017 Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HyhbYrGYe.

Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Smola, A. J., Bartlett, P. L., Schölkopf, B., and Schuurmans, D. (eds.), *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 2000.

Qin, Y., Wang, X., Beutel, A., and Chi, E. Improving calibration through the relationship with adversarial robustness. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14358–14369, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/78421a2e0e1168e5cd1b7a8d23773ce6-Abstract.html.

Rosenfeld, E. and Risteski, A. Outliers with opposing signals have an outsized effect on neural network optimization. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kIZ3S3tel6. ICLR 2024 Poster.

Song, M., Ahn, K., and Yun, C. Does SGD really happen in tiny subspaces? In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=v6iLQBoIJw. ICLR 2025 Poster.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70): 1–57, 2018. URL https://jmlr.org/papers/v19/18-188.html.

Stutz, D., Hein, M., and Schiele, B. Confidence-calibrated adversarial training: Generalizing to unseen attacks. In Daumé III, H. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9155–9166. PMLR, 2020. URL https://proceedings.mlr.press/v119/stutz20a.html.

Stutz, D., Hein, M., and Schiele, B. Relating adversarially robust generalization to flat minima. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7807–7817, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/papers/Stutz_Relating_Adversarially_Robust_Generalization_to_Flat_Minima_ICCV_2021_paper.pdf.

Tan, C., Zhou, Y., Ye, H., Dai, G., Liu, J., Song, Z., Zhang, J., Zhao, Z., Hao, Y., and Xu, Y. Towards understanding the calibration benefits of sharpness-aware minimization. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=c0ERcCz6lD. ICLR 2026 Poster.

Thulasidasan, S., Chennupati, G., Bilmes, J. A., Bhattacharya, T., and Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/36ad8b5f42db492827016448975cc22d-Abstract.html.

Tsuzuku, Y., Sato, I., and Sugiyama, M. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using PAC-bayesian analysis. In Daumé III, H. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9636–9647. PMLR, 2020. URL https://proceedings.mlr.press/v119/tsuzuku20a.html.

Wu, J., Bartlett, P., Telgarsky, M., and Yu, B. Benefits of early stopping in gradient descent for overparameterized logistic regression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67081–67110. PMLR, 2025. URL https://proceedings.mlr.press/v267/wu25b.html.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002. doi: 10.1145/775047.775151.

Zhang, J., Kailkhura, B., and Han, T. Y.-J. Mix-n-Match: Ensemble and compositional methods for uncertainty calibration in deep learning. In Daumé III, H. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11117–11128. PMLR, 2020. URL https://proceedings.mlr.press/v119/zhang20k.html.

Zheng, Y., Zhang, R., and Mao, Y. Regularizing neural networks via adversarial model perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8152–8161, 2021. doi: 10.1109/CVPR46437.2021.00806. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zheng_Regularizing_Neural_Networks_via_Adversarial_Model_Perturbation_CVPR_2021_paper.html.

Zhou, Z., Wang, M., Mao, Y., Li, B., and Yan, J. Sharpness-aware minimization efficiently selects flatter minima late in training. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=aD2uwhLbnA. ICLR 2025 Spotlight.

A. Further Related Work

A.1. Mitigating Miscalibration: Extended Discussion

Post-hoc calibration. Post-hoc methods learn a mapping from model scores to probabilities on a held-out set. Classical approaches include Platt scaling and isotonic/binning methods (Platt, 2000; Zadrozny & Elkan, 2002), with temperature scaling the de facto recipe for deep networks (Guo et al., 2017). More expressive calibrators—such as Dirichlet calibration and compositional strategies like Mix-n-Match—correct class- or confidence-dependent distortions while preserving accuracy (Kull et al., 2019; Zhang et al., 2020). Because post-hoc methods do not influence training dynamics, they provide limited mechanistic insight and can degrade under distribution shift (Ovadia et al., 2019).

Intrinsic methods. Intrinsic methods incorporate calibration objectives directly into training. These include entropy-based regularization (Pereyra et al., 2017), label smoothing (Müller et al., 2019), augmentation schemes such as mixup that soften targets (Thulasidasan et al., 2019), and focal loss—originally introduced for class imbalance—which yields better calibrated classifiers even before post-hoc scaling (Lin et al., 2017; Mukhoti et al., 2020). Differentiable surrogates of calibration metrics further enable joint training for accuracy and calibration (Kumar et al., 2018; Bohdal et al., 2023).

High-dimensional perspectives on miscalibration. Recent theoretical work highlights that miscalibration can arise intrinsically from high-dimensional statistical effects, even in well-specified problems. Li & Sur (2025) analyze confidence estimates in high-dimensional classification and show that predictive probabilities can systematically deviate from true correctness likelihoods due to margin concentration and estimation noise, suggesting that miscalibration need not stem from optimization failures alone. Our perspective is complementary: rather than asymptotic statistical limits, we study the finite-sample training-time evolution of margins and curvature.

A.2. Robust Margins, Sharpness, and Calibration

Robust margins and calibration. Cross-entropy training pushes predictions toward extreme softmax outputs. On linearly separable data, gradient descent drives margins to infinity, converging to a max-margin classifier (Soudry et al., 2018). While large margins aid classification, they also amplify overconfidence: Qin et al. (2021) showed that inputs with small robust margin are more likely to be miscalibrated, and proposed adaptive label smoothing on such points. Focal loss (Mukhoti et al., 2020) and label smoothing (Müller et al., 2019) can likewise curb overconfidence on hard examples. Foret et al. (2021)'s SAM, which biases optimization toward flatter minima, has been observed to lower calibration error (Zheng et al., 2021; Möllenhoff & Khan, 2023). These results share a common theme: controlling the growth or fragility of margins tends to improve calibration. Achieving both robustness and calibration is nevertheless non-trivial—standard adversarial training can degrade calibration without targeted interventions (Stutz et al., 2020).

Flat minima and margins. Loss-landscape geometry has long been linked to generalization, with flat minima hypothesized to be preferable to sharp ones (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017). Dinh et al. (2017) complicated this picture: scaling symmetries in deep networks allow arbitrarily sharp solutions with identical outputs, motivating scale-invariant sharpness measures (Tsuzuku et al., 2020; Liang et al., 2019). Under such measures, flatter minima correlate with better generalization in CE-trained models (Jiang et al., 2020; Maddox et al., 2020). A structural correlate is the classification margin: flat basins of the CE loss align with large training margins (Lengyel et al., 2021; Jiang et al., 2020). Small-batch SGD, which implicitly enlarges margins (Hoffer et al., 2017), also finds flatter solutions than large-batch training (Keskar et al., 2017). Adversarially robust models—which have larger input margins—exhibit lower curvature in weight space (Stutz et al., 2021); conversely, weight-space flatness regularizers such as entropy-SGD (Chaudhari et al., 2017) improve adversarial robustness as a side effect (Stutz et al., 2021).

Linear vs. non-linear caveats. In linear models trained with cross-entropy, the notion of "flat vs. sharp" is less meaningful: on separable data the weight norm grows without bound as margins maximize, driving the Hessian to zero while the classifier becomes arbitrarily confident. Meaningful cross-setting flatness comparisons therefore require correcting for reparameterization invariances (Dinh et al., 2017; Neyshabur et al., 2017; Tsuzuku et al., 2020). In deep non-linear networks with such corrections, large margins correspond to flatter minima (Lengyel et al., 2021). This distinction explains why margin-based analyses (Bartlett et al., 2017; Nagarajan & Kolter, 2019) are often preferred for theoretical guarantees in linear settings.

A.3. Positioning Relative to Prior Work

The four literatures above—miscalibration under cross-entropy, sharpness/flatness and optimization stability, margin maximization via implicit bias, and robustness–calibration connections—together with recent work on calibration benefits of sharpness-aware optimizers (Tan et al., 2026), provide the backdrop for our contribution.

Our work extends these literatures in three directions:

- **Trajectory-level analysis.** Most prior work compares final solutions; we track calibration and curvature *pathwise* across training and show that they co-evolve, peaking together near the edge of stability and decaying together.
- **A shared margin-tail mediator.** We prove that a single exponential margin moment and its robust variant simultaneously upper-bound ECE and Gauss–Newton sharpness, with a two-sided sandwich for ECE in the interpolating regime. The triangulation ECE $\leftrightarrow$ GN sharpness $\leftrightarrow$ robust margins is, to our knowledge, new.
- **Directional vs. flat-minima interventions.** We distinguish optimizers that bias toward flat minima (SAM) from those that suppress steep descent directions along the trajectory (Muon, BulkSGD), and show empirically that the latter yield more reliable in-sample calibration gains.

B. Proofs for Section 4

B.1. Definitions and basic reductions

Fixed binning. Fix $M \in \mathbb{N}$ and deterministic bin edges $0 = a_0 < a_1 < \cdots < a_M = 1$. Define bins $I_m := (a_{m-1}, a_m]$ for $m = 1, \ldots, M$.

Population ECE with fixed bins. Let $(X, Y) \sim \pi$ with $Y \in \{1, \ldots, K\}$. For fixed θ, define the predicted label

$$\widehat{Y} := \arg\max_k z_\theta(X)_k$$

(using the deterministic tie-break rule from the main text), and the confidence

$$\widehat{P} := \max_k p_\theta(X)_k = p_\theta(X)_{\widehat{Y}}.$$

Let $B_m := \{\widehat{P} \in I_m\}$. Define binwise accuracy and confidence by

$$\mathrm{acc}(B_m) := \begin{cases} \mathbb{P}(\widehat{Y} = Y \mid B_m), & \mathbb{P}(B_m) > 0, \\ 0, & \mathbb{P}(B_m) = 0, \end{cases} \qquad \mathrm{conf}(B_m) := \begin{cases} \mathbb{E}[\widehat{P} \mid B_m], & \mathbb{P}(B_m) > 0, \\ 0, & \mathbb{P}(B_m) = 0. \end{cases}$$

The population binned calibration error is

$$\mathrm{ECE}_M(\theta; \pi) := \sum_{m=1}^{M} \mathbb{P}(B_m)\, |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|.$$

Equivalently, with $Z := \mathbf{1}\{\widehat{Y} = Y\} - \widehat{P}$,

$$\mathrm{ECE}_M(\theta; \pi) = \sum_{m=1}^{M} \mathbb{P}(B_m)\, \big|\mathbb{E}[Z \mid B_m]\big|.$$

Empirical ECE. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, define

$$\widehat{Y}_i := \arg\max_k z_\theta(x_i)_k \quad \text{(same deterministic tie-break)}, \qquad \widehat{P}_i := \max_k p_\theta(x_i)_k,$$

and bins $B_m := \{i : \widehat{P}_i \in I_m\}$. Let $Z_i := \mathbf{1}\{\widehat{Y}_i = y_i\} - \widehat{P}_i$. Then

$$\mathrm{ECE}_M(\theta; \mathcal{D}) := \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \frac{1}{|B_m|} \sum_{i \in B_m} Z_i \right|,$$

with the convention that the inner average is 0 when $|B_m| = 0$.

