

Curvature-Based Shrinkage for Label-Budgeted Recalibration under Dataset Shift: An Empirical Risk Study

Jufeng Li

q3529416243@gmail.com

China University of Petroleum, East China

Wenjie Li

China University of Petroleum, East China

Research Article

Keywords: Probability calibration, Dataset shift, Temperature scaling, Shrinkage estimation, Label budgets, Deployment risk

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11372249/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Curvature-Based Shrinkage for Label-Budgeted Recalibration under Dataset Shift: An Empirical Risk Study

Jufeng Li^{1*} and Wenjie Li²

¹College of Science, China University of Petroleum (East China),
Qingdao, 266580, China.

²College of Mechanical and Electronic Engineering, China University of
Petroleum (East China), Qingdao, 266580, China.

*Corresponding author(s). E-mail(s): q3529416243@gmail.com;
Contributing authors: dvien5162@gmail.com;

Abstract

Recalibrating a deployed classifier with very few target labels can increase prediction loss under dataset shift. We study the magnitude of this update through curvature-based shrinkage toward a source-calibrated temperature. A quadratic surrogate gives a classical Gaussian shrinkage interpretation and identifies the assumptions needed for a bias–variance analysis; it does not provide an unconditional safety guarantee for the implemented estimator. We compare curvature shrinkage, a significance-gated alternative, and anchored calibration baselines using shared label queries. An archived development benchmark contains 32 domains from eight families. The subsequent three-family evaluation covers sampling differences in remote sensing, population differences in activity recognition, and calendar-based shifts in restaurant inspections. Implementing the predeclared single-class exclusion leaves nine eligible domains at budgets of 8 and 16 labels and eight at 32 labels. Curvature shrinkage improves mean negative log-likelihood by 0.055–0.096 nats over retaining the source calibration. Its additional advantage over a constant anchor is small and statistically inconclusive. At eight labels, gating reduces the frequency of negative gains from 41.5% to 12.6%, but has lower mean gain and a worse minimum domain-average gain. Rankings also depend on the magnitude threshold used to define harm. The results support a deployment trade-off rather than uniform dominance, and motivate explicit reporting of both frequency and severity of harmful recalibration.

Keywords: Probability calibration, Dataset shift, Temperature scaling, Shrinkage estimation, Label budgets, Deployment risk

1 Introduction

Probability estimates are useful only when their numerical values remain meaningful after deployment. Temperature scaling is a simple post-hoc calibration method that changes confidence through a scalar parameter while preserving the predicted class [1]. A temperature fitted on source data may cease to be appropriate when the target population, sampling process, or operating period changes. Acquiring a small set of target labels offers an inexpensive repair, but estimating even one parameter from eight or sixteen observations can be unstable.

This creates a practical decision: how far should the fitted target temperature move from the source value? Keeping the source calibration avoids estimation noise but can preserve genuine miscalibration. Fitting an unconstrained target temperature can reduce bias while increasing variability. An anchored estimate interpolates between these choices. Its usefulness should therefore be assessed through average prediction loss and the frequency and severity of harmful updates, rather than average improvement alone.

We examine budgeted curvature shrinkage (BCS) and a significance-gated, source-centered version of risk-aware shrinkage (gated RAS). Both are lightweight post-hoc rules. Their mathematical motivation comes from classical normal shrinkage and a quadratic loss surrogate; we do not claim that the underlying normal-posterior formula is new. The empirical question is whether these rules offer useful risk trade-offs against a well-tuned constant anchor under limited target-label budgets.

The study contributes an implementation-explicit description of the two rules, a conditional account of their bias–variance behavior, and a traceable comparison across small target-label budgets. The evaluation distinguishes internally frozen primary hypotheses from exploratory comparisons, implements the archived single-class eligibility rule, and reports a secondary harm threshold as a sensitivity analysis. The findings show why reducing the number of negative-gain cells should not be interpreted as minimizing all forms of deployment risk. They also identify limitations of prior centering and of theoretical interpretations based on estimated curvature.

