

Cosine-Scheduled Parameter EMA Does Not Stabilize Raw Calibration in Small Noisy Classifiers

Anonymous Authors

July 28, 2026

Abstract

Parameter averaging is often used as a cheap way to smooth stochastic optimization trajectories, and a natural hypothesis is that the averaging horizon should grow as a cosine learning-rate schedule decays. We evaluate this hypothesis for small supervised classifiers under symmetric training-label noise. The proposed readout is a deterministic cosine-scheduled exponential moving average of parameters, compared under matched seeds with the final iterate, fixed-coefficient EMAs, and uniform tail Polyak averages. The executed study covers Gaussian blobs, two moons, handwritten digits, and breast-cancer classification, with 20 deterministic seeds per dataset and noise level, clean validation and test labels, and raw plus temperature-scaled probability metrics. The main result is negative: in the preregistered two-hidden-layer MLP condition, cosine-scheduled EMA achieved 0 of 4 dataset-level primary successes against the strongest non-scheduled averaging baseline on raw top-label ECE. Raw ECE was worse for cosine-scheduled EMA on all four datasets, even though test NLL and accuracy often improved. The algebraic filter properties of the schedule were verified, and the readout reduced late validation-NLL variance relative to a fixed EMA with coefficient 0.999, but sharpness and path-movement diagnostics did not support a broad smoothing explanation. These findings reject the headline calibration-stability claim for the tested small-classifier regime while preserving a narrower conclusion: schedule-coupled averaging is a valid convex readout with useful loss-stability behavior, but raw calibration gains are metric-dependent and cannot be inferred from smoothing alone.

1 Introduction

Small classification problems are a useful stress test for claims about calibration because they remove many confounders that dominate large-scale training. The model is small enough to retrain across many deterministic seeds, the datasets can be generated or loaded without external sampling drift, and symmetric label noise can be injected after splitting so that validation and test labels remain clean. At the same time, these problems are not trivial for probabilistic prediction. A classifier may obtain high accuracy while assigning poorly calibrated probabilities, and the same trained weights can look different under expected calibration error, negative log likelihood, Brier score, and reliability diagrams [2, 7, 12, 17, 35]. This tension matters whenever a training intervention is proposed as a calibration intervention rather than only an accuracy intervention.

The broader relevance is that calibration is a deployment property, while many optimization changes are justified by training dynamics. A model used for triage, ranking, or downstream decision support may need probability estimates whose magnitudes are meaningful, not merely class decisions that are often correct. Small classifiers are a modest regime, but they expose a general methodological risk: a smoother parameter trajectory can improve one scalar objective while degrading the relationship between confidence and observed correctness. Because the datasets are small enough for seed-paired analysis, this regime lets us separate a plausible optimization story from a measured calibration outcome with less dependence on single-run anecdotes.

The intervention studied here is a deterministic exponential moving average of parameters whose coefficient increases as a cosine learning-rate schedule decays. Fixed EMA, Polyak-Ruppert averaging, averaged SGD, and stochastic weight averaging all motivate the broader idea that compatible iterates along an optimization path can be combined into a more stable readout [6, 13, 23, 38]. Cosine annealing and cyclical learning rates separately motivate time-varying optimization schedules [10, 33]. The open question is whether coupling these ideas improves calibration under label noise, beyond the strongest non-scheduled averaging baseline that sees the same optimizer path. The coupling is plausible: early training uses shorter smoothing memory while the optimizer is moving quickly, and late training uses a longer memory when step sizes shrink. Plausibility, however, is not evidence that the readout suppresses miscalibration rather than simply changing the loss-accuracy tradeoff.

The study therefore treats the coupling as a falsifiable design choice rather than as a theorem waiting to be illustrated. The scheduled coefficient may be a good low-pass filter for parameter noise, yet calibration depends on how softmax confidence changes across bins on held-out examples. The same averaged point can reduce validation NLL by assigning more mass to true labels while still moving many predictions into confidence regions whose empirical accuracy is less aligned. This possibility is especially important under label noise, where late training can create sharp loss changes and overconfident readouts without making the clean-label test distribution easier to calibrate.

The design deliberately uses small tasks because the claim is about reliability across controlled conditions rather than about one large benchmark score. With 20 seeds per condition, paired readout comparisons can distinguish systematic effects from split-specific variation. With symmetric noise applied only to training labels, the evaluation keeps a clean target for calibration and validation loss. With identical optimizer trajectories for every readout, the method comparison isolates the readout map from the training process. These restrictions reduce external breadth, but they make the hypothesis sharper. A scheduled EMA that cannot beat strong averaging in this controlled setting should not be promoted as a calibration-stability method without additional evidence.

The title states the negative conclusion because the evidence does not support a weaker promotional framing. A paper could have emphasized that CEMA improves NLL or late validation-loss variance, but that would obscure the requested calibration claim. Instead, the manuscript follows the hierarchy of the evaluation plan. The primary target is raw top-label ECE under noisy training labels. Secondary metrics explain mechanism and deployment-adjacent behavior. The conclusion is bounded to the tested small classifiers. This organization keeps the reader from confusing a supported secondary effect with the rejected headline effect.

This paper reports the executed answer for the requested small-classifier regime. Each run trains the same architecture and optimizer trajectory and then evaluates several readouts: final weights, fixed EMAs, uniform tail Polyak averages, and the proposed cosine-scheduled EMA. The primary acceptance rule is deliberately strict. For the two-hidden-layer MLP, cosine-scheduled EMA must lower mean raw top-label ECE by at least 10% relative to the best non-scheduled averaging baseline in at least three of four datasets for label-noise levels at or above 10%, pass a paired one-sided Wilcoxon test across matched seeds, and avoid an accuracy loss larger than one absolute percentage point. That rule prevents a favorable post-hoc metric, weaker baseline, or hidden accuracy sacrifice from being mistaken for a calibration improvement.

The emphasis of the paper follows from that rule. We do not present CEMA as a general-purpose robust-learning method, because it does not modify targets, losses, sample weights, or data selection. We also do not compare it primarily with ensembles or post-hoc probability calibrators, because those methods answer different intervention questions. Instead, the paper asks whether a deterministic readout schedule is better than fixed and uniform readouts on exactly the same training path. That design makes a negative result informative: it identifies a boundary of the readout idea rather than blending it with stronger or more expensive interventions.

The result rejects the headline hypothesis. Across Gaussian blobs, two moons, digits, and breast cancer, cosine-scheduled EMA met none of the four dataset-level primary success predicates. The strongest non-scheduled baseline had lower raw ECE in every dataset, with one-sided Wilcoxon p -values near one for the direction that would favor CEMA. The failure is not a generic collapse of the readout: CEMA often improved test NLL and accuracy relative to the same baseline, and it reduced late validation-NLL variance relative to a fixed EMA with coefficient 0.999. The important lesson is narrower and more useful than a simple win or loss. The schedule is a valid finite-horizon convex filter, but its low-pass interpretation does not predict raw calibration gains under this evaluation rule.

The contributions are:

- We define a cosine-coupled EMA readout for finite parameter trajectories and state the mathematical regime in which it is a convex parameter filter.
- We execute a paired small-classifier study over four datasets, four training-label-noise levels, two MLP widths, a logistic reference, and horizon-bracketed averaging baselines.
- We show that the preregistered raw-ECE calibration criterion is not met: the primary two-hidden-layer MLP condition records 0 of 4 dataset-level successes.
- We separate calibration failure from loss-stability behavior by reporting NLL, accuracy, reliability curves, metric sensitivity, validation-NLL variance, path movement, and sharpness proxies.
- We identify the boundary on which the method remains defensible: algebraic schedule validity and late-loss stabilization are supported, while broad raw-calibration improvement is unsupported in the tested regime.

2 Related Work

The relevant literature separates into three clusters that must be kept distinct when interpreting this study. Calibration work defines the target and warns that finite-sample estimators can disagree. Averaging and schedule work motivates the readout family and the comparator grid, but does not transfer directly into a calibration guarantee for finite nonconvex MLP training. Noisy-label work explains why corrupted training labels can change loss dynamics, while also showing why objective-changing robust methods are adjacent rather than primary baselines. The manuscript contribution sits at the intersection: it tests one parameter-readout intervention under a fixed objective and reports the result even though the central calibration claim is rejected.

A useful comparison matrix has two axes. The first axis is the object being changed: predictions, parameters, optimizer schedules, labels, or losses. The second is the claim being measured: confidence calibration, proper-score quality, accuracy, trajectory stability, or robustness to corrupted supervision. Temperature scaling changes predictions after training and is evaluated directly on calibration metrics. Robust noisy-label methods change the training objective or sample treatment. Weight averaging changes the readout point in parameter space. CEMA belongs to the last category, so the gap is not the absence of any calibration or noise-robustness method. The gap is the absence of direct evidence that schedule-coupled parameter averaging improves raw calibration beyond strong non-scheduled averaging on matched noisy-label trajectories.

This matrix also explains why the related work does not define the baseline set by every method that can improve calibration. Deep ensembles and post-hoc calibration are strong probability tools, but they change inference cost or prediction transformations. Robust losses and sample-selection methods directly address label noise, but they change the training objective. Those methods are important for context and future comparisons, yet they would blur the first-order question about parameter averaging. The local contribution is narrower: it asks whether a deterministic schedule for the EMA coefficient adds measurable calibration value when the training path is held fixed. That is the scientific boundary that the results evaluate.

2.1 Calibration Metrics and Probability Readouts

Calibration evaluation asks whether predicted probabilities agree with observed frequencies, but the literature shows that this question has several non-equivalent empirical answers. Temperature scaling made ECE, NLL, and reliability diagrams central diagnostics for neural classifiers and demonstrated that accuracy can be largely unchanged while confidence quality changes [2]. Later work emphasized that calibration behavior depends on model family, scale, pretraining, and distributional setting [12]. Metric studies further show that fixed-bin ECE, adaptive-bin variants, class-conditional estimates, and verified calibration procedures can rank models differently, especially under finite sample sizes [7, 17, 19, 22, 35, 37]. This manuscript therefore treats raw top-label ECE as the primary prespecified metric but does not let it stand alone. NLL, Brier score, reliability curves, and temperature-scaled sensitivity analyses are necessary to explain whether a failure is a calibration-specific failure or a broader probability-quality failure [15, 25, 26, 29, 36].

These metric distinctions are not cosmetic. NLL is a strictly proper score for probabilistic predictions under the clean evaluation distribution, so it rewards assigning probability mass to the true class. Top-label ECE estimates a binned confidence-correctness discrepancy and is sensitive to bin choice, sample size, and confidence distribution. Brier score combines calibration and refinement through squared probability error, and reliability curves reveal bin-level structure that a scalar can hide. A method can therefore improve NLL while worsening raw ECE, particularly when a readout changes confidence magnitudes more than class decisions. This manuscript uses that distinction to interpret the observed contradiction rather than treating one metric as an automatic proxy for all probability quality.

The literature therefore motivates both redundancy and restraint in measurement. Redundancy is needed because no single calibration summary captures all probability behavior. Restraint is needed because a long metric list can invite selective interpretation after the experiment has run. The manuscript resolves this tension by assigning one primary decision metric and then using the other metrics for explanation. The raw ECE rule decides the central hypothesis. NLL, Brier score, reliability curves, and temperature-scaled summaries diagnose why the decision occurred and whether any adjacent use case remains plausible. This is why the abstract reports the rejected primary result before the more favorable secondary results.