B.2. Core lemmas

Lemma B.1 (ECE is bounded by the mean absolute correctness–confidence gap)**.** ***(Population).*** *Let $Z := \mathbf{1}\{\widehat{Y} = Y\} - \widehat{P}$ and let $\mathcal{G} := \sigma(B_1, \dots, B_M)$ be the σ-algebra generated by the bin events $B_m := \{\widehat{P} \in I_m\}$. Note that $Z \in [-1, 1]$, hence Z is integrable. Then*

$$\mathrm{ECE}_M(\theta; \pi) = \mathbb{E}\Big[\,\big|\,\mathbb{E}[Z \mid \mathcal{G}]\,\big|\,\Big] \;\le\; \mathbb{E}[|Z|].$$

(Empirical). *For any dataset $\mathcal{D}$,*

$$\mathrm{ECE}_M(\theta; \mathcal{D}) \;\le\; \frac{1}{n}\sum_{i=1}^{n} \big|\mathbf{1}\{\widehat{Y}_i = y_i\} - \widehat{P}_i\big|.$$

Proof. **Population.** For each bin B_m with $\mathbb{P}(B_m) > 0$,

$$\mathbb{E}[Z \mid B_m] = \mathbb{P}(\widehat{Y} = Y \mid B_m) - \mathbb{E}[\widehat{P} \mid B_m] = \mathrm{acc}(B_m) - \mathrm{conf}(B_m),$$

and we set $\mathbb{E}[Z \mid B_m] := 0$ when $\mathbb{P}(B_m) = 0$. Therefore,

$$\mathrm{ECE}_M(\theta; \pi) = \sum_{m=1}^{M} \mathbb{P}(B_m)\,\big|\mathbb{E}[Z \mid B_m]\big| = \mathbb{E}\Big[\,\big|\,\mathbb{E}[Z \mid \mathcal{G}]\,\big|\,\Big].$$

By Jensen's inequality for the convex function $u \mapsto |u|$,

$$\mathbb{E}\Big[\,\big|\,\mathbb{E}[Z \mid \mathcal{G}]\,\big|\,\Big] \le \mathbb{E}\big[\mathbb{E}[|Z| \mid \mathcal{G}]\big] = \mathbb{E}[|Z|].$$

Empirical. For each bin B_m with $|B_m| > 0$, let $Z_i := \mathbf{1}\{\widehat{Y}_i = y_i\} - \widehat{P}_i$. Then

$$\big|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\big| = \left|\frac{1}{|B_m|}\sum_{i \in B_m} Z_i\right| \le \frac{1}{|B_m|}\sum_{i \in B_m} |Z_i|$$

by the triangle inequality. Multiplying by $|B_m|/n$ and summing over m yields the claim. □

Lemma B.2 (Correctness–confidence gap is controlled by the true-class probability)**.** *For any (x, y) and θ, where $\widehat{y}(x)$ and $\widehat{P}(x)$ are defined as above,*

$$\big|\mathbf{1}\{\widehat{y}(x) = y\} - \widehat{P}(x)\big| \;\le\; 1 - p_\theta(x)_y.$$

Proof. Let $\widehat{y} = \widehat{y}(x)$ and $\widehat{P} = \widehat{P}(x) = \max_k p_\theta(x)_k = p_\theta(x)_{\widehat{y}}$. If $\widehat{y} = y$, then $|\mathbf{1}\{\widehat{y} = y\} - \widehat{P}| = |1 - p_y| = 1 - p_y$. If $\widehat{y} \neq y$, then $|\mathbf{1}\{\widehat{y} = y\} - \widehat{P}| = \widehat{P} = p_{\widehat{y}} \le \sum_{j \neq y} p_j = 1 - p_y$, since $p_{\widehat{y}}$ is one of the nonnegative summands in $\sum_{j \neq y} p_j$. □

Lemma B.3 (Softmax tail bound: $1 - p_y$ is exponentially controlled by the true margin)**.** *Let $p = \mathrm{softmax}(z) \in \Delta^{K-1}$ and fix a label y. Define the true margin $m := z_y - \max_{j \neq y} z_j$. Then*

$$1 - p_y \;\le\; \sum_{j \neq y} e^{z_j - z_y} \;\le\; (K-1)e^{-m}. \tag{1}$$

Moreover,

$$\frac{e^{-m}}{1 + (K-1)e^{-m}} \;\le\; 1 - p_y. \tag{2}$$

In particular, if $m \ge 0$ then

$$\frac{1}{K}e^{-m} \;\le\; 1 - p_y \;\le\; (K-1)e^{-m}. \tag{3}$$

Proof. Write

$$p_y = \frac{e^{z_y}}{\sum_{k=1}^{K} e^{z_k}} = \frac{1}{1+\sum_{j\neq y} e^{z_j - z_y}}, \qquad 1 - p_y = \frac{\sum_{j\neq y} e^{z_j - z_y}}{1+\sum_{j\neq y} e^{z_j - z_y}}.$$

Let $S := \sum_{j\neq y} e^{z_j - z_y} \geq 0$. Then $1 - p_y = S/(1+S) \leq S$, proving the first inequality in (1). For each $j \neq y$, $z_j - z_y \leq \max_{k\neq y} z_k - z_y = -m$, hence $e^{z_j - z_y} \leq e^{-m}$ and $S \leq (K-1)e^{-m}$, proving the second inequality in (1). For (2), pick $j^\star \in \arg\max_{j\neq y} z_j$ so that $z_{j^\star} - z_y = -m$ and hence $S \geq e^{-m}$. Therefore

$$1 - p_y = \frac{S}{1+S} \geq \frac{e^{-m}}{1+S} \geq \frac{e^{-m}}{1+(K-1)e^{-m}},$$

where the last step uses $S \leq (K-1)e^{-m}$. If $m \geq 0$ then $e^{-m} \leq 1$ and thus $1 + (K-1)e^{-m} \leq 1 + (K-1) = K$, giving (3). □

Lemma B.4 (Cross-entropy logit Hessian top eigenvalue is controlled by $1 - p_{\max}$)**.** *Let $p \in \Delta^{K-1}$ and define $H_z(p) := \mathrm{diag}(p) - pp^\top$. Then*

$$\lambda_{\max}\big(H_z(p)\big) \;\leq\; 2\big(1 - p_{\max}\big), \qquad p_{\max} := \max_k p_k.$$

Proof. We use Gershgorin's circle theorem for symmetric matrices. Write $A := H_z(p)$, so that for each i,

$$A_{ii} = p_i(1-p_i), \qquad A_{ij} = -p_i p_j \quad (i \neq j).$$

Let $R_i := \sum_{j\neq i} |A_{ij}| = \sum_{j\neq i} p_i p_j = p_i(1-p_i)$. Gershgorin implies every eigenvalue λ of A lies in at least one interval

$$\lambda \in [A_{ii} - R_i,\, A_{ii} + R_i] = [0,\, 2p_i(1-p_i)] \quad \text{for some } i.$$

Hence

$$\lambda_{\max}(A) \leq \max_i 2p_i(1-p_i).$$

Now fix $k^\star \in \arg\max_k p_k$ so that $p_{k^\star} = p_{\max}$. If $i = k^\star$, then $p_i(1-p_i) = p_{\max}(1-p_{\max}) \leq 1 - p_{\max}$. If $i \neq k^\star$, then $p_i \leq 1 - p_{\max}$ and $p_i(1-p_i) \leq p_i \leq 1 - p_{\max}$. Therefore $\max_i p_i(1-p_i) \leq 1 - p_{\max}$ and consequently

$$\lambda_{\max}(H_z(p)) \leq 2(1 - p_{\max}),$$

as claimed. □

Lemma B.5 (Robust margin comparisons (trivial upper bound; Lipschitz lower bound))**.** *For all (x, y),*

$$m_{\varepsilon,\theta}(x,y) \leq m_\theta(x,y) \qquad \Longrightarrow \qquad e^{-m_\theta(x,y)} \leq e^{-m_{\varepsilon,\theta}(x,y)}. \tag{4}$$

If moreover there exists $L_m(x,y) \in [0,\infty)$ such that

$$|m_\theta(x+\delta, y) - m_\theta(x,y)| \leq L_m(x,y)\,\|\delta\| \qquad \forall\, \|\delta\| \leq \varepsilon,$$

then

$$m_{\varepsilon,\theta}(x,y) \geq m_\theta(x,y) - \varepsilon L_m(x,y) \qquad \Longrightarrow \qquad e^{-m_\theta(x,y)} \geq e^{-\varepsilon L_m(x,y)}\, e^{-m_{\varepsilon,\theta}(x,y)}. \tag{5}$$

Proof. **Trivial robust-vs-clean comparison.** By definition of the infimum and because $\delta = 0$ is feasible, we have

$$m_{\varepsilon,\theta}(x,y) = \inf_{\|\delta\|\leq\varepsilon} m_\theta(x+\delta, y) \leq m_\theta(x,y).$$

Since the map $t \mapsto e^{-t}$ is *decreasing*, this implies

$$e^{-m_\theta(x,y)} \leq e^{-m_{\varepsilon,\theta}(x,y)},$$

which is (4).

Lipschitz lower bound. Assume the stated local Lipschitz condition at (x, y). Then for any $\|\delta\| \le \varepsilon$,

$$m_\theta(x+\delta, y) \ge m_\theta(x, y) - L_m(x, y)\,\|\delta\| \ge m_\theta(x, y) - \varepsilon L_m(x, y).$$

Taking the infimum over all $\|\delta\| \le \varepsilon$ yields

$$m_{\varepsilon,\theta}(x, y) \ge m_\theta(x, y) - \varepsilon L_m(x, y),$$

equivalently

$$m_\theta(x, y) \le m_{\varepsilon,\theta}(x, y) + \varepsilon L_m(x, y).$$

Multiply by -1 (which flips the inequality) to get

$$-m_\theta(x, y) \ge -m_{\varepsilon,\theta}(x, y) - \varepsilon L_m(x, y),$$

and exponentiate to obtain

$$e^{-m_\theta(x,y)} \ge e^{-m_{\varepsilon,\theta}(x,y)}\, e^{-\varepsilon L_m(x,y)},$$

which is (5). □

Remark B.6 (Label-free GN bound via predicted margin). Because $H_z(p_\theta(X))$ depends only on X, one can avoid the label Y in the GN bound. Let $\widehat{y}(x) \in \arg\max_k z_\theta(x)_k$ (with the deterministic tie-break rule) and define the *predicted margin*

$$\widehat{m}_\theta(x) := z_\theta(x)_{\widehat{y}(x)} - \max_{j \ne \widehat{y}(x)} z_\theta(x)_j \;\ge 0.$$

Applying Lemma B.3 with $y = \widehat{y}(x)$ yields

$$1 - p_{\max}(x) = 1 - p_\theta(x)_{\widehat{y}(x)} \le (K-1)e^{-\widehat{m}_\theta(x)}.$$

Combining with Lemma B.4 gives

$$\lambda_{\max}\big(H_z(p_\theta(x))\big) \le 2\big(1 - p_{\max}(x)\big) \le 2(K-1)e^{-\widehat{m}_\theta(x)}.$$

cConsequently, under $\|J_\theta(X)\|_{\mathrm{op}} \le C_J$ π-a.s.,

$$\lambda_{\max}\big(H_{\mathrm{GN}}(\theta; \pi)\big) \le 2C_J^2(K-1)\,\mathbb{E}\big[e^{-\widehat{m}_\theta(X)}\big].$$

This can be substantially tighter than bounds routing through Y when the model is confidently incorrect.

B.3. Rigorous restatement of the main theorems

Notation alignment with Section 4. Fix a robust radius $\varepsilon > 0$. To match the main-text notation, we use

$$Q(\theta; \pi) \;:=\; \mathbb{E}_{(X,Y)\sim\pi}\big[e^{-m_{\varepsilon,\theta}(X,Y)}\big] \qquad \text{and} \qquad Q_{\mathcal{D}}(\theta) \;:=\; \frac{1}{n}\sum_{i=1}^{n} e^{-m_\theta(x_i,y_i)}.$$

When π (or $\mathcal{D}$) is clear from context, we may drop it from the notation. For comparison with alternative functionals used in some intermediate lemmas, note that $Q(\theta; \pi)$ coincides with the quantity previously denoted $\Psi^0_\varepsilon(\theta; \pi)$, and $Q_{\mathcal{D}}(\theta)$ coincides with the quantity previously denoted $\mu(\theta; \mathcal{D})$. If a pointwise margin Lipschitz constant $L_m(\cdot, \cdot)$ is available, we also define the (generally looser) population functional

$$Q^+(\theta; \pi) \;:=\; \mathbb{E}_{(X,Y)\sim\pi}\big[e^{\varepsilon L_m(X,Y)}\, e^{-m_{\varepsilon,\theta}(X,Y)}\big],$$

and the finite-sample robust moments

$$Q^0_{\varepsilon,\mathcal{D}}(\theta) \;:=\; \frac{1}{n}\sum_{i=1}^{n} e^{-m_{\varepsilon,\theta}(x_i,y_i)}, \qquad Q^-_{\varepsilon,\mathcal{D}}(\theta) \;:=\; \frac{1}{n}\sum_{i=1}^{n} e^{-\varepsilon L_m(x_i,y_i)}\, e^{-m_{\varepsilon,\theta}(x_i,y_i)}.$$

The proofs appear in Subsections B.4 and B.5.