2 Related work

2.1 Calibration and dataset shift

Temperature scaling provides an effective low-capacity baseline for modern classifiers [1]. More flexible calibration maps, including Dirichlet calibration, can model class-dependent distortions but require estimating additional parameters [2]. Under dataset shift, multi-domain temperature scaling exploits heterogeneity among calibration domains [3]. A recent benchmark across natural image shifts found that specialized calibration methods do not consistently improve on simpler alternatives

[4]. These findings motivate careful comparison with a strong anchored baseline rather than only an unstable unconstrained fit.

Label-shift adaptation addresses changes in class priors under assumptions on the class-conditional feature distribution. Maximum likelihood combined with bias-corrected calibration is a strong baseline in that setting [5]. Our experiments do not establish pure label shift, and scalar recalibration should not be interpreted as a complete correction for arbitrary distributional changes. In particular, the remote-sensing evaluation includes differing sampling procedures and class proportions.

2.2 Shrinkage, selection, and risk control

The normal-prior and quadratic-likelihood calculation used here is a classical shrinkage construction, related to empirical-Bayes estimation [6]. Source-temperature estimation variance and target-source discrepancy are distinct quantities: a precisely estimated source temperature does not imply that it is appropriate for a new domain. This distinction motivates a discrepancy-scale parameter, but its numerical value remains a modeling choice that must be assessed empirically.

Selective classification trades coverage for accuracy [7]. Our gate instead chooses whether to deploy a calibration update; it does not reject individual predictions. Conformal risk control provides guarantees for suitable monotone loss families under specified calibration assumptions [8]. A heuristic comparison of upper confidence bounds computed on the same labels used to fit candidate temperatures is not equivalent to that procedure. We therefore do not interpret such a comparator as valid conformal risk control or infer that distribution-free risk control is intrinsically ineffective at small budgets.

3 Problem setting and evaluation criteria

Let a trained classifier produce raw probabilities $p_c(x)$ for classes $c = 1, \dots, K$. An absolute temperature $T = e^\theta$ defines

$$q_{\theta,c}(x) = \frac{p_c(x)e^{-\theta}}{\sum_{j=1}^K p_j(x)e^{-\theta}}. \quad (1)$$

Probabilities are clipped below at 10^{-12} before taking logarithms. The source temperature is $\theta_s = \log T_s$. On a target domain, a shared query set Q_B contains B labeled observations, with $B \in \{8, 16, 32, 64, 128\}$. The empirical objective is the mean negative log-likelihood (NLL),

$$\bar{L}_B(\theta) = \frac{1}{B} \sum_{(x_i, y_i) \in Q_B} -\log q_{\theta, y_i}(x_i). \quad (2)$$

Scalar temperatures are fitted within $T \in [0.1, 10]$. Every method acts on the same raw probabilities; a fitted target temperature is not applied to already source-calibrated probabilities.

For a domain, backbone, seed, and budget, define the held-out gain

$$G = L_{\text{test}}(\theta_s) - L_{\text{test}}(\hat{\theta}). \quad (3)$$

Positive gain is beneficial. A *cell* is one such experimental combination. The primary harm indicator is $\mathbf{1}\{G < 0\}$. Seeds and backbones are averaged within each domain before computing the domain-equal mean, domain-bootstrap interval, and minimum domain-average gain. The fraction of domains with negative average gain is reported separately from the fraction of harmful cells. The minimum is over observed domains and is not a population worst-case bound. NLL is a proper scoring rule but does not isolate calibration from all other aspects of probabilistic quality; Brier score and ten-bin expected calibration error are supplied as secondary metrics.

As an exploratory sensitivity analysis, we also report $\mathbf{1}\{G < -0.01\}$ to distinguish small numerical losses from losses exceeding 0.01 nats. This threshold is not a predeclared primary endpoint or a universal operational tolerance.