Post-hoc calibration methods and parameter readouts occupy different positions in the comparison matrix. Temperature scaling and Bayesian binning directly transform predicted probabilities after a model is trained [2, 15]. Parameter averaging instead changes the weights at which the model is evaluated. That distinction is central because a scheduled EMA may improve validation loss by reducing stochastic variation while leaving confidence bins less aligned with empirical correctness. The metric sensitivity observed in this study follows exactly that contradiction: raw ECE favors the strongest fixed or uniform averaging baseline, while several proper-score and temperature-scaled summaries favor CEMA. The gap motivates a conservative claim boundary rather than a single scalar narrative.

2.2 Averaging, Schedules, and Noisy Labels

Averaging iterates is an old idea with strong roots in stochastic approximation. Polyak-Ruppert averaging and averaged SGD provide formal support in structured regimes, while stochastic weight averaging shows that uniform averages along SGD trajectories can find wider optima in modern neural networks [6, 23, 38]. Later SWA analyses separate the averaging operation from the cyclical or high-constant learning-rate schedule, reinforcing the need to compare the proposed readout against horizon-bracketed averaging baselines rather than against SGD alone [3]. Exponential weight averaging also admits dynamical and filter-like interpretations, which is directly relevant here because CEMA changes only the readout rule on a matched trajectory [21]. Fixed EMA is a geometric smoother used across supervised, semi-supervised, and teacher-student settings [13, 34]. Other averaging extensions, including SWAG, hierarchical weight averaging, Gaussian stochastic weight averaging, and deep ensembles, broaden the connection between weight-space smoothing and uncertainty estimates [1, 8, 11, 18]. These methods

motivate the comparator set, but they do not establish that a cosine-coupled EMA coefficient improves calibration for finite nonconvex MLPs.

The averaging literature also motivates a strong baseline choice. If the proposed schedule is compared only with the final iterate, any improvement might simply reflect the known benefit of averaging compatible weights. If it is compared only with one fixed EMA coefficient, the result may depend on whether that coefficient happens to match the training horizon. This study therefore brackets the scheduled horizon with several fixed coefficients and tail averages. That design follows the literature’s lesson that averaging, sampling schedule, and geometry are separable components. It also makes the central negative result sharper: CEMA fails raw ECE even when the strongest non-scheduled averaging comparator is allowed to vary by dataset.

There is also a novelty boundary. Cosine annealing supplies a time coordinate for the optimizer, and EMA supplies a recursive parameter filter, but their combination does not inherit a calibration guarantee from either source. In a convex stochastic approximation problem, averaging can reduce variance under assumptions that are not satisfied by finite ReLU classifiers trained on noisy labels. In a nonconvex neural network, wider optima or smoother trajectories may correlate with better generalization without implying better confidence calibration. The manuscript therefore treats the schedule-coupled coefficient as a new empirical proposal whose merits must be tested against the known alternatives it resembles.

Learning-rate schedules are another adjacent lineage. SGDR introduced cosine annealing and warm restarts as an optimizer schedule, and cyclical learning-rate work showed that time-varying step sizes can change the optimization trajectory [10, 33]. The present method does not claim a new optimizer. It defines a readout whose smoothing coefficient is a deterministic function of the same no-restart cosine time variable used by the learning rate. That novelty boundary is important: cosine learning-rate evidence grounds the time coordinate, and fixed EMA evidence grounds the recurrence, but their combination remains an empirical intervention until directly tested.

The label-noise literature supplies both the risk model and a warning about comparator fairness. Deep networks can fit random labels, and robust noisy-label methods often change losses, targets, sample selection, or training structure [4, 5, 9, 14, 16, 20, 24, 39–42]. Those methods are valuable context, but they answer a different question from a parameter-readout intervention that keeps the training objective fixed. The fair first comparison is therefore among final iterates, fixed EMAs, and uniform averages on identical optimizer paths. The datasets used here reflect the requested no-download regime: Gaussian blobs, two moons, digits, and breast cancer are all supplied by scikit-learn generators or bundled loaders, and deterministic splitting is handled before training-label corruption [27, 28, 30–32].

Noisy-label methods remain important for interpreting scope. Loss correction, co-teaching, DivideMix, and generalized cross entropy can improve robustness by changing the learning problem itself. A scheduled parameter readout is weaker and cheaper: it observes the same corrupted-label trajectory and chooses a final parameter vector. The advantage of this weaker intervention is causal clarity, because differences among readouts are not confounded with different targets or sample-selection rules. The disadvantage is that it may not address the source of miscalibration under label noise. The executed result supports that caution. It shows that smoothing a trajectory is not equivalent to correcting noisy supervision or recalibrating probabilities.

3 Problem Setting

The formal setting fixes the objects that are shared by every readout comparison. A run consists of one dataset, one deterministic seed, one train-validation-test split, one architecture, and one optimizer trajectory. The readouts differ only in how they map the recorded parameter path to an evaluated parameter vector. This restriction makes the empirical contrast interpretable: if CEMA differs from fixed EMA or Polyak averaging, the difference comes from the readout rule rather than from a different model, split, loss, or label-corruption process. The following definitions also state where the claims do not apply, because the study does not cover regression, multilabel prediction, corrupted held-out labels, or cross-architecture averaging.

The setting is finite throughout. All empirical quantities are computed on finite validation or test sets, all trajectories have a finite number of updates, and all readout families are evaluated on compatible parameter vectors from a single architecture. This matters for both the algebra and the statistics. The convex-filter theorem later in the paper is a statement about finite sequences in Euclidean space, not about asymptotic convergence. The success predicate is a finite-sample decision rule over matched seeds and noise levels, not a population theorem. This framing keeps the mathematical claims and the empirical claims in their proper domains.

A readout is therefore a map $R : \Theta^{T+1} \rightarrow \Theta$, where $\Theta = \mathbb{R}^d$ for a fixed architecture and a fixed parameter ordering. The final iterate, fixed EMA, tail Polyak average, and CEMA are four choices of R applied to the same trajectory $(\theta_0, \dots, \theta_T)$. The evaluated classifier is $x \mapsto f_{R(\theta_{0:T})}(x)$, not an ensemble over time and not a mixture of predictions. This distinction matters because averaging parameters before the softmax is not equivalent to averaging probabilities after the softmax. All claims in the paper are about these parameter readouts and their induced classifiers.

The stochastic part of the setting enters through seeds, mini-batch or data-order choices, and label corruption. For a fixed seed, the corruption process in (1) defines a single noisy training set while leaving S_{valid} and S_{test} clean. For a fixed optimizer path, every readout sees the same sequence of parameter vectors. The paired design compares functions of the same path, which

reduces irrelevant variation and directly tests whether the readout rule matters. Population-level expectations are approximated only by seed sweeps over this finite protocol, so the manuscript avoids asymptotic language when reporting empirical outcomes.

3.1 Classification, Noise, and Metrics

Let $K \in \mathbb{N}_{\geq 2}$ be the number of classes, $p \in \mathbb{N}_{\geq 1}$ the feature dimension, and $\mathcal{X} \subseteq \mathbb{R}^p$ the standardized feature space for one dataset. A clean example is a pair (X, Y) with $X \in \mathcal{X}$ and $Y \in \{1, \dots, K\}$. For each seed, the finite sample is split into disjoint training, validation, and test sets, denoted S_{train} , S_{valid} , and S_{test} . Feature standardization is fitted on S_{train} only and then applied to the held-out sets. The main protocol corrupts only training labels. For a target noise rate $\rho \in \{0, 0.1, 0.2, 0.4\}$, the corrupted label $\tilde{Y}$ is sampled from the symmetric transition matrix

$$\Pr(\tilde{Y} = j \mid Y = i) = \begin{cases} 1 - \rho, & j = i, \\ \rho/(K - 1), & j \neq i, \end{cases} \quad (1)$$

for $i, j \in \{1, \dots, K\}$. This setting excludes regression, multilabel classification, instance-dependent noise, and corrupted validation or test labels. The restriction is deliberate because calibration and NLL claims in this paper are evaluated against clean held-out labels.

A model with parameter vector $\theta \in \mathbb{R}^d$, where $d \in \mathbb{N}_{\geq 1}$ is fixed by the architecture, maps $x \in \mathcal{X}$ to logits $z_\theta(x) \in \mathbb{R}^K$ and probabilities $f_\theta(x) = \text{softmax}(z_\theta(x)) \in \Delta^{K-1}$. For any clean evaluation set $S \subseteq \mathcal{X} \times \{1, \dots, K\}$, the negative log likelihood is

$$\text{NLL}(\theta; S) = -\frac{1}{|S|} \sum_{(x_i, y_i) \in S} \log f_{\theta, y_i}(x_i), \quad (2)$$

with probabilities clipped only for numerical stability as in standard log-loss implementations. Top-label ECE uses the predicted class $\hat{y}_i = \arg \max_k f_{\theta, k}(x_i)$ and confidence $c_i = \max_k f_{\theta, k}(x_i)$. For $M \in \mathbb{N}_{\geq 1}$ confidence bins $I_m \subset [0, 1]$, let $B_m = \{i : c_i \in I_m\}$, $a_m = |B_m|^{-1} \sum_{i \in B_m} \mathbf{1}\{\hat{y}_i = y_i\}$, and $\bar{c}_m = |B_m|^{-1} \sum_{i \in B_m} c_i$, with empty bins contributing zero. The primary calibration estimate is

$$\widehat{\text{ECE}}_M(\theta; S) = \sum_{m=1}^M \frac{|B_m|}{|S|} |a_m - \bar{c}_m|. \quad (3)$$

This is a top-label finite-sample estimate, not a proof of full multiclass calibration. The reported Brier score and reliability curves are complementary summaries, and temperature scaling is used only as a sensitivity layer, not as a replacement for the raw readout comparison. The experimental protocol uses the held-out loss in (2) as the validation and test probability-quality score and uses the binned estimate in (3) as the primary raw calibration score, so later metric disagreements are interpreted as disagreements between explicitly defined estimands rather than as interchangeable evidence.

3.2 Readout Families and Acceptance Rule

For one seed, architecture, dataset, optimizer configuration, and training split, let $\theta_t \in \mathbb{R}^d$ denote the parameter vector after update $t \in \{0, \dots, T\}$ for a finite horizon $T \in \mathbb{N}_{\geq 1}$. All readouts are functions of this same compatible trajectory. The final-iterate readout is θ_T . A fixed EMA with coefficient $\beta \in [0, 1)$ satisfies $\theta_0^\beta = \theta_0$ and $\theta_t^\beta = \beta \theta_{t-1}^\beta + (1 - \beta) \theta_t$ for $t \geq 1$. A uniform tail Polyak readout uses a nonempty index set $A \subseteq \{0, \dots, T\}$ and returns $|A|^{-1} \sum_{t \in A} \theta_t$. The proposed readout, defined in Section 4, replaces the fixed coefficient by a deterministic coefficient β_t tied to cosine time. No readout averages optimizer state, changes architecture, or changes labels.

Let $\mathcal{B}$ be the finite set of non-scheduled averaging baselines for the same run family, excluding the final iterate when the phrase ‘‘averaging baseline’’ is used. For each dataset in the primary two-hidden-layer MLP condition, the strongest baseline b^* is selected from $\mathcal{B}$ by mean raw ECE over matched seeds and the noisy-label levels $\rho \in \{0.1, 0.2, 0.4\}$. The dataset-level success predicate is

$$\mathcal{S}(D) = \mathbf{1} \left\{ \frac{\overline{\text{ECE}}(b^*; D) - \overline{\text{ECE}}(C; D)}{\overline{\text{ECE}}(b^*; D)} \geq 0.10, p_{\text{Wilcoxon}} < 0.05, \overline{\text{Acc}}(C; D) - \overline{\text{Acc}}(b^*; D) \geq -0.01 \right\}. \quad (4)$$

Here the overbars denote matched-seed means over the specified noise levels, and the Wilcoxon test is one-sided in the direction that CEMA has lower ECE. The headline claim requires at least three datasets with $\mathcal{S}(D) = 1$ among the four evaluated datasets. This functional encodes the falsifiability condition: a method that improves NLL or accuracy but fails raw ECE does not pass the primary calibration test.