Theorem 4.1 *(Overlap regime: simultaneous robust-margin upper bounds).*
Let π be any distribution on $\mathcal{X} \times \{1, \dots, K\}$ and let θ be any parameter vector.

(i) Calibration upper bound. For the population binned calibration error $\mathrm{ECE}_M(\theta;\pi)$,

$$\mathrm{ECE}_M(\theta;\pi) \;\le\; (K-1)\,\mathbb{E}\big[e^{-m_\theta(X,Y)}\big] \;\le\; (K-1)\,Q(\theta;\pi).$$

If L_m is defined, then also $\mathrm{ECE}_M(\theta;\pi) \le (K-1)\,Q^+(\theta;\pi)$, but this is never tighter than the $Q(\theta;\pi)$ bound since $Q^+(\theta;\pi) \ge Q(\theta;\pi)$.

(ii) Gauss–Newton curvature (top eigenvalue) upper bound. Assume additionally that the logit Jacobian is uniformly bounded in operator norm,

$$\|J_\theta(X)\|_{\mathrm{op}} \le C_J \qquad \pi\text{-a.s.}$$

Then the population Gauss–Newton matrix

$$H_{\mathrm{GN}}(\theta;\pi) := \mathbb{E}_{(X,Y)\sim\pi}\big[J_\theta(X)^\top H_z(p_\theta(X))J_\theta(X)\big]$$

satisfies

$$\lambda_{\max}\big(H_{\mathrm{GN}}(\theta;\pi)\big) \;\le\; 2C_J^2\,(K-1)\,\mathbb{E}\big[e^{-m_\theta(X,Y)}\big] \;\le\; 2C_J^2\,(K-1)\,Q(\theta;\pi).$$

If L_m is defined, then also $\lambda_{\max}(H_{\mathrm{GN}}(\theta;\pi)) \le 2C_J^2\,(K-1)\,Q^+(\theta;\pi)$, again a looser bound than the one via $Q(\theta;\pi)$.

(iii) What the bound can (and cannot) certify. If along a training trajectory $\{\theta_t\}$ the robust moment $Q(\theta_t;\pi)$ fails to converge to 0, then the bounds in (i)–(ii) do not certify that $\mathrm{ECE}_M(\theta_t;\pi) \to 0$ or $\lambda_{\max}(H_{\mathrm{GN}}(\theta_t;\pi)) \to 0$.

(iv) Remarks (label dependence and trivial clamping).

- **Label dependence vs. label-free curvature.** $\lambda_{\max}(H_{\mathrm{GN}}(\theta;\pi))$ depends only on the marginal law of X (since $H_z(p_\theta(X))$ is label-free), whereas the bound above routes through Y via $m_\theta(X,Y)$ (equivalently $1-p_\theta(X)_Y$). This is valid but can be loose, especially when the model is confidently wrong. A label-free alternative (via the predicted margin) is given in Remark B.6.
- **Bounds may exceed 1.** Since $\mathrm{ECE}_M(\theta;\pi) \in [0,1]$, any upper bound U can be trivially tightened to $\min\{1,U\}$.

Theorem 4.2 *(Interpolating regime: two-sided ECE control and coupling to $\lambda_{\max}$)***.**
Assume $\gamma(\theta;\mathcal{D}) > 0$, i.e. the training set $\mathcal{D} = \{(x_i,y_i)\}_{i=1}^n$ is correctly classified with strictly positive *true* margin. (Strictness ensures $\widehat{Y}_i = y_i$ without tie-breaking subtleties.)

(i) Two-sided control of in-sample ECE by the exponential margin moment.

$$\frac{1}{K}\,Q_{\mathcal{D}}(\theta) \;\le\; \mathrm{ECE}_M(\theta;\mathcal{D}) \;\le\; (K-1)\,Q_{\mathcal{D}}(\theta) \;\le\; (K-1)\,e^{-\gamma(\theta;\mathcal{D})}.$$

(As always, one may clamp the upper bound by $\min\{1,\cdot\}$.)

(ii) In-sample GN curvature bound in terms of the same moment. Assume additionally that $\|J_\theta(x_i)\|_{\mathrm{op}} \le C_J$ for all $i = 1,\dots,n$. Then

$$\lambda_{\max}\big(H_{\mathrm{GN}}(\theta;\mathcal{D})\big) \;\le\; 2C_J^2\,(K-1)\,Q_{\mathcal{D}}(\theta) \;\le\; 2C_J^2\,K(K-1)\,\mathrm{ECE}_M(\theta;\mathcal{D}).$$

(iii) Consequence: in the interpolating regime, curvature and ECE are forced to co-vary. Under the same assumptions,

$$\mathrm{ECE}_M(\theta;\mathcal{D}) \;\ge\; \frac{\lambda_{\max}(H_{\mathrm{GN}}(\theta;\mathcal{D}))}{2C_J^2\,K(K-1)}.$$

Thus, once the training set is correctly classified and Jacobians remain bounded, large GN curvature cannot occur without large in-sample ECE.

(iv) Robust-margin variant (optional; requires local Lipschitzness at (x_i, y_i)). Assume moreover that for each i there exists $L_m(x_i, y_i) \in [0, \infty)$ such that

$$|m_\theta(x_i + \delta, y_i) - m_\theta(x_i, y_i)| \le L_m(x_i, y_i)\,\|\delta\| \qquad \forall\, \|\delta\| \le \varepsilon.$$

Then, with the robust moments $Q^0_{\varepsilon,\mathcal{D}}(\theta)$ and $Q^-_{\varepsilon,\mathcal{D}}(\theta)$ defined above,

$$\frac{1}{K}\, Q^-_{\varepsilon,\mathcal{D}}(\theta) \;\le\; \mathrm{ECE}_M(\theta; \mathcal{D}) \;\le\; (K-1)\, Q^0_{\varepsilon,\mathcal{D}}(\theta),$$

and, under $\max_i \|J_\theta(x_i)\|_{\mathrm{op}} \le C_J$,

$$\lambda_{\max}\big(H_{\mathrm{GN}}(\theta; \mathcal{D})\big) \;\le\; 2C_J^2\,(K-1)\, Q^0_{\varepsilon,\mathcal{D}}(\theta).$$

(v) Remark (binning irrelevance under perfect accuracy). Under $\gamma(\theta; \mathcal{D}) > 0$, every nonempty bin has empirical accuracy 1, so $\mathrm{ECE}_M(\theta; \mathcal{D})$ reduces to the *mean misconfidence* and becomes independent of the choice of bins.

B.4. Proof of Theorem 4.1

Proof of Theorem 4.1. **(i) Calibration bound.** By Lemma B.1 (population version),

$$\mathrm{ECE}_M(\theta; \pi) \le \mathbb{E}_{(X,Y)\sim\pi}\Big[\big|\mathbf{1}\{\widehat{Y} = Y\} - \widehat{P}\big|\Big].$$

By Lemma B.2,

$$\big|\mathbf{1}\{\widehat{Y} = Y\} - \widehat{P}\big| \le 1 - p_\theta(X)_Y,$$

hence

$$\mathrm{ECE}_M(\theta; \pi) \le \mathbb{E}_{(X,Y)\sim\pi}\big[1 - p_\theta(X)_Y\big].$$

Applying Lemma B.3 with $z = z_\theta(X)$ and $y = Y$ yields

$$1 - p_\theta(X)_Y \le (K-1)e^{-m_\theta(X,Y)} \qquad \pi\text{-a.s.},$$

so

$$\mathrm{ECE}_M(\theta; \pi) \le (K-1)\,\mathbb{E}_{(X,Y)\sim\pi}\big[e^{-m_\theta(X,Y)}\big].$$

Finally, by Lemma B.5 (the trivial robust-vs-clean comparison),

$$e^{-m_\theta(X,Y)} \le e^{-m_{\varepsilon,\theta}(X,Y)} \qquad \pi\text{-a.s.},$$

and therefore

$$\mathrm{ECE}_M(\theta; \pi) \le (K-1)\,\mathbb{E}_{(X,Y)\sim\pi}\big[e^{-m_{\varepsilon,\theta}(X,Y)}\big] = (K-1)\,Q(\theta; \pi).$$

If the local Lipschitz condition in Lemma B.5 holds so that $Q^+(\theta; \pi)$ is defined, then $Q(\theta; \pi) \le Q^+(\theta; \pi)$ since $e^{\varepsilon L_m(X,Y)} \ge 1$.

(ii) GN curvature bound. Define the random PSD matrix

$$A(X) := J_\theta(X)^\top H_z(p_\theta(X)) J_\theta(X) \succeq 0.$$

Then $H_{\mathrm{GN}}(\theta; \pi) = \mathbb{E}_{(X,Y)\sim\pi}[A(X)]$. Since $\lambda_{\max}$ is convex on the PSD cone (equivalently, by the variational characterization),

$$\lambda_{\max}\big(H_{\mathrm{GN}}(\theta; \pi)\big) = \lambda_{\max}\big(\mathbb{E}[A(X)]\big) \le \mathbb{E}\big[\lambda_{\max}(A(X))\big].$$

For each realization X,

$$\lambda_{\max}(A(X)) \le \|J_\theta(X)\|_{\mathrm{op}}^2\, \lambda_{\max}\big(H_z(p_\theta(X))\big).$$

Under $\|J_\theta(X)\|_{\mathrm{op}} \le C_J$ π-a.s.,

$$\lambda_{\max}(A(X)) \le C_J^2\, \lambda_{\max}\big(H_z(p_\theta(X))\big).$$

By Lemma B.4,

$$\lambda_{\max}(H_z(p_\theta(X))) \le 2(1 - p_{\max}(X)), \qquad p_{\max}(X) := \max_k p_\theta(X)_k.$$

Since $p_{\max}(X) \ge p_\theta(X)_Y$, we have $1 - p_{\max}(X) \le 1 - p_\theta(X)_Y$, hence

$$\lambda_{\max}(H_z(p_\theta(X))) \le 2\big(1 - p_\theta(X)_Y\big) \le 2(K-1)e^{-m_\theta(X,Y)} \qquad \pi\text{-a.s.}$$

Therefore,

$$\lambda_{\max}\big(H_{\mathrm{GN}}(\theta;\pi)\big) \le 2C_J^2(K-1)\,\mathbb{E}_{(X,Y)\sim\pi}\big[e^{-m_\theta(X,Y)}\big] \le 2C_J^2(K-1)\,\mathbb{E}_{(X,Y)\sim\pi}\big[e^{-m_{\varepsilon,\theta}(X,Y)}\big] = 2C_J^2(K-1)\,Q(\theta;\pi),$$

where the last inequality uses Lemma B.5. If the local Lipschitz condition in Lemma B.5 holds, then also $Q(\theta;\pi) \le Q^+(\theta;\pi)$.

(iii) Certification statement. Immediate from (i)–(ii): if $Q(\theta_t;\pi) \not\to 0$ (or likewise $Q^+(\theta_t;\pi) \not\to 0$), then the corresponding right-hand sides do not converge to 0 and therefore cannot certify $\mathrm{ECE}_M(\theta_t;\pi) \to 0$ nor $\lambda_{\max}(H_{\mathrm{GN}}(\theta_t;\pi)) \to 0$. □

B.5. Proof of Theorem 4.2

Proof of Theorem 4.2. Assume $\gamma(\theta;\mathcal{D}) > 0$, i.e. $m_\theta(x_i, y_i) > 0$ for all i. Hence $\widehat{Y}_i = y_i$ for all i (no tie-breaking occurs).