4 Curvature shrinkage and gated updating

4.1 Budgeted curvature shrinkage

The unanchored target fit is $\hat{\theta}_{\text{TS}} = \arg \min_{\theta} \bar{L}_B(\theta)$. In the implemented BCS rule, the curvature is measured at the *source* temperature, not at the target optimum:

$$\hat{H}_s = \text{clip}\left(\frac{\bar{L}_B(\theta_s + h) - 2\bar{L}_B(\theta_s) + \bar{L}_B(\theta_s - h)}{h^2}, 10^{-3}, 50\right), \quad h = 0.05. \quad (4)$$

With the development-frozen constant scale $\tau_0^2 = 0.05$, BCS deploys

$$\hat{\kappa}_s = \frac{B\hat{H}_s\tau_0^2}{1 + B\hat{H}_s\tau_0^2}, \quad \hat{\theta}_{\text{BCS}} = \text{clip}\{\theta_s + \hat{\kappa}_s(\hat{\theta}_{\text{TS}} - \theta_s), \log 0.1, \log 10\}. \quad (5)$$

The clipping makes the weight well defined even when the finite-difference curvature is small or negative. The constant scale is a frozen regularization choice; the experiments do not establish that it equals the true cross-domain discrepancy second moment.

4.2 Source-centered gated RAS

Gated RAS evaluates curvature at $\hat{\theta}_{\text{TS}}$ with the same finite-difference and clipping settings. It uses a global residual scale $\tau_g^2 = 0.4151041376$ from a ridge discrepancy model fitted on five development families. The archived model stores 315 training records across seeds and backbones; these are not 315 independent domains. Its scale is *constant across target domains*. Although the mean model uses label-free drift descriptors, the primary source-centered variant sets its mean correction to zero. Consequently, the residual scale should not automatically be interpreted as the source-centered discrepancy second moment.

Let $\hat{\kappa}_t = B\hat{H}_t\tau_g^2/(1 + B\hat{H}_t\tau_g^2)$ and $s_B = (B\hat{H}_t)^{-1/2}$. The frozen gate uses $z = 1$ and deploys

$$\hat{\theta}_g = \begin{cases} \text{clip}\{\theta_s + \hat{\kappa}_t(\hat{\theta}_{\text{TS}} - \theta_s), \log 0.1, \log 10\}, & |\hat{\theta}_{\text{TS}} - \theta_s| > zs_B, \\ \theta_s, & \text{otherwise.} \end{cases} \quad (6)$$

For $B < 4$, the implementation retains the source temperature. The gate is a heuristic based on a local uncertainty scale; $z = 1$ is not a finite-sample confidence guarantee. BCS and gated RAS differ in curvature location and scale as well as gating, so their comparison cannot isolate the gate alone.

4.3 Anchored and capacity baselines

The principal constant anchor minimizes

$$\bar{L}_B(\theta) + \gamma(\theta - \theta_s)^2, \quad \gamma = 0.5. \quad (7)$$

Its value was frozen using development results. We also report a stronger anchor ($\gamma = 4$), a budget schedule, source-variance and weighted-source empirical-Bayes anchors, class-bias calibration, regularized per-class temperatures, and empirical selection rules in Online Resource 1. The per-class temperature baseline uses a penalty of 4 toward the source temperature; class-bias calibration uses an ℓ_2 penalty of 2. These are the archived implementations, not exhaustive representatives of all calibration methods. Methods share the same queried labels and held-out evaluation observations.

5 Conditional interpretation of shrinkage

The following statements explain a simplified shrinkage mechanism. They do not establish an unconditional risk guarantee for Eqs. (5) and (6).

Proposition 1 (Quadratic Gaussian surrogate) *For an unconstrained quadratic surrogate $\bar{L}_B(\theta) = C + \frac{a}{2}(\theta - \hat{\theta}_{\text{TS}})^2$, $a > 0$, and $\gamma \geq 0$, the minimizer of $\bar{L}_B(\theta) + \gamma(\theta - \theta_p)^2$ is $\theta_p + w(\hat{\theta}_{\text{TS}} - \theta_p)$ with $w = a/(a + 2\gamma)$. Under a normal prior with center θ_p and variance $\tau^2 > 0$ and likelihood proportional to $\exp\{-B\bar{L}_B(\theta)\}$, both the posterior mean and mode use $w = Ba\tau^2/(1 + Ba\tau^2)$.*

Proof Differentiation gives $(a + 2\gamma)\theta = a\hat{\theta}_{\text{TS}} + 2\gamma\theta_p$. Completing the square in the Gaussian posterior gives precision $Ba + 1/\tau^2$ and the stated mean. The conversion is $\gamma = 1/(2B\tau^2)$, which does not itself depend on curvature. $\square$

Even a constant γ has a curvature-dependent weight in this surrogate. Thus, a constant penalty does not imply a constant interpolation weight. The exact penalized NLL fit and the explicit interpolation need not agree outside a quadratic surrogate, particularly when the curvature is measured at a different point or bounds are active.