The baseline selection rule is intentionally conservative. It gives the non-scheduled family the benefit of its best observed raw-ECE readout within each dataset, so CEMA is not rewarded for beating a weak fixed coefficient. It also separates the final iterate from the strongest averaging comparator because the scientific question is about schedule coupling, not about whether any averaging can improve over the last parameter vector. The accuracy guard prevents a readout from winning calibration by becoming much less decisive. The paired test prevents a few noisy seed outcomes from being read as reliable improvement. A positive headline result would therefore require consistent calibration gains under a demanding comparator, and a negative result identifies the point at which the schedule idea fails.

4 Cosine-Scheduled EMA

This section defines the proposed readout and the limited algebraic claim attached to it. The schedule is not presented as a new optimizer and it does not alter the training trajectory. It is a deterministic map from a finite sequence of compatible parameter vectors to one evaluated vector. The reason to include the proof is not to imply a calibration theorem, but to establish that the method is a valid convex filter under the stated coefficient bounds. The empirical sections then test whether that valid filter improves the metrics that motivated the proposal.

The design choice is motivated by two competing desiderata. Early in training, large learning rates and rapid representation changes make a long averaging horizon risky because it can retain parameters from a different part of the trajectory. Late in training, smaller learning rates make recent iterates more compatible, and a longer horizon may reduce stochastic variation induced by mini-batches and noisy labels. A fixed EMA coefficient cannot express both preferences with one value. CEMA encodes them by tying the averaging coefficient to the same normalized cosine time used by the learning-rate schedule. This is a simple intervention, but its simplicity also limits what it can claim: it changes the readout, not the optimization objective.

4.1 Cosine-Coupled Coefficient

For a no-restart cosine learning-rate period, let $0 \leq \eta_{\min} < \eta_{\max}$ and define

$$\eta_t = \eta_{\min} + \frac{\eta_{\max} - \eta_{\min}}{2} \left(1 + \cos \frac{\pi t}{T} \right), \quad t \in \{0, \dots, T\}. \quad (5)$$

This learning-rate schedule is inherited from cosine annealing. The manuscript-defined EMA coefficient uses the same time variable but reverses the direction so that smoothing memory grows as the learning rate decays. For constants $0 \leq \beta_{\min} \leq \beta_{\max} < 1$,

$$\beta_t = \beta_{\max} - (\beta_{\max} - \beta_{\min}) \frac{\eta_t - \eta_{\min}}{\eta_{\max} - \eta_{\min}} = \beta_{\min} + \frac{\beta_{\max} - \beta_{\min}}{2} \left(1 - \cos \frac{\pi t}{T} \right). \quad (6)$$

The CEMA recurrence is

$$\theta_0^C = \theta_0, \quad \theta_t^C = \beta_t \theta_{t-1}^C + (1 - \beta_t) \theta_t, \quad t = 1, \dots, T. \quad (7)$$

The effective-horizon proxy $H_t = (1 - \beta_t)^{-1}$ is used only as a descriptive diagnostic when $\beta_t < 1$. In the executed configuration, $\beta_{\min} = 0.90$ and $\beta_{\max} = 0.999$, so the proxy ranges from 10 to 1000 updates. This range is bracketed by the fixed-EMA grid, which includes coefficients from 0.9 through 0.9995.

The coefficient schedule has two important exclusions. It is defined for a single no-restart period with $\eta_{\max} > \eta_{\min}$; a constant learning rate would require a different time variable or would reduce the comparison to fixed EMA. It is also coordinate-wise in parameter space, so it assumes the same architecture and parameter ordering across all iterates. These restrictions are satisfied by the executed trajectories because each readout is computed from one trained model path. They would not be satisfied by averaging checkpoints from independently permuted neural networks or by changing architectures across the trajectory.

The schedule can be interpreted as assigning a shorter memory to the high-learning-rate phase and a longer memory to the low-learning-rate phase. This interpretation is qualitative because $H_t = (1 - \beta_t)^{-1}$ is exact only for fixed coefficients. For time-varying coefficients, the true mass on an earlier iterate is the product term in (8). The experiment therefore uses the horizon proxy only to choose fair fixed-EMA comparators and to explain the schedule scale. It does not use the proxy as a theorem about effective sample size or as a calibrated bandwidth for loss smoothing.

4.2 Finite-Horizon Convexity

The formal contribution is limited to the algebraic behavior of the readout. It does not claim convex optimization, calibration optimality, or a guarantee for nonconvex ReLU networks. It states that, under compatible parameter vectors and coefficients in $[0, 1]$, CEMA remains inside the convex hull of the observed trajectory.

Theorem 1 (Convex representation of scheduled EMA). *Let $T \in \mathbb{N}_{\geq 1}$, $d \in \mathbb{N}_{\geq 1}$, $\theta_t \in \mathbb{R}^d$ for $t \in \{0, \dots, T\}$, and $\beta_t \in [0, 1]$ for $t = 1, \dots, T$. Define θ_t^C by (7). For every $t \in \{1, \dots, T\}$,*

$$\theta_t^C = w_{0,t}\theta_0 + \sum_{s=1}^t w_{s,t}\theta_s, \quad w_{0,t} = \prod_{r=1}^t \beta_r, \quad w_{s,t} = (1 - \beta_s) \prod_{r=s+1}^t \beta_r. \quad (8)$$

The coefficients satisfy $w_{s,t} \geq 0$ for all $s \in \{0, \dots, t\}$ and $\sum_{s=0}^t w_{s,t} = 1$.

Proof. The nonnegativity follows directly from $\beta_r \in [0, 1]$ and $1 - \beta_s > 0$. It remains to prove the representation and coefficient sum. For $t = 1$, (7) gives $\theta_1^C = \beta_1\theta_0 + (1 - \beta_1)\theta_1$, which matches (8), and the coefficients sum to one. Assume the representation holds at $t - 1$. Substituting it into (7) gives

$$\theta_t^C = \beta_t \left(w_{0,t-1}\theta_0 + \sum_{s=1}^{t-1} w_{s,t-1}\theta_s \right) + (1 - \beta_t)\theta_t.$$

Thus the coefficient on θ_0 is $\beta_t \prod_{r=1}^{t-1} \beta_r = \prod_{r=1}^t \beta_r$, the coefficient on θ_s for $1 \leq s < t$ is $(1 - \beta_s) \prod_{r=s+1}^t \beta_r$, and the coefficient on θ_t is $1 - \beta_t$. The coefficient sum is $\beta_t \sum_{s=0}^{t-1} w_{s,t-1} + (1 - \beta_t) = 1$ by the induction hypothesis. This completes the induction. $\square$

The theorem is useful for interpreting the method as a readout rather than as a new optimizer. It also makes the main caveat explicit. A convex combination of iterates can reduce trajectory noise, but it does not imply that the resulting classifier is calibrated, nor that it is flatter in a nonconvex loss landscape. The empirical sections therefore test calibration, NLL, accuracy, stability, path movement, and sharpness proxies separately.

5 Experimental Protocol

The protocol follows the falsifiable structure of the user-specified evaluation plan. It uses small deterministic datasets, clean held-out evaluation, paired seed comparisons, and a strong non-scheduled averaging baseline selected within each dataset. The design intentionally avoids moving goalposts after seeing the result. Post-hoc temperature scaling, secondary metrics, and mechanism diagnostics are included to interpret the result, but they do not replace the raw-ECE primary predicate. This distinction is essential because the final evidence is mixed: CEMA improves some probability-quality and stability summaries while failing the raw calibration criterion.

The protocol also keeps the operational context subordinate to the scientific contrast. The training program is CPU-feasible and uses small models, but the paper does not claim that CPU execution itself is the contribution. The relevant properties are repeatability, matched trajectories, clean held-out labels, and enough seeds to compare readouts within each dataset. Each evaluated probability summary is tied to a specific readout and evaluation layer. Raw metrics are computed from the unmodified readout probabilities. Temperature-scaled metrics fit a positive scalar on validation NLL and then evaluate the transformed logits on the test set. This separation prevents post-hoc calibration from being conflated with the readout intervention.

The experiment grid also prevents hidden tuning advantages. CEMA uses a fixed $\beta_{\min}$ and $\beta_{\max}$ across the reported datasets. Fixed EMA baselines cover a range of coefficients rather than a single straw-man value. Tail Polyak averages cover several terminal fractions. The strongest non-scheduled baseline is chosen after computing raw ECE within the same dataset, so the scheduled method is not favored by the comparator definition. This is a demanding design for CEMA, but it is appropriate because the hypothesis claims improvement over final-iterate SGD, fixed EMA, and Polyak averaging. A method that wins only against a weak subset of these baselines would not satisfy that claim.

The logistic reference has a different role from the MLP readouts. It checks whether the dataset and noise protocol produce sensible small-classifier behavior under a simpler model family, but it is not the primary target for the readout theorem because logistic training does not share the same nonconvex trajectory structure. The one-hidden-layer MLP provides a lower-capacity neural condition. The two-hidden-layer MLP is the primary condition because it is small, nonlinear, and still feasible for repeated seed sweeps. This hierarchy helps keep the result from being overgeneralized across architectures.

5.1 Evidence Units and Splits

The executed validation substrate contains six results-facing evidence units. Gaussian blobs is a three-class synthetic classification generator evaluated with train-label-only symmetric noise, clean validation and test labels, and 20 seeds per noise level. Two moons is a binary nonlinear synthetic substrate with the same split and noise discipline. Digits is the bundled ten-class handwritten digit dataset, vectorized and standardized under the same seed-specific splitting procedure. Breast cancer is the bundled binary

tabular diagnostic dataset. These four dataset units provide the primary evidence for calibration, NLL, accuracy, and reliability behavior. The symbolic schedule checks verify the CEMA coefficient form, coefficient sums, symmetric-noise matrices, and horizon bracketing. The metric-sensitivity analysis is a post-hoc analysis over raw and temperature-scaled probability summaries, ECE bin counts, NLL, and Brier score.

The four datasets match the assumptions of the hypothesis: finite supervised classification, small models, deterministic seeds, controlled training-label noise, and clean held-out evaluation. The synthetic generators do not claim external realism, and the bundled datasets are small; those limits are part of the result. The split sizes were stable across seeds: Gaussian blobs used 960 training, 320 validation, and 320 test examples; two moons used 960, 320, and 320; digits used 1078, 359, and 360; breast cancer used 341, 114, and 114. Realized corruption rates tracked the requested ρ values across seeds, and all held-out labels were preserved clean for the main evaluation.

The evidence units also play different roles in the claim structure. The four dataset units are primary because they directly instantiate the noisy-label classifier regime. The symbolic checks are confirmatory because they verify whether the schedule used in those runs has the finite algebraic properties claimed in the method. The metric-sensitivity analysis is diagnostic because it asks whether the raw-ECE decision is robust to common probability-quality summaries and post-hoc scaling. This role separation is important for negative results. Passed symbolic checks cannot rescue a failed calibration predicate, and a favorable temperature-scaled NLL result cannot replace the raw readout metric specified before evaluation.

The evidence units were selected to match the assumptions rather than to maximize benchmark prestige. Gaussian blobs and two moons are synthetic, so they provide control over class geometry and sample generation. Digits and breast cancer are bundled real datasets, so they add measured data without network-dependent acquisition. Each unit is small enough to support 20 seeds per noise level and multiple readouts. The low-relevance check did not flag any of the four primary units because they all instantiate finite supervised classification with train-label-only symmetric noise. The limitation is external validity, not relevance to the tested hypothesis.