(i) Two-sided ECE–moment bounds. Because $\widehat{Y}_i = y_i$, every nonempty bin B_m has empirical accuracy $\mathrm{acc}(B_m) = 1$. Therefore, for each nonempty bin,

$$\big|1 - \mathrm{conf}(B_m)\big| = 1 - \mathrm{conf}(B_m) \qquad \text{since } \mathrm{conf}(B_m) \in [0,1].$$

Hence

$$\mathrm{ECE}_M(\theta;\mathcal{D}) = \sum_{m=1}^{M} \frac{|B_m|}{n}\Big(1 - \mathrm{conf}(B_m)\Big) = 1 - \frac{1}{n}\sum_{i=1}^{n} \widehat{P}_i.$$

Since $\widehat{P}_i = \max_k p_\theta(x_i)_k = p_\theta(x_i)_{\widehat{Y}_i}$ and $\widehat{Y}_i = y_i$, we have $\widehat{P}_i = p_\theta(x_i)_{y_i}$, and thus

$$\mathrm{ECE}_M(\theta;\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}\big(1 - p_\theta(x_i)_{y_i}\big). \tag{6}$$

Recalling $Q_{\mathcal{D}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} e^{-m_\theta(x_i,y_i)}$, Lemma B.3 implies (since $m_\theta(x_i,y_i) \ge 0$ for all i) that

$$\frac{1}{K}\, e^{-m_\theta(x_i,y_i)} \le 1 - p_\theta(x_i)_{y_i} \le (K-1)\, e^{-m_\theta(x_i,y_i)}.$$

Averaging over i and using (6) yields

$$\frac{1}{K}\, Q_{\mathcal{D}}(\theta) \le \mathrm{ECE}_M(\theta;\mathcal{D}) \le (K-1)\, Q_{\mathcal{D}}(\theta).$$

Finally, $m_\theta(x_i,y_i) \ge \gamma(\theta;\mathcal{D})$ implies $Q_{\mathcal{D}}(\theta) \le e^{-\gamma(\theta;\mathcal{D})}$.

(ii) GN curvature bound. For each i, define $H_{z,i} := H_z(p_\theta(x_i))$ and $J_i := J_\theta(x_i)$. Then

$$H_{\mathrm{GN}}(\theta;\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} J_i^\top H_{z,i} J_i \succeq 0.$$

Since $\lambda_{\max}$ is convex on the PSD cone,

$$\lambda_{\max}\big(H_{\mathrm{GN}}(\theta;\mathcal{D})\big) \le \frac{1}{n}\sum_{i=1}^{n} \lambda_{\max}(J_i^\top H_{z,i} J_i) \le \frac{1}{n}\sum_{i=1}^{n} \|J_i\|_{\mathrm{op}}^2\, \lambda_{\max}(H_{z,i}).$$

Under $\|J_i\|_{\mathrm{op}} \le C_J$ for all i,

$$\lambda_{\max}\big(H_{\mathrm{GN}}(\theta;\mathcal{D})\big) \le C_J^2 \cdot \frac{1}{n}\sum_{i=1}^{n} \lambda_{\max}(H_{z,i}).$$

By Lemma B.4, $\lambda_{\max}(H_{z,i}) \le 2(1 - p_{\max,i})$. Since $\widehat{Y}_i = y_i$, we have $p_{\max,i} = p_\theta(x_i)_{y_i}$, hence by Lemma B.3,

$$\lambda_{\max}(H_{z,i}) \le 2\big(1 - p_\theta(x_i)_{y_i}\big) \le 2(K-1)e^{-m_\theta(x_i,y_i)}.$$

Therefore

$$\lambda_{\max}\big(H_{\mathrm{GN}}(\theta;\mathcal{D})\big) \le 2C_J^2(K-1)\cdot\frac{1}{n}\sum_{i=1}^{n} e^{-m_\theta(x_i,y_i)} = 2C_J^2(K-1)\,Q_{\mathcal{D}}(\theta).$$

(iii) Coupling to ECE_M (rearranged lower bound). Combining the bound in (ii) with $\mathrm{ECE}_M(\theta;\mathcal{D}) \ge \frac{1}{K}Q_{\mathcal{D}}(\theta)$ from (i) yields

$$\lambda_{\max}\big(H_{\mathrm{GN}}(\theta;\mathcal{D})\big) \le 2C_J^2K(K-1)\,\mathrm{ECE}_M(\theta;\mathcal{D}), \qquad \text{equivalently} \qquad \mathrm{ECE}_M(\theta;\mathcal{D}) \ge \frac{\lambda_{\max}(H_{\mathrm{GN}}(\theta;\mathcal{D}))}{2C_J^2K(K-1)}.$$

(iv) Robust-margin variant. Assume that for each (x_i, y_i) there exists $L_m(x_i, y_i) \in [0, \infty)$ such that

$$|m_\theta(x_i+\delta, y_i) - m_\theta(x_i, y_i)| \le L_m(x_i, y_i)\|\delta\| \qquad \forall\|\delta\| \le \varepsilon.$$

Define the robust moments (as in Appendix E.3)

$$Q^0_{\varepsilon,\mathcal{D}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} e^{-m_{\varepsilon,\theta}(x_i,y_i)}, \qquad Q^-_{\varepsilon,\mathcal{D}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} e^{-\varepsilon L_m(x_i,y_i)}e^{-m_{\varepsilon,\theta}(x_i,y_i)}.$$

By Lemma B.5,

$$e^{-m_\theta(x_i,y_i)} \ge e^{-\varepsilon L_m(x_i,y_i)}e^{-m_{\varepsilon,\theta}(x_i,y_i)} \qquad \text{and} \qquad e^{-m_\theta(x_i,y_i)} \le e^{-m_{\varepsilon,\theta}(x_i,y_i)}.$$

Insert these bounds into $\frac{1}{K}Q_{\mathcal{D}}(\theta) \le \mathrm{ECE}_M(\theta;\mathcal{D}) \le (K-1)Q_{\mathcal{D}}(\theta)$ to obtain

$$\frac{1}{K}Q^-_{\varepsilon,\mathcal{D}}(\theta) \le \mathrm{ECE}_M(\theta;\mathcal{D}) \le (K-1)Q^0_{\varepsilon,\mathcal{D}}(\theta).$$

For $\lambda_{\max}(H_{\mathrm{GN}}(\theta;\mathcal{D}))$, repeat the argument in (ii) and apply $e^{-m_\theta(x_i,y_i)} \le e^{-m_{\varepsilon,\theta}(x_i,y_i)}$ in the final step to get

$$\lambda_{\max}\big(H_{\mathrm{GN}}(\theta;\mathcal{D})\big) \le 2C_J^2(K-1)\,Q^0_{\varepsilon,\mathcal{D}}(\theta).$$

□

B.6. Connection to the Observed Train/Test Split

Theorems 4.1–4.2 give a mechanism-level account of the dynamics in Section 3. Throughout training, both ECE and Gauss–Newton sharpness are controlled by the same margin-tail functional; once the training set enters the interpolating regime, this coupling becomes two-sided. This explains the strong co-evolution of training ECE and sharpness well before convergence.

The same picture also explains why held-out behavior can decouple. First, the bounds depend on the *true* margin $m_\theta(x, y)$, not the predicted one: a model can therefore become more confidently wrong on a subset of test points, so predicted margins increase while test ECE worsens. Second, the sharpness control carries a geometry factor through the Jacobian. Even when $H_z(p)$ contracts as predictions become more one-hot, large Jacobian norms can keep curvature proxies large.

Theorem 4.1 bounds ECE on any distribution – including the test distribution – by the robust-margin moment on that same distribution. As a target for training, this is not directly actionable: bounding test ECE through Theorem 4.1 requires Q computed on test data. The next result closes this gap by replacing the test-side moment with an empirical, training-side one, under a local label-preserving shift assumption.

Proposition B.7 (Robust margins as a local distributional worst case)**.** *Let $\hat{P}_n := \frac{1}{n}\sum_{i=1}^{n}\delta_{(x_i,y_i)}$ and $Q^0_{\varepsilon,\mathcal{D}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} e^{-m_{\varepsilon,\theta}(x_i,y_i)}$, the robust counterpart of $Q_{\mathcal{D}}(\theta)$. Let $\Pi(P, \hat{P}_n)$ denote the set of couplings between P and $\hat{P}_n$, and define*

$$\mathcal{B}^{\mathrm{lp}}_\varepsilon(\hat{P}_n) := \Big\{P : \exists\mu \in \Pi(P,\hat{P}_n) \text{ such that } \mu\big(\{((x,y),(x',y')) : y = y',\ \|x - x'\| \le \varepsilon\}\big) = 1\Big\}.$$

Then, for every θ,

$$Q^0_{\varepsilon,\mathcal{D}}(\theta) = \sup_{P\in\mathcal{B}^{\mathrm{lp}}_\varepsilon(\hat{P}_n)} \mathbb{E}_P\big[e^{-m_\theta(X,Y)}\big].$$

Under local label-preserving shift, the empirical robust margin moment is exactly the worst clean-margin tail over the corresponding uncertainty set.

Corollary B.8 (Conditional transfer to out-of-sample calibration)**.** *If* $P_{\text{test}} \in \mathcal{B}_\varepsilon^{\text{lp}}(\hat{P}_n)$*, then*

$$\text{ECE}_M(\theta; P_{\text{test}}) \le (K-1)\, Q^0_{\varepsilon,\mathcal{D}}(\theta).$$

B.7. Proof of Proposition B.7

Proof. We prove the two inequalities separately.

Upper bound. Fix $P \in \mathcal{B}_\varepsilon^{\text{lp}}(\hat{P}_n)$, and let $\mu \in \Pi(P, \hat{P}_n)$ be a feasible coupling. For μ-a.e. pair $((x,y),(x',y'))$, we have $y = y'$ and $\|x - x'\| \le \varepsilon$. Hence

$$m_\theta(x,y) \ge \inf_{\|\delta\| \le \varepsilon} m_\theta(x'+\delta, y') = m_{\varepsilon,\theta}(x',y').$$

Since $t \mapsto e^{-t}$ is decreasing,

$$e^{-m_\theta(x,y)} \le e^{-m_{\varepsilon,\theta}(x',y')}.$$

Taking expectation under μ,

$$\mathbb{E}_P\big[e^{-m_\theta(X,Y)}\big] = \mathbb{E}_\mu\big[e^{-m_\theta(X,Y)}\big] \le \mathbb{E}_\mu\big[e^{-m_{\varepsilon,\theta}(X',Y')}\big] = \mathbb{E}_{\hat{P}_n}\big[e^{-m_{\varepsilon,\theta}(X,Y)}\big] = Q^0_{\varepsilon,D}(\theta).$$

Therefore

$$\sup_{P \in \mathcal{B}_\varepsilon^{\text{lp}}(\hat{P}_n)} \mathbb{E}_P\big[e^{-m_\theta(X,Y)}\big] \le Q^0_{\varepsilon,D}(\theta).$$

Lower bound. Fix $\eta > 0$. By definition of the infimum, for each $i \in [n]$ there exists x_i^η such that

$$\|x_i^\eta - x_i\| \le \varepsilon, \qquad m_\theta(x_i^\eta, y_i) \le m_{\varepsilon,\theta}(x_i, y_i) + \eta.$$

Define

$$P^\eta := \frac{1}{n}\sum_{i=1}^n \delta_{(x_i^\eta, y_i)}.$$

Then $P^\eta \in \mathcal{B}_\varepsilon^{\text{lp}}(\hat{P}_n)$, and

$$\mathbb{E}_{P^\eta}\big[e^{-m_\theta(X,Y)}\big] = \frac{1}{n}\sum_{i=1}^n e^{-m_\theta(x_i^\eta, y_i)} \ge e^{-\eta}\,\frac{1}{n}\sum_{i=1}^n e^{-m_{\varepsilon,\theta}(x_i,y_i)} = e^{-\eta}\, Q^0_{\varepsilon,D}(\theta).$$