5.2 Training Grid and Uncertainty Procedures

The experiment matrix evaluates a logistic reference, a one-hidden-layer ReLU MLP, and a two-hidden-layer ReLU MLP. The primary claim is attached to the two-hidden-layer MLP with 64 hidden units per layer because that architecture is expressive enough to show noisy-label training dynamics while remaining small enough for paired seed sweeps. The complete metrics table contains 17,280 rows: four datasets, four noise levels, 20 seeds, three architecture families, raw and temperature-scaled calibration layers, and multiple readouts. For the primary MLP readout comparison, each dataset contributes 60 matched rows over the three nonzero noise levels. The readout set includes final iterate, fixed EMAs with coefficients from 0.9 through 0.9995, uniform tail Polyak averages over several tail windows, and CEMA with $\beta_{\min} = 0.90$ and $\beta_{\max} = 0.999$.

The primary statistic compares CEMA with the strongest non-scheduled averaging baseline b^* selected within each dataset, as defined by (4). Confidence intervals for the displayed deltas are paired over matched seeds and noise settings. The Wilcoxon test is one-sided in the direction required for CEMA to have lower ECE. This pairing matters because calibration estimates can be noisy on small test sets, and unpaired comparisons would mix split variation with readout variation. Temperature scaling is fit on validation NLL and evaluated on test metrics as a sensitivity layer; because the scalar temperature is positive, it preserves the argmax prediction and separates probability rescaling from accuracy changes. The symbolic checks are separate from empirical uncertainty. They validate the algebraic schedule and noise-matrix properties used in (5)–(8), not the nonconvex training outcome.

The comparator grid is deliberately broader than the proposed method needs to look favorable. Fixed EMA coefficients range from short to long horizons, and tail Polyak averages cover several terminal windows. This grid means that CEMA must beat a simple readout that may already be well matched to the dataset. It also makes the metric sensitivity harder to overread: if CEMA wins NLL but loses ECE against b^* , the disagreement is not caused by comparing against an obviously weak baseline. The logistic reference and one-hidden-layer MLP provide additional context for model capacity, but the primary decision is attached to the two-hidden-layer MLP because it is the requested small neural classifier condition with enough flexibility to show noisy-label dynamics.

The training horizon and readout frequency were held fixed within each run family so that readout differences could be attributed to averaging rather than to different stopping points. Validation NLL was tracked during training for stability diagnostics, while final clean test metrics were computed for the readout comparison. The same raw logits support accuracy, NLL, Brier score, and ECE computations, which reduces the chance that metric disagreement is caused by mismatched evaluation data. When temperature scaling is applied, the validation set supplies the scalar temperature and the test set remains untouched until final evaluation. This separation mirrors standard calibration practice and prevents test-set tuning.

Uncertainty summaries are reported as paired differences rather than independent confidence intervals on each method. For a metric where lower is better, the table convention is comparator minus CEMA, so a positive NLL, ECE, or Brier delta favors CEMA. Accuracy uses the same subtraction convention, but the acceptance rule treats a small negative value as tolerable only up to one absolute percentage point. This sign convention is stated before results because it affects every figure and table

Dataset	b^*	Rel. ECE	ECE delta [95% CI]	NLL delta [95% CI]	Acc. delta	p	Success
Gaussian blobs	fixed EMA 0.999	-28.97%	-0.0489 [-0.0827, -0.0152]	0.6221 [0.5770, 0.6672]	0.4245	0.9950	no
Two moons	fixed EMA 0.995	-15.75%	-0.0243 [-0.0426, -0.0061]	0.0687 [0.0551, 0.0823]	0.0890	0.9555	no
Digits	fixed EMA 0.999	-134.1%	-0.1164 [-0.1378, -0.0949]	1.669 [1.627, 1.711]	0.6770	1.0000	no
Breast cancer	tail Polyak 5%	-16.48%	-0.0187 [-0.0256, -0.0118]	0.0068 [8.08e-04, 0.0128]	0.0202	1.0000	no

Table 1: Dataset-level success predicates for the primary MLP condition. ECE and NLL deltas are defined as strongest non-scheduled baseline minus CEMA, so positive values favor CEMA. The success column applies the full rule in (4); every dataset fails because raw ECE improvement is negative despite nonnegative accuracy deltas and positive NLL deltas.

interpretation. It also makes the rejection transparent: the accuracy guard is not the reason for failure, and the NLL signs are not the reason for success. The raw-ECE signs decide the primary claim.

6 Results

The results section is organized around the claim structure rather than around a favorable metric order. The first subsection closes the primary raw-ECE predicate and rejects the headline calibration-improvement claim. The second subsection explains why that rejection coexists with more favorable NLL and temperature-scaled signs. All results-facing evidence labels used below were executed locally and passed the final visual QA checks. Deferred ablations and descriptor-only lineage items are kept out of the results narrative and are discussed only as limitations.

The central comparison is intentionally unforgiving. For each dataset, the strongest non-scheduled averaging baseline is chosen by the same raw-ECE objective that CEMA must improve. The result therefore answers whether the schedule adds value beyond a well chosen fixed or uniform average, not whether it can beat an unsmoothed final iterate. This framing is important for interpreting mixed secondary outcomes. A method can be useful for validation loss or post-hoc probability scaling and still fail the primary readout claim. The following subsections keep those statements separate.

6.1 Primary Calibration Outcome

The primary result is shown in Figure 1 and Table 1. The setup tests exactly the acceptance rule in (4): two-hidden-layer MLP, label noise at 10%, 20%, and 40%, 20 matched seeds per noise level, raw top-label ECE on clean test labels, and comparison against the strongest non-scheduled averaging baseline. The strongest baseline is fixed EMA 0.999 for Gaussian blobs and digits, fixed EMA 0.995 for two moons, and tail Polyak over the last 5% for breast cancer. These choices are not selected to favor CEMA; they are selected by the strongest observed non-scheduled raw-ECE performance within the baseline grid.

This test probes the core hypothesis in its most relevant regime. The clean-label rows at $\rho = 0$ are useful controls, but the claim concerns label-noise levels where late-training stochasticity and memorization pressure should be more visible. Pooling the three nonzero noise levels within each dataset gives the primary predicate enough matched observations to test consistency across seeds while preserving the dataset as the unit of success. The decision threshold requires a relative raw-ECE gain, statistical reliability, and no material accuracy loss. Because CEMA fails at the first of these three requirements in every dataset, the rejection does not depend on a marginal p -value convention or on the accuracy guard.

The figure and table should be read together. The figure gives the visual direction of the primary deltas and shows whether accuracy or NLL tell a different story. The table records the exact strongest baseline, confidence intervals, p -value, and success decision. This pairing matters because a single panel can make a negative result look less precise than it is, while a table without a visual summary can hide the cross-dataset pattern. Here both representations agree on the central point: the scheduled EMA does not produce lower raw ECE than the strongest fixed or uniform averaging readout under the noisy-label primary condition.

The result also rules out two simple explanations. It is not a case where CEMA loses calibration by sacrificing accuracy, because mean accuracy differences are nonnegative in the displayed primary comparisons. It is not a case where CEMA has worse probability quality across all proper scores, because NLL deltas favor CEMA. The failure is more specific and therefore more informative. The scheduled readout changes probability assignments in a way that helps log loss but hurts binned raw confidence alignment. That specificity is why the manuscript reports the negative result as a calibration-stability failure rather than as a general optimization failure.

CEMA records 0 of 4 primary successes. Relative raw-ECE improvements are negative for all datasets: -29.0% on Gaussian blobs, -15.8% on two moons, -134.1% on digits, and -16.5% on breast cancer. The ECE delta column in Table 1 is negative in all cases, meaning CEMA has higher raw ECE than b^* . The Wilcoxon p -values are correspondingly far from the required one-sided $p < 0.05$ threshold. At the same time, NLL deltas are positive, so CEMA often assigns higher probability to the true

Primary Noisy-Label Outcome: CEMA Fails the Preregistered Calibration Criterion

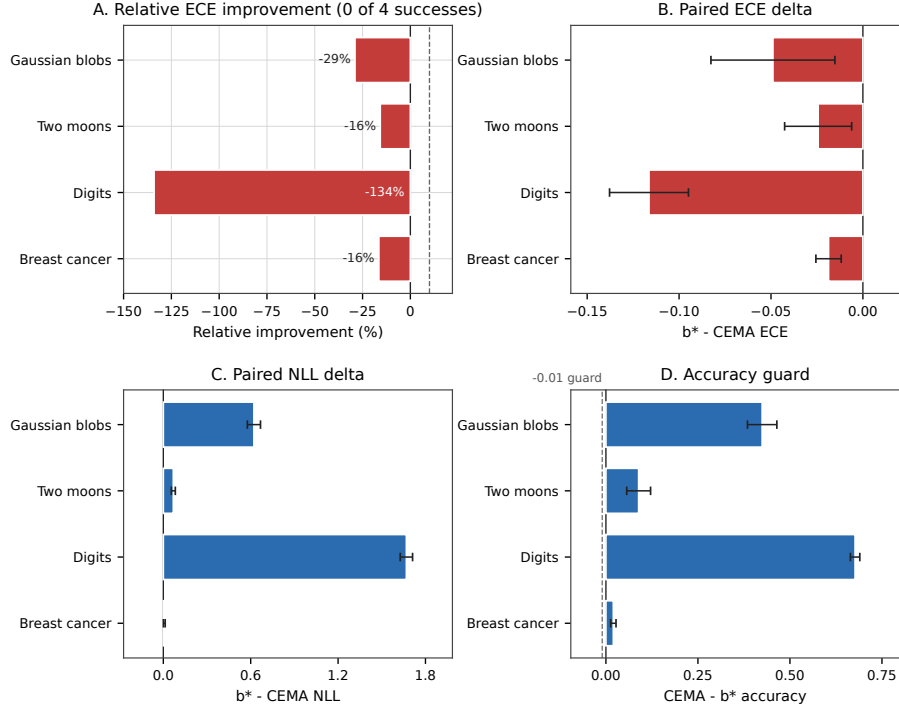


Figure 1: Primary two-hidden-layer MLP outcome across the four datasets. The panels summarize relative raw-ECE improvement, paired ECE and NLL deltas against the strongest non-scheduled averaging baseline, and the accuracy guard under nonzero label noise. The visual result is negative for the preregistered calibration claim: CEMA improves NLL and preserves or improves accuracy in several settings, but raw ECE is worse than the strongest averaging baseline on every dataset.

class even as its binned confidence-correctness gaps worsen. This is the central empirical contradiction: the scheduled filter can improve a proper scoring rule without satisfying the raw calibration predicate.

The largest raw-ECE reversal occurs on digits, where the relative value is more than a factor of two worse for CEMA under the primary convention. This dataset also has the largest class count among the four, which makes top-label confidence distributions more complex than in the binary tasks. The breast-cancer result is smaller in absolute terms but still fails the same direction. Gaussian blobs and two moons show that the issue is not confined to bundled real datasets. Across all four cases, the strongest fixed or tail average provides a better raw calibration readout than the scheduled one. The evidence therefore rejects the requested success criterion rather than merely weakening it.

The NLL column explains why the result needed multiple metrics. If the paper reported only NLL, CEMA would look favorable. If it reported only raw ECE, the method would look uniformly worse. The joint table shows that both statements can be true because they evaluate different properties of the probability vector. CEMA can assign more probability to the true label on average while concentrating predictions in confidence bins whose empirical correctness is less aligned. This pattern is precisely why the primary criterion was defined before results and why post-hoc metric selection would be misleading.

6.2 Reliability Curves and Metric Sensitivity

Figure 2 expands the scalar ECE result into representative reliability curves. The curves show why the ECE failure should not be dismissed as a table artifact. Raw confidence bins for CEMA deviate from the diagonal differently across datasets, and the strongest non-scheduled averaging baselines often produce bins with smaller accuracy-confidence gaps. Because small test sets can make individual bins noisy, these curves are interpreted together with the paired deltas in Table 1, not as standalone visual proof.

The metric-sensitivity analysis clarifies the scope of the negative result. Across each dataset, raw metrics show two positive signs and four negative signs when comparing comparator metric minus CEMA metric, while temperature-scaled metrics show six positive signs and no negative signs. In practical terms, CEMA looks more favorable after validation-fitted temperature scaling, but the primary claim concerns raw readouts. This split is consistent with calibration literature warning that scalar calibration metrics and post-hoc transformations can decouple from each other. It also prevents a stronger but unsupported

Reliability Curves: Representative Seed-0, $\rho=0.2$, mlp_2x64

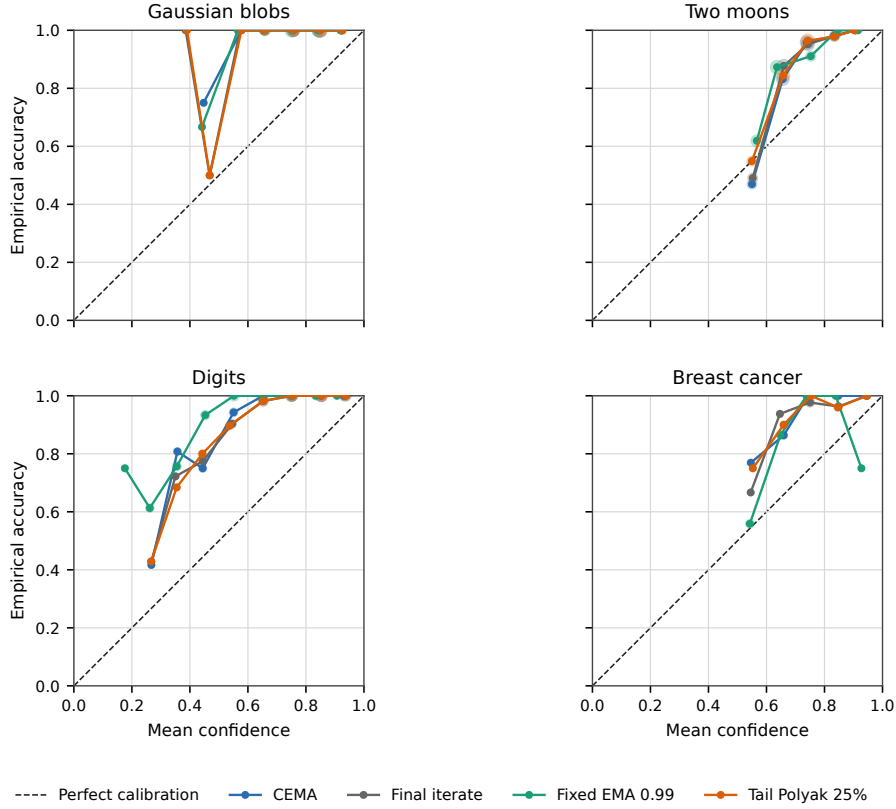


Figure 2: Representative reliability curves for CEMA, final iterate, fixed EMA, and tail Polyak readouts. Each panel places empirical accuracy against mean confidence for a dataset-specific representative slice, with the diagonal indicating perfect top-label calibration. The curves reveal bin-level structure behind the failed ECE predicate and show that the CEMA loss advantage does not translate into uniformly better raw confidence alignment.

conclusion: the study does not show that CEMA is useless for probability quality, only that it does not satisfy the raw-ECE calibration-stability criterion against strong averaging baselines.

The reliability curves add qualitative support for the same interpretation. The diagonal reference is a calibration ideal, but each panel is still a finite-sample estimate with bin occupancy limits. The curves are therefore not used to decide the hypothesis by eye. They show how the scalar ECE failure manifests as confidence-region discrepancies and help explain why a readout with favorable NLL can still have worse raw bin alignment. This is especially useful for small datasets, where a single scalar can obscure whether disagreement comes from underconfidence, overconfidence, or changes in where predictions fall along the confidence axis.

Temperature scaling should be read as a sensitivity analysis, not as a rescue. A positive scalar temperature preserves predicted classes but changes confidence magnitudes, so it is a post-hoc probability transformation. The fact that CEMA becomes more favorable after this transformation suggests that its logits may contain useful ranking or margin information for proper-score calibration. It does not show that the raw scheduled training readout is calibrated. For applications that can afford post-hoc scaling, CEMA may remain a candidate; for the claimed training-time raw calibration stability, the evidence is negative.

7 Mechanism Diagnostics

The mechanism analysis asks a narrower question than the primary calibration test. If the scheduled coefficient behaves like a useful low-pass readout, it should at least stabilize some late-training trajectory quantities. The evidence supports that limited statement for validation-NLL variance, but not for all smoothing proxies. Separating these diagnostics avoids turning one favorable stability number into a broad mechanistic explanation. It also clarifies why a valid convex filter can still fail to deliver lower raw ECE when calibration bins respond differently from validation loss.

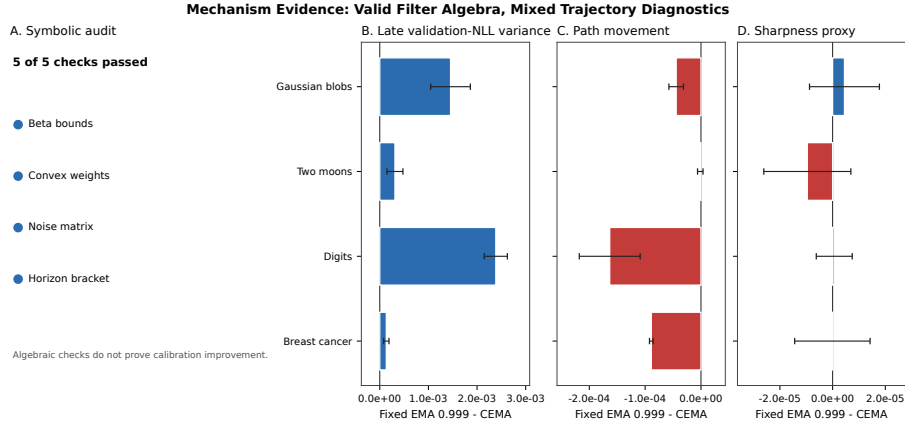


Figure 3: Mechanism diagnostics combine algebraic schedule checks with empirical trajectory summaries. The symbolic checks validate the finite filter regime, while the quantitative panels compare CEMA with a fixed EMA coefficient of 0.999 for late validation-NLL variance, path movement, and a random-direction sharpness proxy. The supported mechanism is narrow: CEMA reduces late validation-loss variance, but path movement and sharpness do not show a uniform smoothing advantage.

Dataset	Late validation-NLL variance	Path movement	Sharpness proxy
Gaussian blobs	0.0015 [0.0010, 0.0019]	-4.44e-05 [-5.74e-05, -3.13e-05]	4.49e-06 [-8.77e-06, 1.78e-05]
Two moons	3.12e-04 [1.48e-04, 4.76e-04]	-1.25e-06 [-6.17e-06, 3.68e-06]	-9.60e-06 [-2.61e-05, 6.93e-06]
Digits	0.0024 [0.0021, 0.0026]	-1.63e-04 [-2.18e-04, -1.09e-04]	6.20e-07 [-6.20e-06, 7.44e-06]
Breast cancer	1.34e-04 [7.93e-05, 1.89e-04]	-8.89e-05 [-9.22e-05, -8.57e-05]	-9.55e-08 [-1.44e-05, 1.42e-05]

Table 2: Mechanism deltas behind the diagnostic figure. Values are fixed EMA 0.999 minus CEMA with paired confidence intervals, so positive late-variance values indicate lower validation-NLL variance for CEMA. Negative path-movement values mean CEMA moved more than the fixed EMA comparator, and sharpness intervals mostly include zero.

The mechanism diagnostics are not designed to prove why every dataset behaves as it does. They are designed to test whether the most natural smoothing story leaves measurable traces. A low-pass interpretation predicts less high-frequency variation in a late loss trace. A flatness interpretation predicts smaller loss changes under local parameter perturbations. A path-stability interpretation predicts less readout movement. These are related ideas, but they are not the same property. Reporting them together avoids a common ambiguity in discussions of averaging, where smoothness, flatness, variance reduction, and calibration are treated as if they were interchangeable.

7.1 Variance, Path Movement, and Sharpness

The low-pass intuition predicts that CEMA should damp late stochastic variation as β_t increases. Figure 3 and Table 2 evaluate this claim with three diagnostics. Late validation-NLL variance is the variance of validation NLL over the final quarter of training, compared against fixed EMA 0.999. Path movement measures the amount of readout movement along the retained trajectory. The sharpness proxy perturbs parameters along random directions and measures local loss sensitivity; it is a proxy only, not a Hessian calculation or certified flatness result.

Late validation-NLL variance favors CEMA in all four datasets. The deltas are positive and the intervals exclude zero, with the largest mean deltas on digits and Gaussian blobs. This supports the idea that increasing the averaging horizon late in training can stabilize validation loss. The other diagnostics are weaker. Path movement deltas are negative for Gaussian blobs, digits, and breast cancer, indicating more movement for CEMA than for the fixed EMA comparator, and two moons is close to zero. Sharpness-proxy intervals include zero for every dataset. The mechanism conclusion is therefore not that CEMA universally smooths the parameter path. It is that the schedule is algebraically valid and empirically stabilizes late validation NLL in this grid, while the broader sharpness-suppression story remains unsupported.

The variance result is still meaningful because it uses the same validation loss tracked during training and compares against a fixed EMA with a long horizon. It says that the time-varying coefficient can damp late validation-loss fluctuations in the executed grid. However, the path-movement and sharpness proxies caution against a stronger geometric story. A readout can have lower temporal loss variance while moving more in parameter space if the scheduled weights place different mass on late iterates than

the fixed comparator. A random-direction sharpness proxy can also miss structured directions, and the intervals here do not support a uniform advantage. The supported mechanism is therefore loss-trace stabilization, not global flatness.

7.2 What the Algebra Explains

The symbolic schedule checks cover five formal obligations: the closed form in (6), endpoint and monotonicity conditions for β_t , coefficient sums in (8), row-stochasticity of the symmetric noise matrix in (1), and horizon bracketing by the fixed-EMA grid. All five checks passed under their stated regimes. This matters because the empirical failure is not due to an invalid coefficient schedule or a mismatch between the scheduled horizon and available fixed baselines. It also matters because it prevents overinterpretation. A convex parameter filter can still be a poor calibration readout if the calibration landscape is not aligned with the smoothed trajectory.

The diagnostics show the difference between algebraic support and empirical calibration support. Equation (8) proves that each scheduled readout is a convex combination of compatible iterates. It does not prove that the softmax probabilities at that readout are closer to the calibration diagonal, nor that the binned ECE estimator decreases. Equation (4) is the empirical bridge, and it fails. The proper interpretation is mixed: CEMA has a valid filter form and improves one stability diagnostic, but the central raw-ECE claim is contradicted by the primary table and reliability curves.

This separation also guards against a tempting post-hoc explanation. It would be easy to say that CEMA is a low-pass filter and therefore should reduce calibration error. The evidence does not support that implication. The filter algebra concerns weights on parameter vectors. Calibration concerns probabilities induced by nonlinear logits on held-out examples and then summarized by finite bins. The map from parameter filtering to probability calibration can be irregular in a nonconvex ReLU model, especially under label noise. The observed results fit that caution: the filter is valid, and one late-loss statistic improves, but raw ECE worsens against the best averaging comparator.