Taking the supremum over $P \in \mathcal{B}_\varepsilon^{\text{lp}}(\hat{P}_n)$ and then letting $\eta \downarrow 0$ yields

$$\sup_{P \in \mathcal{B}_\varepsilon^{\text{lp}}(\hat{P}_n)} \mathbb{E}_P\big[e^{-m_\theta(X,Y)}\big] \ge Q^0_{\varepsilon,D}(\theta).$$

Combining the two inequalities proves the claim. □

B.8. Proof of Corollary B.8

Proof. By the clean-margin part of Theorem 4.1,

$$\text{ECE}_M(\theta; P_{\text{test}}) \le (K-1)\,\mathbb{E}_{P_{\text{test}}}\big[e^{-m_\theta(X,Y)}\big].$$

If $P_{\text{test}} \in \mathcal{B}_\varepsilon^{\text{lp}}(\hat{P}_n)$, Proposition B.7 implies

$$\mathbb{E}_{P_{\text{test}}}\big[e^{-m_\theta(X,Y)}\big] \le Q^0_{\varepsilon,D}(\theta).$$

Substituting proves the result. □

C. Additional Sharpness–Calibration Experiments

C.1. Sharpness–Calibration Correlation Analysis

We present detailed training dynamics for each optimizer on CIFAR-10 and CIFAR-100, showing the co-evolution of loss, accuracy, ECE, margin, and sharpness throughout training. Since computing the full Hessian eigenvalue is expensive, these experiments use an MLP with a 5K/5K train/validation split. For each dataset, a scatter summary visualizes the temporal coupling across optimizers in a single view; per-optimizer figures report training (left) and validation (right) metrics across learning rates.

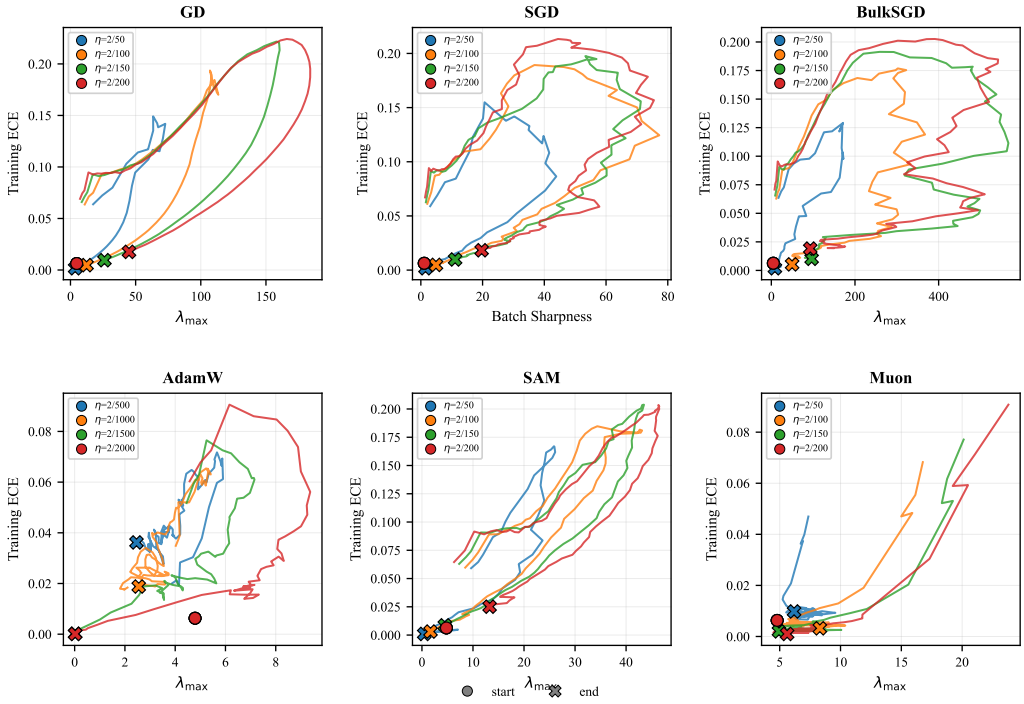


Figure A1. **ECE vs. GN sharpness trajectories (CIFAR-10).** Each curve traces the joint evolution of ECE and GN sharpness ($\lambda_{\max}$) across training steps for one optimizer and learning rate, with a filled circle marking the first training step and a cross ($\times$) the last; color encodes learning rate. Trajectories lie near the diagonal, visualizing the temporal coupling between the two quantities.

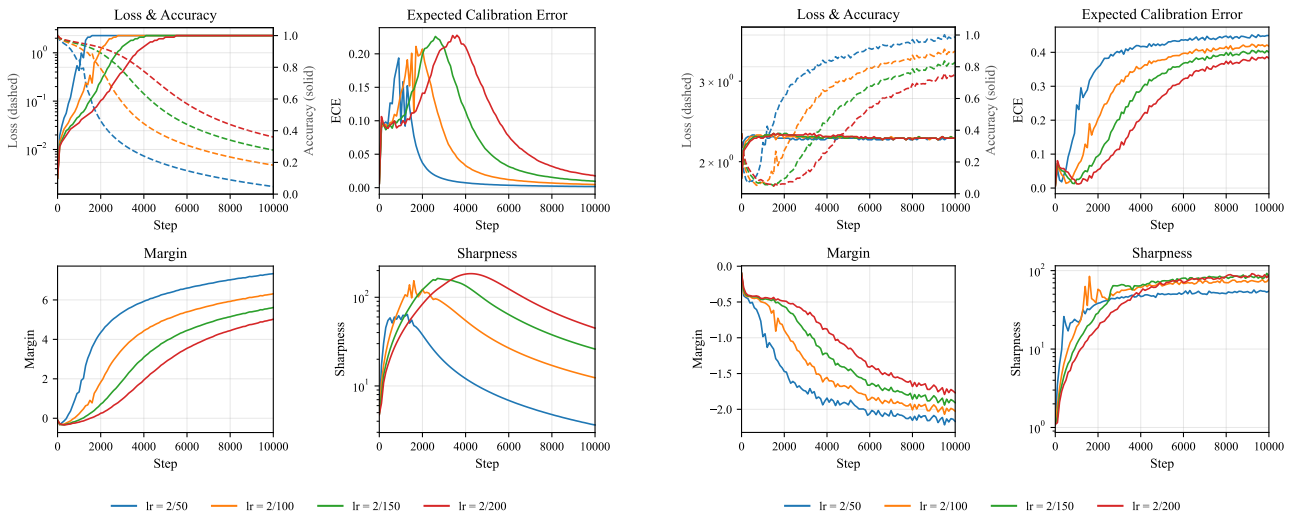
C.2. Optimizer Details: SAM, Muon, and BulkSGD

There is literature to support the notion that SAM may lead to improved calibration metrics, specifically that SAM act as an implicit regularizer and therefore prevents overfitting during training (Tan et al., 2026). At every step, SAM solves

$$\min_{w} \max_{||\epsilon|| \leq \rho} L(w + \epsilon)$$

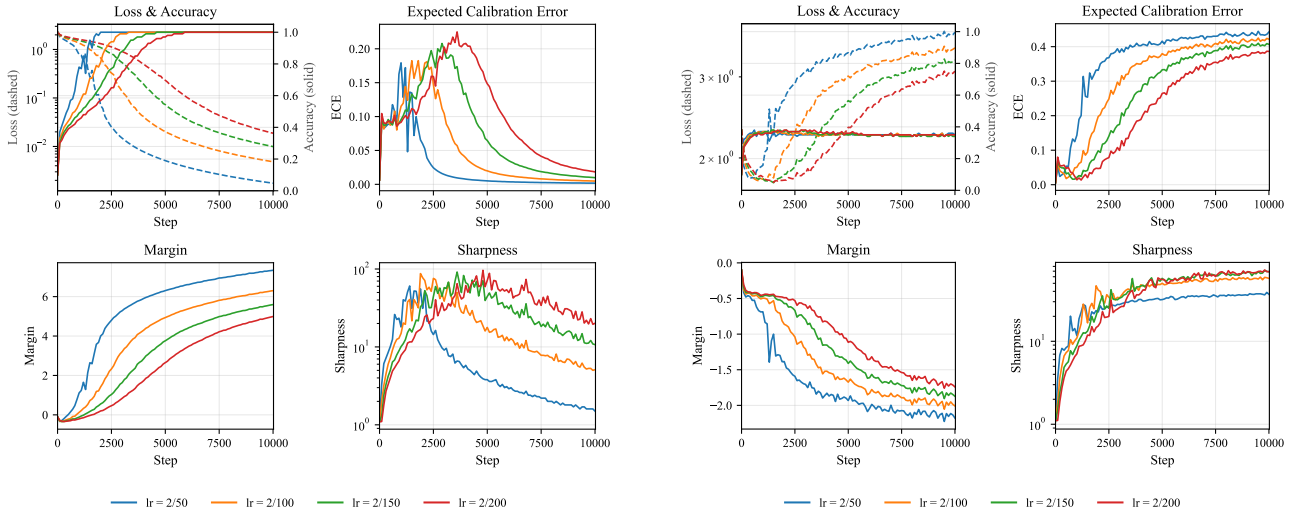
which explicitly penalizes the sharpness of the Hessian and leads to convergence to flatter minima (Zhou et al., 2025). We train networks using SAM to test the first hypothesis, looking to confirm that flat minima lead to lower calibration error.

To test the second hypothesis, we apply optimizers that explicitly suppress the contribution of eigenvectors associated with directions of steep descent, through Muon and BulkSGD. Muon rescales the gradient components at each update, so all directions contribute with comparable magnitude. This means directions of steep descent are clamped, while the flatter directions are amplified (Jordan et al., 2024). Another method to suppress directions of steepest descent is using BulkSGD, which at each step projects the gradient to the space orthogonal to the subspace spanned by the top eigenvectors. We try projecting out the top eigenvector, as well as the top three and five eigenvectors (Song et al., 2025). We note that with BulkSGD, we entirely omit the directions of steepest descent, while with Muon we still allow small updates to be made in those directions.



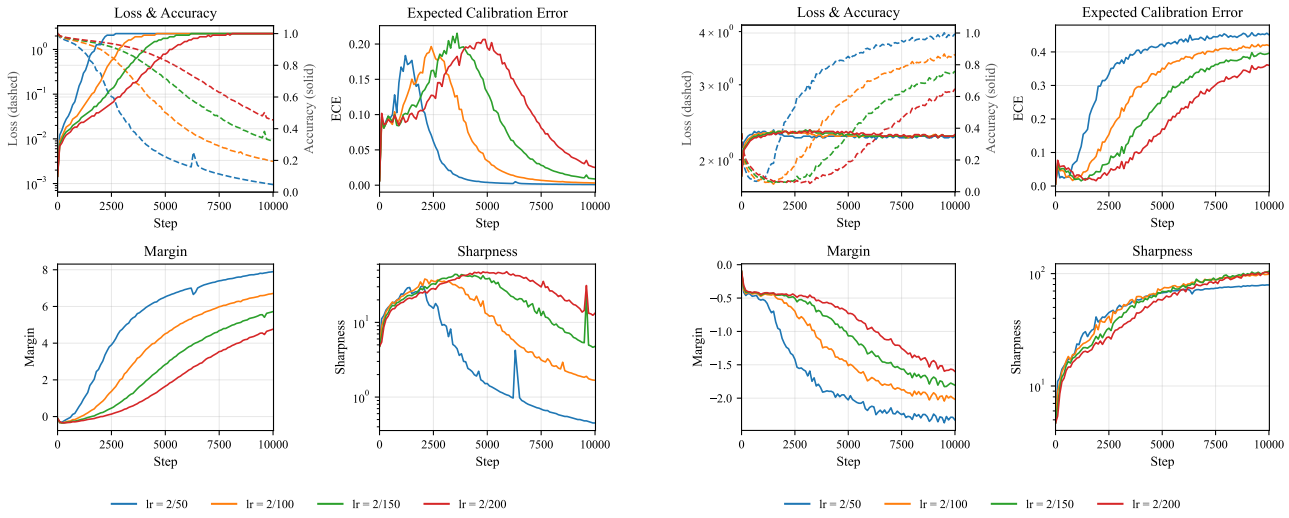
(a) Training metrics *(b)* Validation metrics

Figure A2. **Gradient Descent (GD).** Training dynamics (loss, accuracy, ECE, margin, sharpness) across four learning rates on CIFAR-10; training (left) and validation (right).