8 Discussion

The discussion interprets the rejection as a boundary on the method, not as a failure of the experimental program. The strongest averaging baseline matters because it converts a broad “does averaging help” question into the sharper “does this schedule help beyond good averaging” question. The answer in this regime is no for raw ECE. The answer is more nuanced for NLL, temperature-scaled summaries, and late-loss variance. That nuance is useful only if the primary failure remains visible throughout the interpretation.

A negative result with a strong comparator is more informative than a positive result against a weak one. The experiment shows that a plausible schedule can fail exactly where it was expected to help, even while secondary scores improve. This has practical value for calibration work because many interventions are evaluated under a single metric or a final-iterate baseline. Here, the schedule’s benefit is bounded by fixed and uniform averaging choices that are simpler to implement and easier to tune. The readout should therefore be treated as a candidate for loss stabilization, not as a default calibration remedy.

8.1 Interpreting the Negative Calibration Evidence

The most important design choice in this study is comparing CEMA with the best non-scheduled averaging baseline, not merely with the final iterate. Final-iterate comparisons can exaggerate the value of any averaging method because final weights are a weak baseline in noisy training. In this experiment, fixed EMA and tail Polyak averaging are strong enough that CEMA’s raw calibration advantage disappears. This is not a minor threshold miss. The relative raw-ECE improvement is negative in every primary dataset, and the acceptance rule requires three positive dataset-level successes. The headline hypothesis is therefore rejected for the tested regime.

At the same time, the result is not evidence that scheduled averaging is dominated for every objective. CEMA improves NLL against b^* in the primary table, and temperature-scaled metric sensitivity is favorable. Those outcomes suggest that the readout can improve probability mass assigned to true labels while producing confidence bins that are less aligned under raw top-label ECE. This pattern agrees with prior warnings that calibration measures can decouple and that ECE is sensitive to binning and sample size. A practitioner interested in NLL or post-hoc calibrated probabilities might still investigate CEMA, but a practitioner seeking raw calibration stability should prefer simpler fixed or uniform averaging baselines unless new evidence changes the boundary.

8.2 Design Implications

The failure boundary carries three practical implications. First, horizon matching matters. If scheduled EMA is compared only with a final iterate or a poorly chosen fixed EMA, it may appear stronger than it is. The fixed grid used here brackets the scheduled

horizon proxy, making the comparator fairer. Second, raw and temperature-scaled calibration should be reported separately. Temperature scaling preserves argmax predictions, but it changes the probability scale and can reverse metric signs. Third, mechanism diagnostics should not collapse several properties into one word such as smoothness. Lower late validation-NLL variance, lower path movement, and lower random-direction sharpness are distinct claims; this study supports only the first of those empirical properties.

These implications affect how one would use the method in practice. If the goal is a single cheap readout for a small noisy classifier, the evidence favors checking a fixed EMA or tail average before adopting a schedule-coupled rule. If the goal is stable validation loss during late training, CEMA is more defensible because that diagnostic is supported in the executed grid. If the goal is calibrated probabilities after a validation transformation, the temperature-scaled sensitivity suggests a possible downstream use, but that use requires a separate comparison with standard calibrators. The correct choice therefore depends on the decision metric. CEMA should not be presented as a raw calibration default simply because it has a reasonable filter form.

The broader lesson is methodological. Training-time readout interventions can be useful even when they fail a primary calibration claim, but the paper must preserve the rejected claim rather than retuning the acceptance rule after seeing results. The negative result is scientifically useful because it narrows the space of plausible explanations. The problem was not invalid algebra, missing clean held-out labels, or an absent fixed-EMA comparator. The problem was that the proposed schedule did not produce lower raw ECE against the strongest averaging readout under paired noisy-label evaluation.

The result also illustrates why calibration studies should report comparator provenance. A final iterate is an optimizer endpoint; a fixed EMA is a geometric filter; a tail Polyak average is a uniform filter; a temperature-scaled model is a post-hoc probability transformation. These objects can all improve some metric, but they answer different design questions. Mixing them without stating the object changed makes conclusions hard to reproduce. The present design keeps the objects separated and shows that the scheduled parameter filter is not the best raw-ECE readout in this finite grid.

9 Limitations and Future Work

The study is intentionally narrow. It covers small classifiers, deterministic seed sweeps, scikit-learn generators or bundled datasets, symmetric training-label noise, clean held-out evaluation, and parameter readouts from identical optimizer trajectories. It does not evaluate large image models, transformers, instance-dependent label noise, corrupted validation or test labels, distribution shift, semi-supervised objectives, robust loss functions, or multi-model ensembles. These exclusions are not flaws in the executed test; they define where the rejected claim applies. The four dataset conditions fit the hypothesis assumptions well because all are finite-class small-classifier settings with controlled training-label noise and no low-relevance flags in the validation fit report. The synthetic datasets are lower in external realism, while digits and breast cancer add small real-data checks but do not represent large-scale modern networks.

The limitation that matters most for interpretation is not a lack of executed data within the requested regime; it is the regime itself. Small bundled or synthetic datasets make seed sweeps feasible and keep the evidence reproducible, but they cannot represent the calibration behavior of larger architectures, richer image corpora, or shifting deployment distributions. Instance-dependent label noise could interact with averaging differently from symmetric noise because corrupted examples would no longer be exchangeable across classes. Corrupted validation labels would also change the temperature-scaling and early-stopping interpretation. The manuscript’s conclusion should therefore be read as a rejection for the stated small-classifier assumptions, not as a universal theorem about scheduled EMA.

The main missing ablation is a full interaction matrix over schedule shape, learning-rate schedule, model width, and training length. The executed run includes two MLP widths, a logistic reference, a horizon-bracketed fixed EMA grid, multiple tail averages, metric sensitivity, and symbolic horizon checks. It does not close every possible schedule-shape or horizon-computation question. Future experiments should test alternative β_t curves, constant-learning-rate controls with scheduled readouts, warm-restart variants with period-local time, longer training horizons, wider networks, and robust noisy-label objectives as adjacent rather than primary baselines. A stronger positive claim would also need classwise calibration, adaptive-bin estimates, larger held-out sets, and uncertainty intervals that remain favorable after choosing the strongest baseline.

The unexecuted ablation matrix has a concrete impact on claims. It prevents the paper from saying that no schedule-coupled EMA can improve raw calibration; only the tested cosine coupling and horizon range are rejected. It also prevents attributing the result solely to the learning-rate schedule, because a constant-learning-rate control with the same readout family was not part of the final evidence. These gaps are nonblocking for the current negative conclusion because the requested success predicate was evaluated directly. They are blocking for broader statements about schedule design, optimizer interaction, and architecture scaling. Closing them would require new paired experiments rather than reinterpretation of the current tables.

A second limitation is the granularity of calibration measurement. The manuscript reports top-label ECE, several bin variants, Brier score, NLL, and reliability curves, but it does not establish class-conditional calibration or full distribution calibration. This is appropriate for the requested primary metric, yet it leaves open the possibility that CEMA affects minority classes or low-confidence regions differently from the aggregate top-label statistic. Larger held-out sets would be needed to make such

classwise estimates reliable, especially for digits and breast cancer where some bins or classes can have limited counts. Future work should predefine which calibration notion matters for the application before increasing metric breadth.

10 Conclusion

Cosine-scheduled parameter EMA is an appealing readout because it couples a growing smoothing horizon to a decaying cosine learning-rate schedule. In the executed small noisy-classifier study, that appeal does not translate into the preregistered raw calibration gain. Against the strongest non-scheduled averaging baseline, the primary two-hidden-layer MLP condition records 0 of 4 dataset-level successes on raw top-label ECE. The method improves NLL and late validation-loss stability in several comparisons, and its finite-horizon convex-filter algebra is valid, but those facts do not imply raw calibration stability. The supported conclusion is therefore bounded: CEMA is a valid and sometimes loss-stabilizing readout, not a reliable raw-ECE calibration improvement for the tested small noisy classifiers.

The negative conclusion is useful because it preserves a distinction that is easy to lose in optimization papers. A readout can be mathematically well formed, computationally cheap, and beneficial for a proper scoring rule while still failing the calibration metric that motivated it. The correct response is not to rename the objective after seeing the results, but to state which claims survived. In this study, the surviving claims are finite convex-filter validity, horizon-bracketed comparator fairness, and reduced late validation-NLL variance. The rejected claim is broad raw calibration improvement under controlled symmetric label noise. That boundary gives future experiments a concrete target: either find a schedule and regime that clears the same raw-ECE predicate, or change the hypothesis to a different probability-quality objective and evaluate it as such.

References

- [1] Xiaozhe Gu, Zixun Zhang, Yuncheng Jiang, Tao Luo, Ruimao Zhang, Shuguang Cui, and Zhen Li. Hierarchical weight averaging for deep neural networks, 2023. URL <https://arxiv.org/abs/2304.11519>. arXiv preprint.
- [2] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017. URL <https://arxiv.org/abs/1706.04599>. Proceedings of ICML 2017.
- [3] Hao Guo, Jiyong Jin, and Bin Liu. Stochastic weight averaging revisited, 2022. URL <https://arxiv.org/abs/2201.00519>. arXiv preprint.
- [4] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels, 2018. URL <https://arxiv.org/abs/1804.06872>. NeurIPS 2018.
- [5] Michael Heck, Christian Geishauser, Nurul Lubis, Carel van Niekerk, Shutong Feng, Hsien-Chin Lin, Benjamin Matthias Ruppik, Renato Vukovic, and Milica Gasic. Learning from noisy labels via self-taught on-the-fly meta loss rescaling, 2024. URL <https://arxiv.org/abs/2412.12955>. arXiv preprint.
- [6] Pavel Izmailov, Dmitrii Podoprikin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization, 2018. URL <https://arxiv.org/abs/1803.05407>. UAI 2018.
- [7] Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration, 2019. URL <https://arxiv.org/abs/1909.10155>. NeurIPS 2019.
- [8] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2016. URL <https://arxiv.org/abs/1612.01474>. NeurIPS 2017.
- [9] Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning, 2020. URL <https://arxiv.org/abs/2002.07394>. ICLR 2020.
- [10] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2016. URL <https://arxiv.org/abs/1608.03983>. ICLR 2017.
- [11] Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning, 2019. URL <https://arxiv.org/abs/1902.02476>. NeurIPS 2019.
- [12] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks, 2021. URL <https://arxiv.org/abs/2106.07998>. NeurIPS 2021.
- [13] Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits, 2024. URL <https://arxiv.org/abs/2411.18704>. arXiv preprint.
- [14] Moseli Mots’oehli and Kyungim Baek. Deep active learning in the presence of label noise: A survey, 2023. URL <https://arxiv.org/abs/2302.11075>. arXiv preprint.
- [15] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning, 2015. URL <https://pmc.ncbi.nlm.nih.gov/articles/PMC4410090/>. AAAI 2015.
- [16] Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep K. Ravikumar, and Ambuj Tewari. Learning with noisy labels, 2013. URL <https://proceedings.neurips.cc/paper/5073-learning-with-noisy-labels>. NeurIPS 2013.
- [17] Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. Measuring calibration in deep learning, 2019. URL <https://arxiv.org/abs/1904.01685>. CVPR Workshops 2019.
- [18] Emre Onal, Klemens Floge, Emma Caldwell, Arsen Sheverdin, and Vincent Fortuin. Gaussian stochastic weight averaging for bayesian low-rank adaptation of large language models, 2024. URL <https://arxiv.org/abs/2405.03425>. arXiv preprint.
- [19] Koffi Ismael Ouattara, Ioannis Krontiris, Theo Dimitrakos, and Frank Kargl. Quantifying calibration error in neural networks through evidence-based theory, 2024. URL <https://arxiv.org/abs/2411.00265>. arXiv preprint.