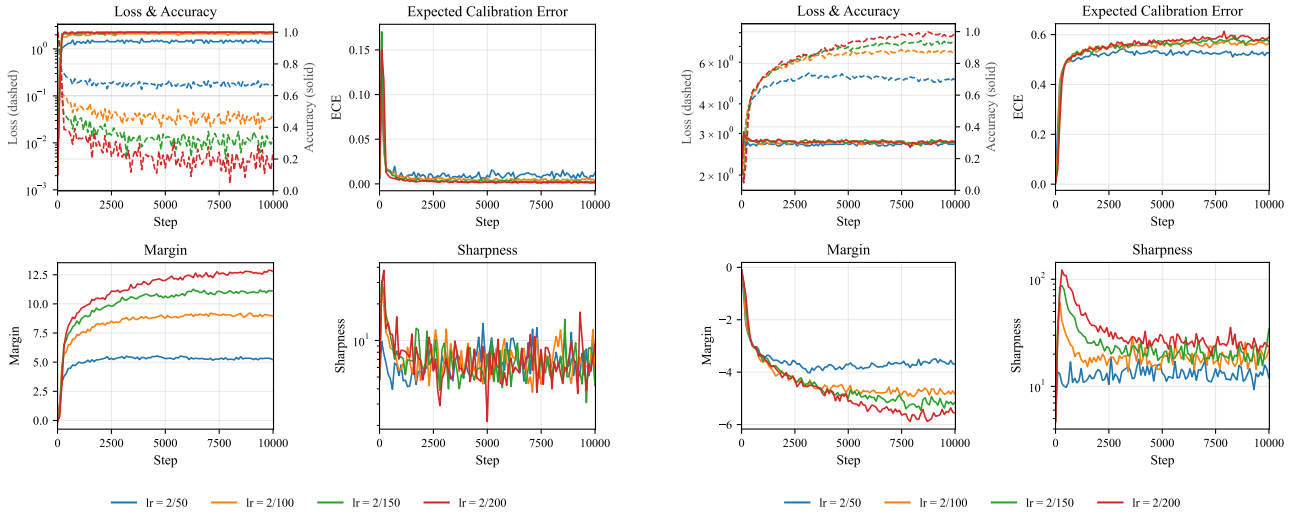
(a) Training metrics *(b)* Validation metrics

Figure A3. **Stochastic Gradient Descent (SGD).** Training dynamics (loss, accuracy, ECE, margin, sharpness) across four learning rates on CIFAR-10; training (left) and validation (right).



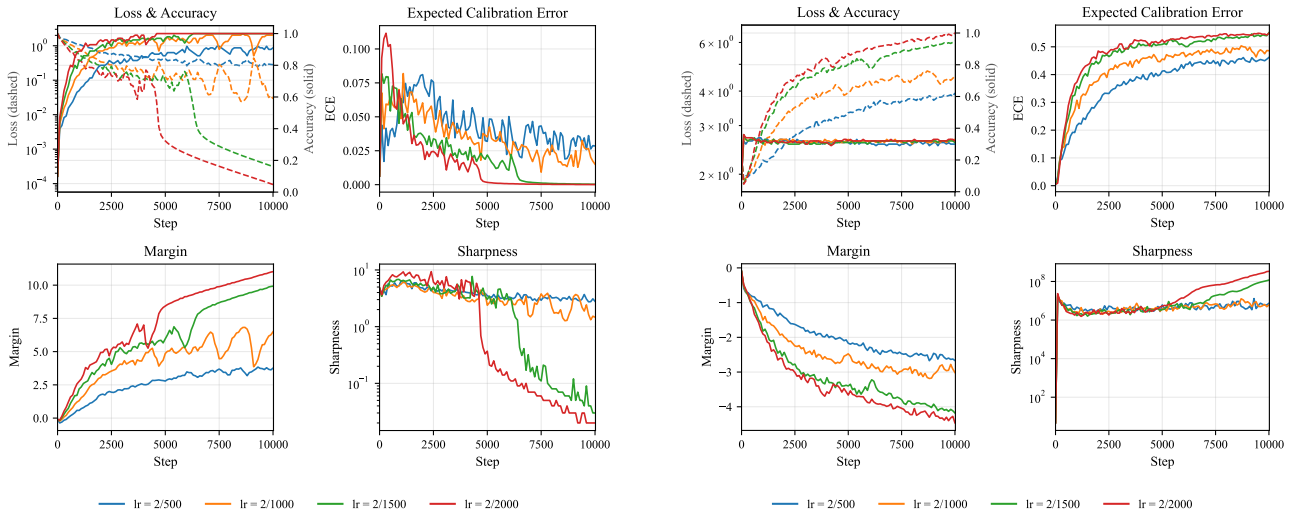
(a) Training metrics *(b)* Validation metrics

Figure A4. **Sharpness-Aware Minimization (SAM).** Training dynamics (loss, accuracy, ECE, margin, sharpness) across four learning rates on CIFAR-10; training (left) and validation (right).



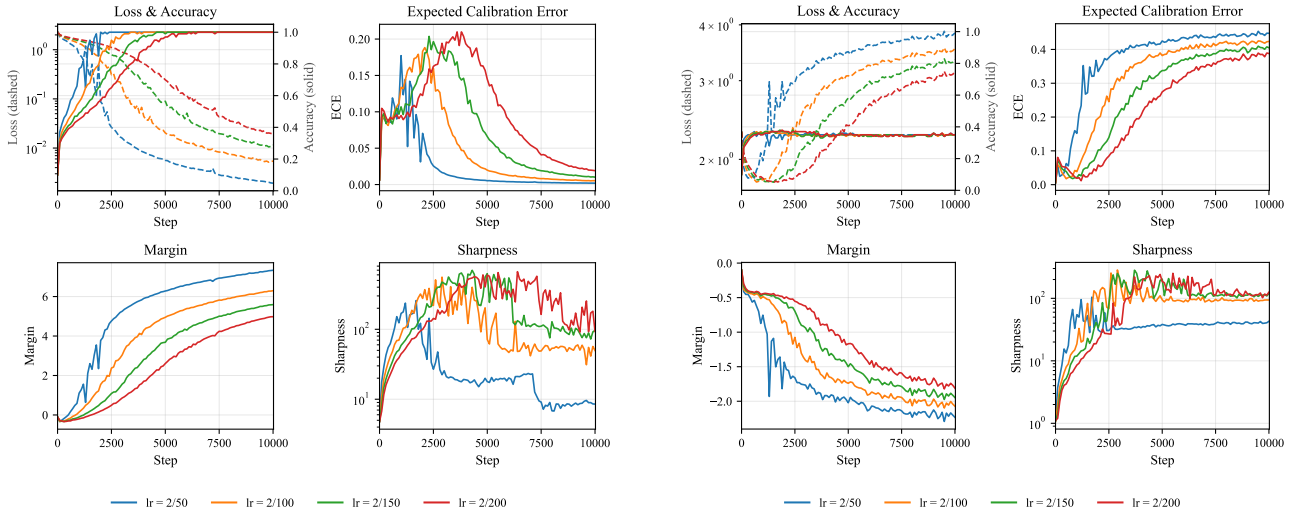
(a) Training metrics *(b)* Validation metrics

Figure A5. **Muon.** Training dynamics (loss, accuracy, ECE, margin, sharpness) across four learning rates on CIFAR-10; training (left) and validation (right).



(a) Training metrics *(b)* Validation metrics

Figure A6. **AdamW.** Training dynamics (loss, accuracy, ECE, margin, sharpness) across four learning rates on CIFAR-10; training (left) and validation (right).



(a) Training metrics *(b)* Validation metrics

Figure A7. **BulkSGD.** Training dynamics (loss, accuracy, ECE, margin, sharpness) across four learning rates on CIFAR-10; training (left) and validation (right).

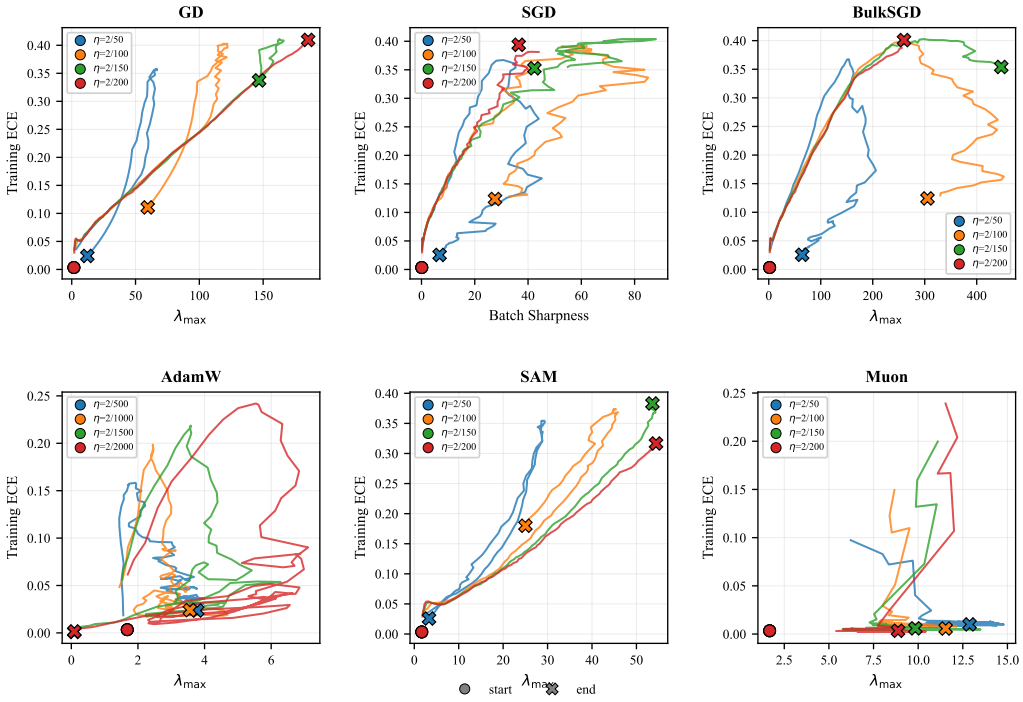
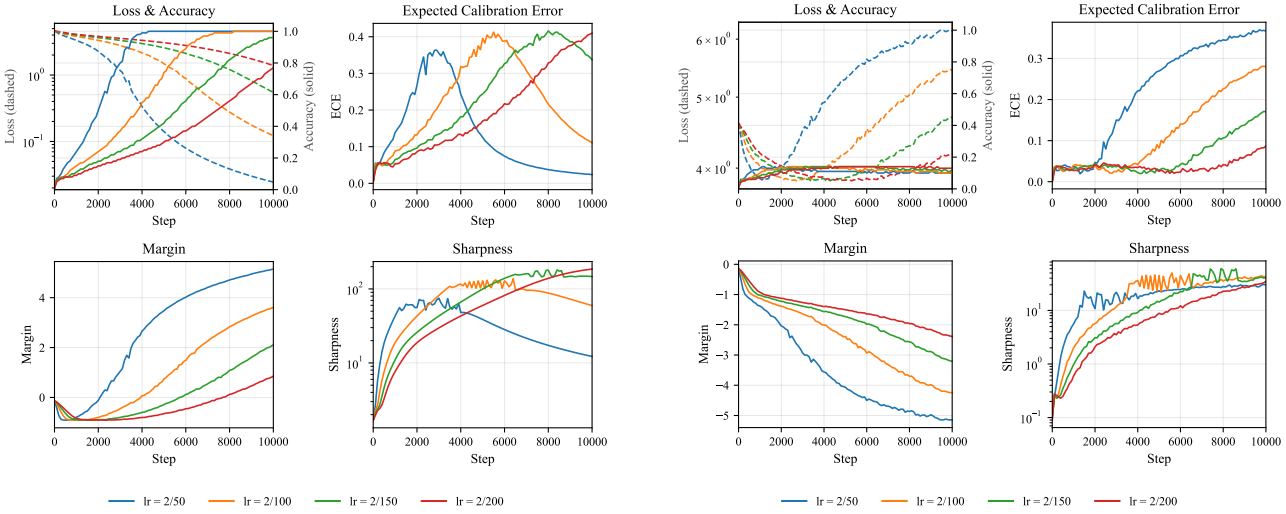


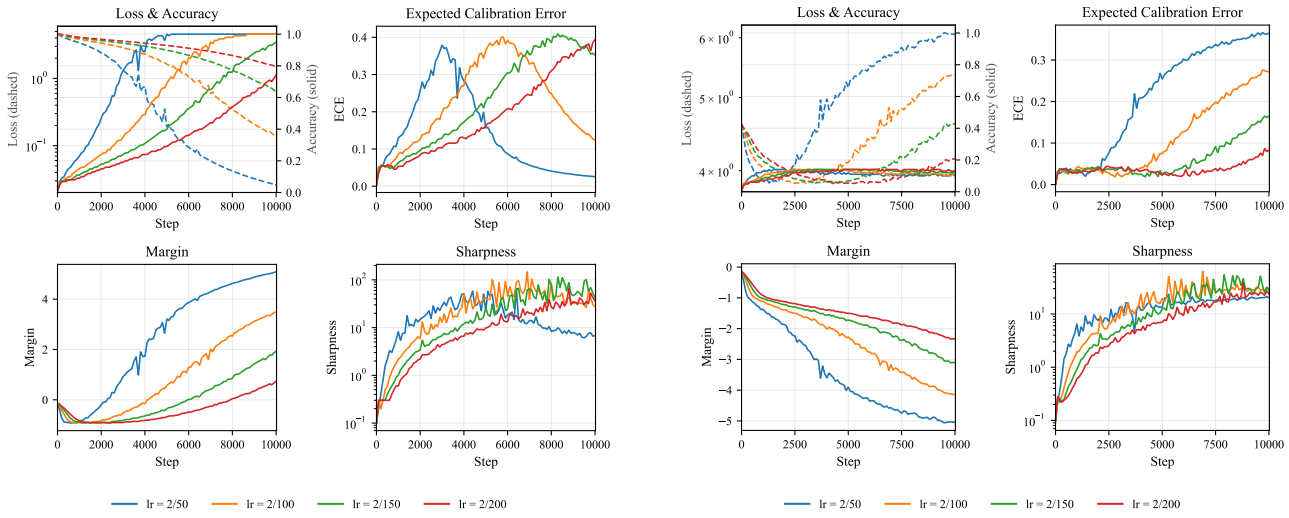
Figure A8. **ECE vs. GN sharpness trajectories (CIFAR-100).** Same format as Figure A1.



(a) Training metrics

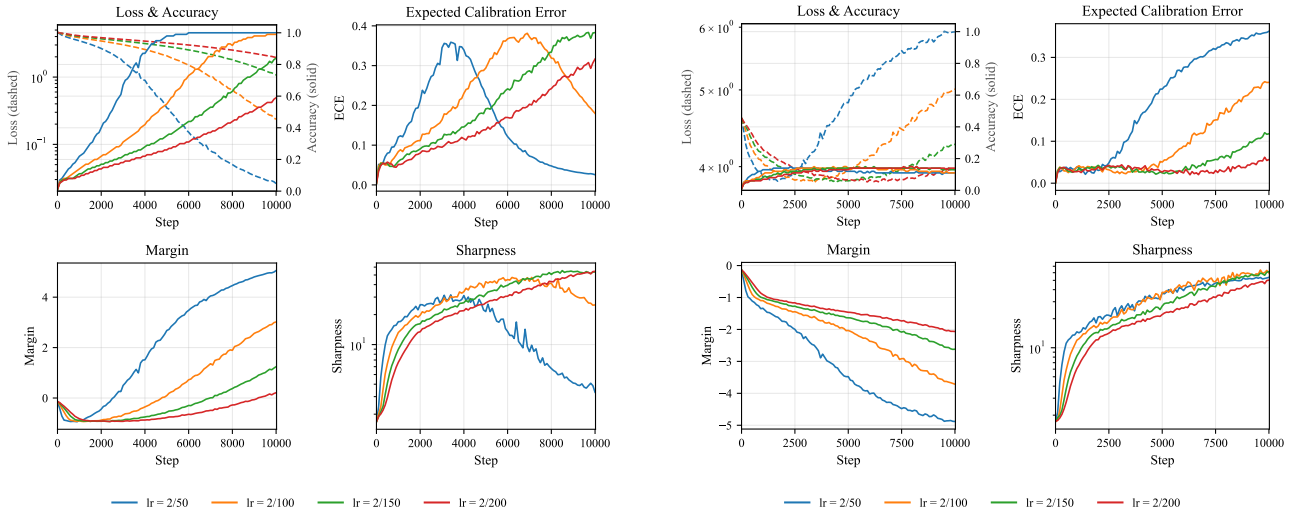
(b) Validation metrics

Figure A9. **Gradient Descent (GD) — CIFAR-100.** Training dynamics (loss, accuracy, ECE, margin, sharpness) across learning rates on CIFAR-100; training (left) and validation (right).



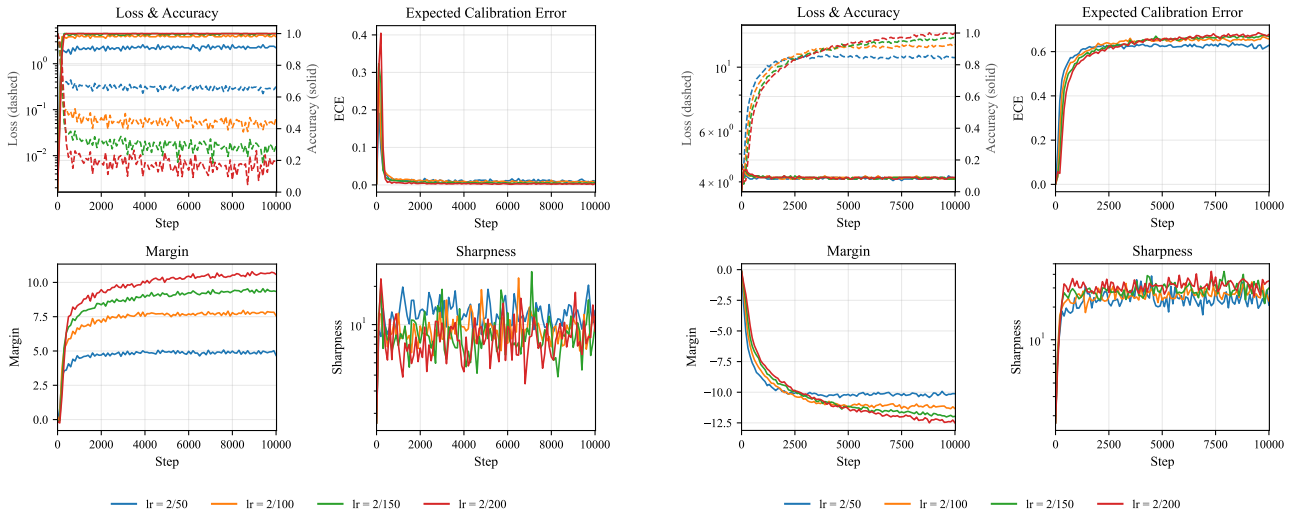
(a) Training metrics *(b)* Validation metrics

Figure A10. **Stochastic Gradient Descent (SGD) — CIFAR-100.** Training dynamics (loss, accuracy, ECE, margin, sharpness) across learning rates on CIFAR-100; training (left) and validation (right).



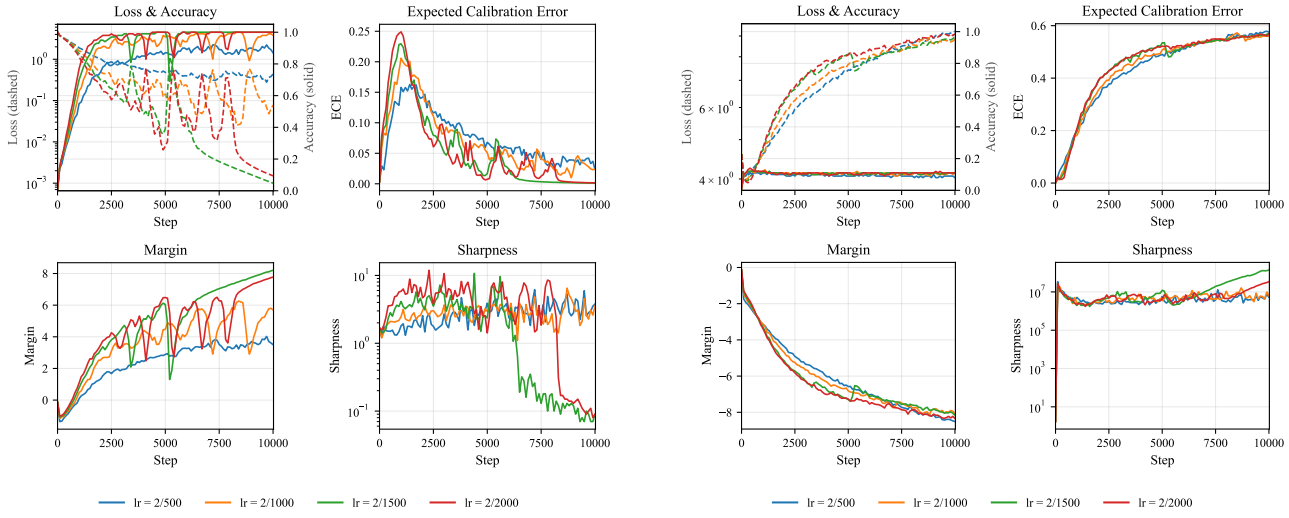
(a) Training metrics *(b)* Validation metrics

Figure A11. **Sharpness-Aware Minimization (SAM) — CIFAR-100.** Training dynamics (loss, accuracy, ECE, margin, sharpness) across learning rates on CIFAR-100; training (left) and validation (right).