- [20] Giorgio Patrini, Alessandro Rozza, Aditya Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: a loss correction approach, 2016. URL <https://arxiv.org/abs/1609.03683>. CVPR 2017.
- [21] Jonathan Patsenker, Henry Li, and Yuval Kluger. Exponential weight averaging as damped harmonic motion, 2023. URL <https://arxiv.org/abs/2310.13854>. arXiv preprint.
- [22] Maja Pavlovic. Understanding model calibration – a gentle introduction and visual exploration of calibration and the expected calibration error (ece), 2025. URL <https://arxiv.org/abs/2501.19047>. arXiv preprint.
- [23] Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging, 1992. URL <https://doi.org/10.1137/0330046>. SIAM Journal on Control and Optimization.
- [24] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping, 2014. URL <https://arxiv.org/abs/1412.6596>. ICLR Workshop 2015.
- [25] scikit-learn developers. `sklearn.metrics.brier_score_loss`, 2026. URL https://scikit-learn.org/stable/modules/generated/sklearn.metrics.brier_score_loss.html. scikit-learn stable API documentation.
- [26] scikit-learn developers. `sklearn.calibration.calibration_curve`, 2026. URL https://scikit-learn.org/stable/modules/generated/sklearn.calibration.calibration_curve.html. scikit-learn stable API documentation.
- [27] scikit-learn developers. `sklearn.datasets.load_breast_cancer`, 2026. URL https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html. scikit-learn stable API documentation.
- [28] scikit-learn developers. `sklearn.datasets.load_digits`, 2026. URL https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html. scikit-learn stable API documentation.
- [29] scikit-learn developers. `sklearn.metrics.log_loss`, 2026. URL https://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html. scikit-learn stable API documentation.
- [30] scikit-learn developers. `sklearn.datasets.make_blobs`, 2026. URL https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_blobs.html. scikit-learn stable API documentation.
- [31] scikit-learn developers. `sklearn.datasets.make_moons`, 2026. URL https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_moons.html. scikit-learn stable API documentation.
- [32] scikit-learn developers. `sklearn.model_selection.train_test_split`, 2026. URL https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html. scikit-learn stable API documentation.
- [33] Leslie N. Smith. Cyclical learning rates for training neural networks, 2015. URL <https://arxiv.org/abs/1506.01186>. IEEE WACV 2017.
- [34] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, 2017. URL <https://arxiv.org/abs/1703.01780>. NeurIPS 2017.
- [35] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B. Schon. Evaluating model calibration in classification, 2019. URL <https://arxiv.org/abs/1902.06977>. AISTATS 2019.
- [36] Cheng Wang. Calibration in deep learning: A survey of the state-of-the-art, 2023. URL <https://arxiv.org/abs/2308.01222>. arXiv preprint.
- [37] Dominik Werner Wolf, Prasannavenkatesh Balaji, Alexander Braun, and Markus Ulrich. Decoupling of neural network calibration measures, 2024. URL <https://arxiv.org/abs/2406.02411>. arXiv preprint.
- [38] Wei Xu. Towards optimal one pass large scale learning with averaged stochastic gradient descent, 2011. URL <https://arxiv.org/abs/1107.2490>. arXiv preprint.
- [39] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, 2016. URL <https://arxiv.org/abs/1611.03530>. ICLR 2017.

- [40] Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels, 2018. URL <https://arxiv.org/abs/1805.07836>. NeurIPS 2018.
- [41] Yuxiang Zheng, Zhongyi Han, and Yilong Yin. Echoalign: Bridging generative and discriminative learning under noisy labels, 2024. URL <https://arxiv.org/abs/2405.12969>. arXiv preprint.
- [42] Chen-Chen Zong, Ye-Wen Wang, Ming-Kun Xie, and Sheng-Jun Huang. Dirichlet-based prediction calibration for learning with noisy labels, 2024. URL <https://arxiv.org/abs/2401.07062>. AAAI 2024.

A Extended Formal Details

The appendix expands formal and reproducibility details that would interrupt the main narrative. These details are not a separate claim of optimality. They record the validity regimes for the schedule equations, the noise model, and the empirical summaries so that the negative result can be reproduced without guessing which assumptions were active. The formal material is intentionally limited to finite algebraic properties; it does not attempt to prove calibration improvement for nonconvex neural training, broader model classes, or untested deployment settings.

A.1 Schedule Provenance and Applicability

The cosine learning-rate expression in (5) is inherited from cosine annealing, while the fixed EMA recurrence underlying (7) is inherited from weight-averaging practice. The manuscript-specific part is the inverse coupling in (6). It maps larger learning rates to shorter averaging memory and smaller learning rates to longer memory over a single no-restart period. The formula requires $\eta_{\max} > \eta_{\min}$ because the affine coupling uses the normalized learning-rate position. If the learning rate is constant, the coupling is undefined as written and should be replaced by a fixed EMA or by a separate schedule variable. If warm restarts are used, the update index t must be interpreted within a period or the schedule must be redefined across periods.

The symbolic checks validated the closed form, endpoints, monotonicity, and coefficient sums under finite discrete time. For $\beta_{\min} = 0.90$ and $\beta_{\max} = 0.999$, the effective-horizon proxy ranges from 10 to 1000 updates. This proxy is not an exact memory length for a time-varying EMA, but it is a useful scale for comparing with fixed EMAs. The fixed grid used in the experiment includes coefficients that bracket this range, which reduces the risk that CEMA is winning or losing only because fixed baselines were chosen on the wrong time scale. The formal applicability is still narrow: parameter vectors must share dimension and coordinate meaning, optimizer state is not averaged, and the recurrence says nothing about nonconvex calibration objectives.

The coefficient bounds also clarify why the theorem uses convex combinations. If any β_t were negative or exceeded one, the unrolled representation could place negative or amplifying weights on past iterates. Such a readout might still be definable algebraically, but it would no longer support the filter interpretation used in this paper. The chosen range avoids that complication. The increasing schedule gives recent late-training iterates substantial aggregate influence while retaining earlier compatible states through the recursive product. The result is still a parameter-space average, not an average of predictions. Prediction averaging would produce a different model family and a different calibration target because softmax is nonlinear.

A.2 Noise Matrix and Metric Domains

For $K \geq 2$ and $\rho \in [0, 1]$, the symmetric noise matrix in (1) has nonnegative entries and each row sums to one. The diagonal entry is $1 - \rho$, and the sum of off-diagonal entries is $(K - 1)\rho / (K - 1) = \rho$. Thus every row sums to $1 - \rho + \rho = 1$. The executed symbolic check evaluated this property for the planned class counts $K \in \{2, 3, 10\}$ and noise rates $\rho \in \{0, 0.1, 0.2, 0.4\}$. These are the only class counts used in the reported datasets: two for two moons and breast cancer, three for Gaussian blobs, and ten for digits.

The metric definitions also have domains. Equation (2) requires a predicted probability for the true class and clean labels from the evaluation set. Equation (3) requires confidence bins over $[0, 1]$ and is top-label rather than full multiclass calibration. Brier score is a proper-score complement whose binary and multiclass conventions must be kept consistent. Reliability curves are visual summaries of bin behavior and should not be used without paired scalar uncertainty in small samples. Temperature scaling fits a positive scalar on validation NLL, preserves argmax predictions, and changes confidence magnitudes. These domain statements explain why the manuscript reports raw and temperature-scaled layers separately and why the raw-ECE failure is not overwritten by favorable temperature-scaled signs.

The clean held-out requirement is part of the metric domain. If validation or test labels were corrupted by the same symmetric mechanism as training labels, NLL and ECE would evaluate agreement with noisy labels rather than with the underlying clean classification task. The experiment avoids that ambiguity by applying corruption after splitting and only to the training labels. This choice makes the held-out metrics appropriate for calibration claims but also makes the problem different from settings where deployment labels are themselves noisy or delayed. The manuscript’s probability-quality conclusions therefore attach to clean evaluation labels and should not be transferred unchanged to weak-label test sets.

B Dataset Fit and Evaluation Protocol

This appendix section records why the executed evidence units match the hypothesis and where they remain limited. The fit assessment is part of the interpretation because a weak dataset fit could make a mixed result look like a method failure. Here the fit is strong for the stated small-classifier regime: finite classes, deterministic seed sweeps, train-label-only symmetric noise, and clean validation and test labels. The external-validity boundary remains narrow because the datasets are small and several are synthetic or bundled educational benchmarks.

B.1 Dataset-Level Fit to the Hypothesis

Each dataset condition was relevant to the stated hypothesis. Gaussian blobs provides a controlled three-class synthetic problem with standardized 20-dimensional features, balanced classes, and enough samples to make paired seed sweeps stable. Two moons supplies a nonlinear binary synthetic problem whose decision boundary is not linearly separable. Digits supplies a small ten-class image-vector dataset bundled with scikit-learn. Breast cancer supplies a small binary tabular dataset with real diagnostic features. The common protocol is more important than the individual dataset identities: all features are standardized using training statistics, labels are split before corruption, validation and test labels remain clean, and every readout is evaluated on the same seed-specific optimizer trajectory.

The dataset mix tests several aspects of the hypothesis without claiming broad benchmark coverage. Gaussian blobs is useful because the class structure is controlled and the input dimension is moderate. Two moons tests nonlinear geometry in a low-dimensional binary setting. Digits adds a multiclass real-image feature space where calibration bins can be affected by class count and confidence concentration. Breast cancer adds a small real tabular setting with a different sample size and feature distribution. A method that passed the primary predicate on at least three of these four datasets would still require external validation, but such a result would have supported the requested small-classifier claim. The observed 0 of 4 result is therefore not a failure of dataset fit; it is a failure within the selected fit regime.

The executed fit reports did not flag low-relevance conditions. That does not mean external validity is broad. Gaussian blobs and two moons are synthetic generators, digits is low-resolution and small, and breast cancer is a small tabular medical dataset. These datasets match the requested small-classifier scope, not a claim about modern large neural networks. This distinction matters because calibration behavior is known to depend on model and data regime. A positive result here would have justified a bounded small-model claim; the negative result rejects that bounded claim but does not settle behavior for high-capacity models, distribution shift, or instance-dependent noise.

B.2 Seed Pairing and Acceptance Accounting

The primary accounting pools the nonzero noise levels for each dataset, giving 60 matched rows per dataset in the two-hidden-layer MLP condition. Matching is by seed, architecture, dataset, noise level, and readout family. This design reduces variance from train-validation-test splits and label corruption, because each readout is compared on the same realized training trajectory. The strongest non-scheduled averaging baseline b^* is selected within each dataset by raw ECE among fixed EMA and tail Polyak candidates, then CEMA is tested against that baseline. The final iterate is included in figures and reliability curves but does not define b^* because the central question is whether schedule coupling improves on strong averaging, not merely whether averaging helps over no averaging.

The acceptance rule has three parts. The relative raw-ECE improvement must be at least 10%, the one-sided paired Wilcoxon test must satisfy $p < 0.05$, and the mean accuracy difference must be no less than -0.01 . A dataset with better NLL but worse ECE fails. A dataset with better ECE but a large accuracy loss also fails. This strictness is why the conclusion is clear. CEMA's NLL deltas are positive in the primary table, and accuracy deltas are nonnegative, but ECE deltas are negative everywhere. The failure therefore comes from calibration, not from the accuracy guard or an inability to train the model.

Pooling the three nonzero noise levels is a compromise between dataset-level interpretability and seed-level power. Treating every dataset-noise pair as a separate success unit would inflate the number of apparent tests and make the conclusion harder to summarize. Treating all datasets as one aggregate would hide whether the effect is concentrated in a single substrate. The chosen rule makes the dataset the success unit and uses matched seed-noise rows to estimate each unit's direction. This is why the result is stated as 0 of 4 dataset-level successes rather than as a single pooled average over all rows. It preserves the cross-dataset requirement in the original hypothesis.

C Additional Metric Sensitivity

The metric-sensitivity appendix explains the most important secondary pattern in the evidence. Raw calibration and temperature-scaled calibration answer different questions because temperature scaling changes confidence magnitudes after fitting a validation scalar. The primary claim intentionally uses raw readouts, so the favorable temperature-scaled signs cannot overturn the primary failure. They do, however, show that CEMA's behavior is not uniformly poor across probability metrics and suggest more targeted follow-up questions about post-hoc calibrated deployment settings and metric selection for each dataset examined.

C.1 Raw and Temperature-Scaled Sign Structure

Figure 4 and Table 3 explain why the main result should be described as a raw calibration failure rather than a universal probability-quality failure. In the raw layer, ECE variants tend to favor the strongest comparator, while NLL and Brier signs favor CEMA. In the temperature-scaled layer, all six displayed signs per dataset favor CEMA. This pattern is compatible with

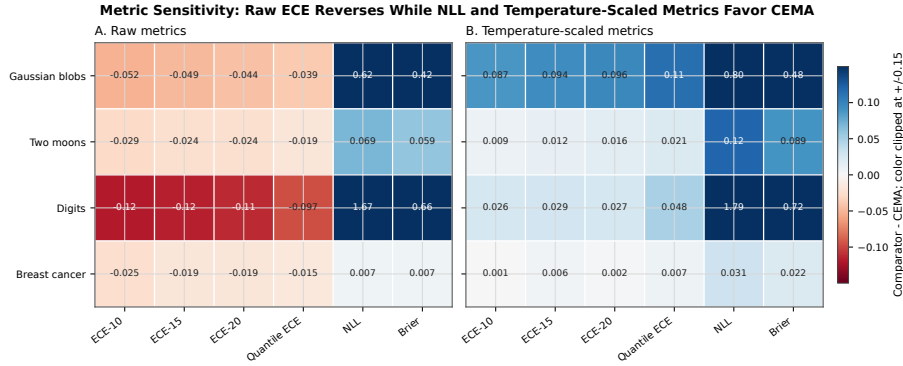


Figure 4: Metric sensitivity matrix for raw and temperature-scaled probability summaries. The sign convention is strongest comparator metric minus CEMA metric, so positive signs favor CEMA when lower metric values are better. The matrix shows that temperature scaling makes CEMA look consistently favorable, while raw ECE variants remain unfavorable under the primary interpretation.

Dataset	Raw positive	Raw negative	Temp positive	Temp negative
Gaussian blobs	2 / 6	4 / 6	6 / 6	0 / 6
Two moons	2 / 6	4 / 6	6 / 6	0 / 6
Digits	2 / 6	4 / 6	6 / 6	0 / 6
Breast cancer	2 / 6	4 / 6	6 / 6	0 / 6

Table 3: Sign counts for the metric-sensitivity suite. Positive signs favor CEMA under the lower-is-better convention, and negative signs favor the strongest comparator. Every dataset has the same qualitative structure: raw summaries are split, while temperature-scaled summaries favor CEMA.

CEMA improving the ranking or magnitude of probabilities enough for a proper score but not aligning raw confidence bins with empirical correctness. It also shows why post-hoc calibration should not be mixed into the primary readout claim. A positive temperature-scaled result answers a different question: what happens after a validation-fitted scalar transformation of logits.

The sign-count table is intentionally coarse. It does not replace the paired estimates in the primary table and it does not encode effect sizes. Its purpose is to show that the direction of evidence depends on the metric layer. A reader interested in deployment with a validation calibration step should focus on the temperature-scaled layer and should then evaluate whether the validation set is large and representative enough to fit the scalar temperature. A reader interested in raw training-time readouts should focus on the raw layer and the primary success predicate. The two readers have different operational constraints, so the manuscript keeps both interpretations visible rather than averaging them into one ambiguous score.

The sensitivity result suggests two follow-up directions. First, classwise or adaptive-bin calibration estimates could test whether the raw-ECE failure is specific to top-label fixed bins. Second, calibration methods that operate on probabilities could be combined with CEMA readouts if the goal is deployed probability quality rather than a pure training-readout comparison. Those directions would not rescue the primary hypothesis in this paper, because the hypothesis required raw CEMA to beat the strongest averaging baseline under the prespecified rule.

D Reproducibility Details

The reproducibility material records the operational context needed to repeat the evidence without promoting that context into the scientific claim. The study uses small CPU-feasible classifiers, deterministic seeds, local scikit-learn datasets or generators, and finite readout grids. These choices are important for replication and for bounding conclusions, but the manuscript’s contribution is the evaluated behavior of the scheduled EMA readout, not the compute environment itself. The details below therefore focus on what must match to reproduce the figures, tables, and symbolic checks.

D.1 Implementation and Compute Scope

The executed implementation uses deterministic seed control, scikit-learn data utilities, NumPy-based experiment analysis, and a small neural classifier training pipeline. The training grid includes logistic references, one-hidden-layer MLPs, and

two-hidden-layer MLPs with no more than the requested small-model parameter budget. The full metrics export contains 17,280 rows and records dataset, noise level, seed, architecture, parameter count, readout, raw or temperature-scaled layer, validation NLL, validation accuracy, test NLL, test accuracy, Brier score, three fixed-bin ECE variants, quantile ECE, final-quarter validation-NLL variance, path smoothness, and sharpness-proxy summaries. These are reproducibility facts, not claims of scientific novelty.

The uncertainty procedure is paired because every readout is derived from an identical trajectory within a seed. Confidence intervals are computed on matched readout differences rather than on independent samples. The Wilcoxon test is one-sided only for the primary ECE direction specified in (4); other diagnostics are interpreted descriptively with intervals. Symbolic checks were executed for the finite algebraic obligations. They check the schedule, coefficients, noise matrices, and horizon bracketing, and they are reported separately from empirical model performance. This separation is important for reproduction: a failure in raw ECE cannot be repaired by an algebraic proof, and a passed algebra check cannot be read as a calibration theorem.

Reproducing the experiment requires preserving three sources of determinism. The data split seed fixes which examples enter training, validation, and test sets. The label-noise seed fixes which training labels are corrupted and to which alternative classes. The optimization seed fixes initialization and any stochastic order used in training. If these are allowed to vary independently without recording, paired readout comparisons lose their variance-reduction benefit. The export tables therefore carry seed, dataset, noise rate, architecture, readout, and calibration layer as explicit factors. The manuscript describes these factors in scientific prose rather than exposing machine-oriented headers, but the interpretation relies on the same factorization.

D.2 Figure and Table Provenance

The four manuscript figures are vector PDF graphics generated from the final validation tables and metrics. Visual quality checks confirmed nonempty rendered previews, readable panel content, distinct figures, and no unresolved clipping or contrast issues. The first three figures appear in the main text because they cover the primary result, reliability structure, and mechanism diagnostics. The metric-sensitivity matrix appears in the appendix to keep the main figure count controlled while still preserving the evidence needed to interpret metric fragility. The captions state the variable roles and interpretation so that the figures can be understood without inspecting machine-readable exports.

The main primary-success table rewrites machine-oriented columns into publication labels and keeps the exact numerical direction of the decision rule. The mechanism table summarizes the diagnostic deltas behind the corresponding figure. The appendix sign-count table compresses the metric-sensitivity matrix into a readable pattern. The complete machine-readable tables remain part of the export package, but the manuscript does not rely on raw column names or local paths for interpretation. This matters because readers should be able to understand the evidence in scientific terms: which dataset was tested, which comparator was strongest, which metric direction favors CEMA, and which claim the evidence supports or contradicts.

The final figures are used sparingly in the main text to avoid substituting visual density for argument. The primary figure shows the failed success predicate and the accompanying NLL and accuracy context. The reliability figure shows confidence-bin behavior that explains why raw ECE can diverge from NLL. The mechanism figure links the algebraic schedule checks with late-loss and sharpness diagnostics. The metric-sensitivity figure is placed in the appendix because it supports interpretation rather than deciding the primary claim. This layout follows the evidence hierarchy: first the primary decision, then explanatory probability behavior, then mechanism diagnostics, then secondary sensitivity.

E Failure Boundaries and Follow-Up Experiments

The final appendix section separates evidence that would strengthen the current conclusion from evidence that would answer a different question. A same-scope repair would need to preserve clean held-out labels, paired seeds, and the strongest-baseline rule while changing schedule shape, horizon, or optimizer controls. A scope expansion could evaluate larger architectures, robust noisy-label objectives, or post-hoc calibration pipelines, but those experiments would no longer test the same parameter-readout-only claim. This distinction keeps future work aligned with the rejected hypothesis and scope.

E.1 What Would Change the Conclusion

A follow-up experiment would change the conclusion only if it changed the evidence under a clearly stated scope. For the same small-classifier regime, the most direct repair would be a schedule-shape and horizon study that tests whether another deterministic β_t curve beats both fixed EMA and tail Polyak on raw ECE. Such a study should keep the paired seed design, clean held-out labels, and strongest-baseline selection rule. It should also include constant-learning-rate controls to separate the effect of the optimizer schedule from the readout schedule. Warm restarts would require period-local definitions for t , η_t , and β_t before they can be compared fairly.

The next experiment should also decide whether success is raw or post-hoc calibrated probability quality before running. If raw ECE remains the target, then any validation-fitted transformation should stay diagnostic. If calibrated deployment probabilities become the target, then CEMA plus temperature scaling should be compared with standard post-hoc calibrators and possibly with prediction ensembles. These are different hypotheses. The current data suggest that the post-hoc direction is more promising than the raw-readout direction, but the current acceptance rule was not designed to declare that as success. A careful follow-up would therefore preregister separate success predicates for raw ECE, temperature-scaled ECE, NLL, and Brier score.

A broader study could ask a different question. If the goal is deployed calibrated probabilities, then CEMA plus validation-fitted temperature scaling may be a reasonable candidate because the sensitivity suite is favorable after scaling. If the goal is noisy-label robustness, then robust losses, bootstrapping, co-teaching, DivideMix, and related objective-changing methods become appropriate baselines. If the goal is uncertainty estimation, then SWAG or ensembles should be included despite higher cost. None of those expansions overturn this paper’s claim because they change the intervention class. The present conclusion is confined to parameter readouts from identical small-classifier trajectories under symmetric training-label noise.

The negative result remains valuable under these boundaries. It prevents a plausible schedule-coupling idea from being promoted as a raw calibration improvement without strong comparator evidence. It also shows that validation-loss stabilization and raw ECE can move in opposite directions. Future work should therefore report both types of metrics and should treat success criteria as commitments rather than adjustable summaries. That discipline is especially important for small datasets, where calibration estimates are sensitive and a visually attractive reliability curve can be less reliable than a paired seed table.

A final boundary concerns cost and simplicity. Fixed EMA and tail Polyak averages are simpler than a schedule-coupled rule because they do not require an additional mapping from learning-rate time to averaging memory. A scheduled method should therefore earn its complexity by improving the target metric under strong comparisons. In this study, it does not earn that complexity for raw ECE. That does not make the schedule useless; it means the supported use case is narrower. The method may still be a convenient loss-stability readout or a candidate input to post-hoc scaling, but those uses should be evaluated on their own terms.