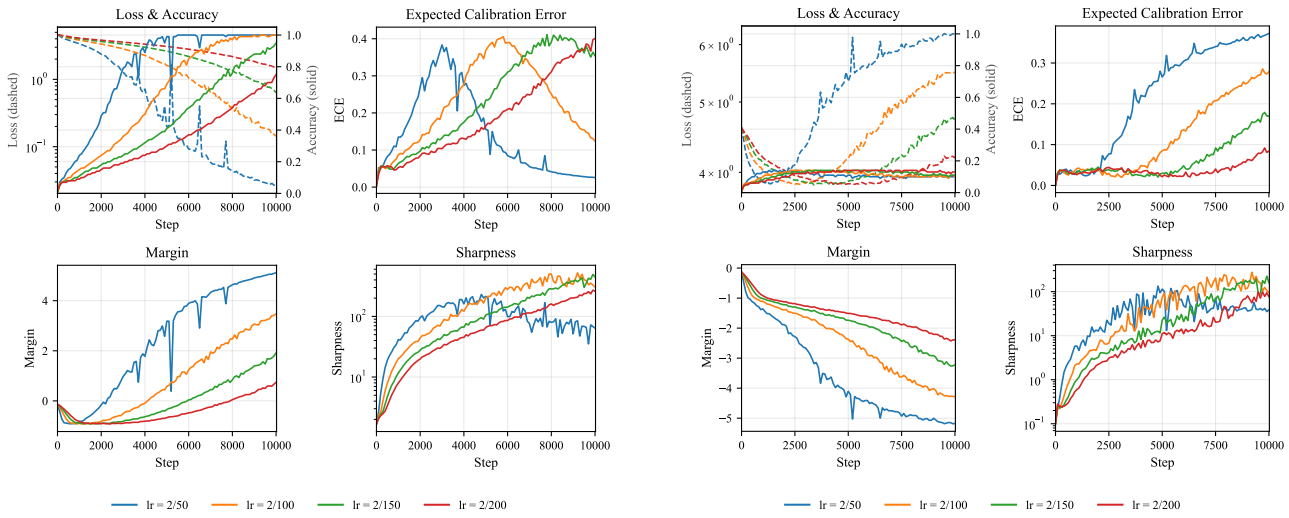
(a) Training metrics *(b)* Validation metrics

Figure A12. **Muon — CIFAR-100.** Training dynamics (loss, accuracy, ECE, margin, sharpness) across learning rates on CIFAR-100; training (left) and validation (right).



(a) Training metrics *(b)* Validation metrics

Figure A13. **AdamW — CIFAR-100.** Training dynamics (loss, accuracy, ECE, margin, sharpness) across learning rates on CIFAR-100; training (left) and validation (right).



(a) Training metrics

(b) Validation metrics

Figure A14. **BulkSGD — CIFAR-100.** Training dynamics (loss, accuracy, ECE, margin, sharpness) across learning rates on CIFAR-100; training (left) and validation (right).

For BulkSGD, training suffers from high levels of instability depending on the number of dominant eigenvectors that are projected out. We observe that training loss is still minimized over 100,000 steps, however the trajectory features steep oscillations. Similarly, sharpness explodes to values in the thousands, which has not been previously observed with other optimizers. This could be due to the fact that, with the dominant eigenvectors projected out, the gradient continues to remain in areas of high curvature, without following the directions of steepest descent.

D. Extension to Mean Squared Error

D.1. GD and SGD Experiments

We rerun the experimental setup of Section 3 with mean squared error (MSE) loss in place of cross-entropy, on 5000 CIFAR-10 training samples. Figures A16 and A17 report the training dynamics for gradient descent and stochastic gradient descent, respectively.

The results show that MSE is extremely miscalibrated, resulting in severely underconfident models as evidenced by the reliability diagrams. This is a consequence of the fact that MSE is not a proper scoring rule.

For L_{MSE}, treating $z_\theta(x_i)$ as free variables, the unique minimizer for a single example is

$$z_k^\star(x_i) = \begin{cases} 1, & k = y_i, \\ 0, & k \neq y_i. \end{cases}$$

Thus the loss penalizes pushing $z_\theta(x_i)_{y_i}$ above 1 or the other logits below 0. However, when using the logits in the softmax before the ECE computation, this finite target pattern leads to underconfidence. In fact, at the optimum, the confidence is

$$\widehat{P}(x_i) = p_\theta(x_i)_{y_i} = \frac{e^1}{e^1 + (K-1)e^0} = \frac{e}{e + K - 1} \approx 0.23 \quad \text{for } K = 10.$$

Consequently, in regimes where training accuracy is close to 1 but logits are near this finite pattern, the model is systematically underconfident on the training set (accuracy ≈ 1 vs. confidence ≈ 0.23 in the main bin), and the training ECE remains large instead of decaying towards zero as in the CE case.

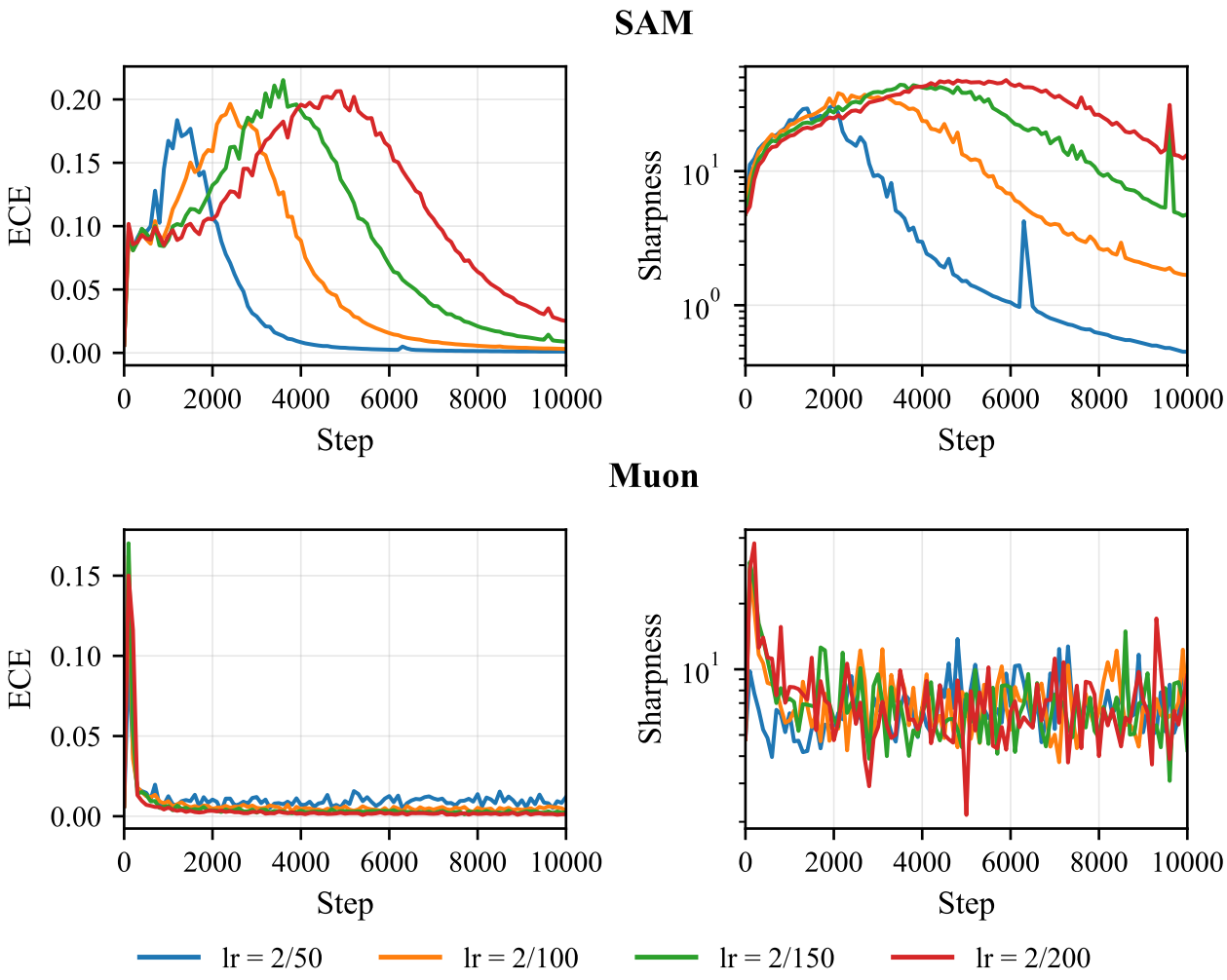


Figure A15. Training dynamics for SAM and Muon across learning rates on CIFAR-10.

D.2. Asymptotics of MSE

For MSE,

$$H_{z,i}^{\mathrm{MSE}}(\theta) \;=\; \nabla_{z_i}^2 L_{\mathrm{MSE}}(z_i, y_i) \;=\; \frac{2}{K} I_K,$$

so the logit-level Hessian is constant and does not depend on the predicted probabilities $p_\theta(x_i)$. Hence, unlike CE, the Gauss–Newton curvature does not attenuate as the model improves its fit:

$$H_{\mathrm{GN}}^{\mathrm{MSE}}(\theta) = \frac{1}{n}\sum_i J_i(\theta)^\top H_{z,i}^{\mathrm{MSE}} J_i(\theta) = \frac{2}{K}\cdot\frac{1}{n}\sum_i J_i(\theta)^\top J_i(\theta),$$

so the eigenvalues of $H_{\mathrm{GN}}^{\mathrm{MSE}}(\theta)$ are governed entirely by the Jacobians $J_i(\theta)$, with no probability-dependent factor $\mathrm{diag}(p) - pp^\top$ to drive curvature toward zero. Combined with the underconfidence analysis of Appendix D.1, this explains why neither sharpness nor training ECE collapse under MSE in the interpolation regime, in contrast with the CE case.

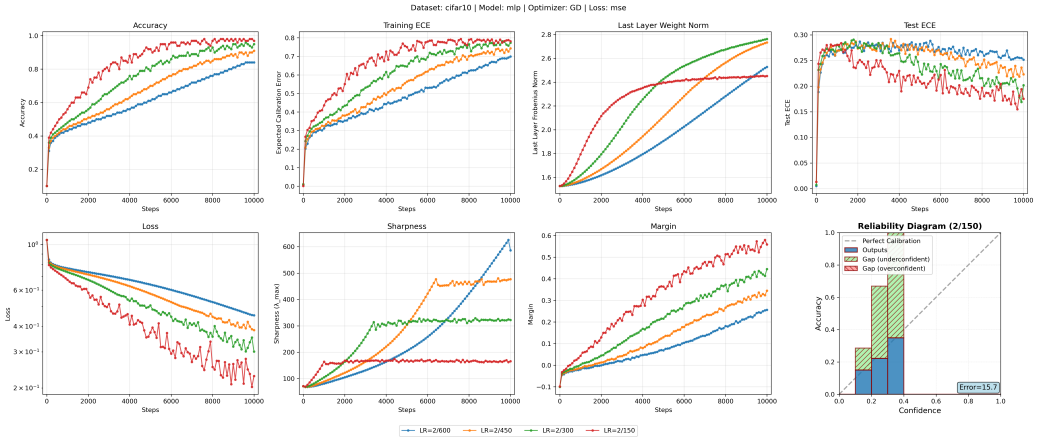


Figure A16. CIFAR-10 | Optimizer: Gradient Descent | Loss: Mean Squared Error

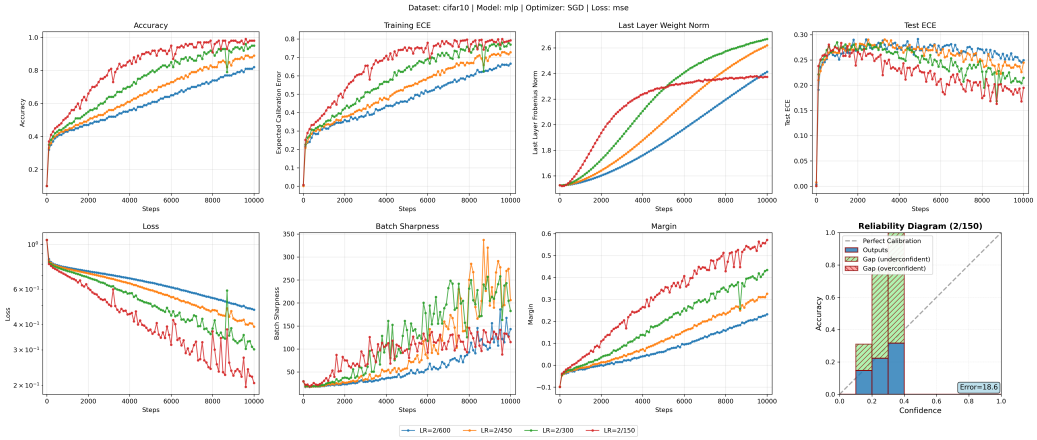


Figure A17. CIFAR-10 | Optimizer: Stochastic Gradient Descent | Loss: Mean Squared Error