Proposition 2 (Risk for a deterministic weight) *Suppose $L(\theta) - L(\theta^*) = \frac{H}{2}(\theta - \theta^*)^2$, $H > 0$. Write $\delta = \theta^* - \theta_p$ and $\epsilon = \hat{\theta}_{\text{TS}} - \theta^*$, with $\mathbb{E}\epsilon = 0$, $\mathbb{E}(\delta\epsilon) = 0$, $\mathbb{E}\delta^2 = \tau_D^2$, $\mathbb{E}\epsilon^2 = v_B$, and $\tau_D^2 + v_B > 0$. For a deterministic weight w , its expected excess risk is*

$$R(w) = \frac{H}{2}[(1-w)^2\tau_D^2 + w^2v_B], \quad w^* = \frac{\tau_D^2}{\tau_D^2 + v_B}. \quad (8)$$

Moreover, $R(w) - R(w^*) = \frac{H}{2}(\tau_D^2 + v_B)(w - w^*)^2$.

Proof The estimation error is $-(1-w)\delta + w\epsilon$. Expanding its square and taking expectations removes the cross term by the stated assumption. Differentiation and completing the square give the remaining identities. $\square$

In an exact quadratic empirical-loss model, if the gradient noise at θ^* is sub-Gaussian with variance proxy σ^2/B , then ϵ has proxy $\sigma^2/(BH^2)$. For a general misspecified smooth likelihood, the local variance is approximately $J/(BH^2)$, where J is the score variance. Substituting $1/(BH)$ requires additional information-equality assumptions. Further, the implemented weight and target fit use the same labels. Their dependence prevents direct substitution of $\hat{\kappa}$ into the deterministic-weight expectation above. Curvature clipping, bounded fitting, prior-scale misspecification, and non-quadratic remainder terms are also outside this proposition.

Proposition 3 (Conditional comparison with retaining the source) *If L is convex on the segment between θ_s and $\hat{\theta}_{\text{TS}}$, then, for any $w \in [0, 1]$, a source-centered interpolation that is harmful relative to θ_s implies that $\hat{\theta}_{\text{TS}}$ is harmful. The statement is pointwise and permits a data-dependent weight.*

Proof Convexity gives $L((1-w)\theta_s + w\hat{\theta}_{\text{TS}}) \leq (1-w)L(\theta_s) + wL(\hat{\theta}_{\text{TS}})$. If $L(\hat{\theta}_{\text{TS}}) \leq L(\theta_s)$, the interpolation cannot have larger loss than $L(\theta_s)$. Taking the contrapositive proves the claim. $\square$

Convexity in log-temperature is not assumed to hold globally for the actual NLL. The proposition therefore supports only a conditional non-increase in harm occurrence relative to target TS. It gives neither strict improvement nor dominance over a constant anchor. Online Resource 1 provides a deterministic-weight tail calculation and states the threshold condition needed to compare its upper bound with target TS.

6 Experimental protocol

6.1 Domains and development decisions

The archived development benchmark contains 32 domains in eight families: occupancy, gas, electricity, bank, semiconductor manufacturing, Polish bankruptcy, ozone, and EEG eye state. Development data were available for method design and variant selection. The discrepancy mean model and its global residual scale used only

five of these families; its predictions were frozen before the subsequent three-family evaluation.

The nominal evaluation comprises ten domains. Wilt uses 4,339 training image segments as source and two contiguous halves of the 500 validation records as targets. The original study used a single QuickBird image acquired on 27 August 2012; its validation set was stratified sampled and contained 187 diseased-tree observations [9, 10]. Training segments were merged across several segmentation levels, while the validation records use a specified level. We describe this comparison as a sampling-distribution difference, not two temporally separated image acquisitions. Spatial independence is not established. The supplied file order is used only to form two evaluation blocks and is not evidence of spatial scan order or chronology.

The wrist-accelerometer dataset contains 16 volunteers performing 14 activities [11]. Non-overlapping 128-sample windows yield signal-statistics features; walking and stair activities are grouped against the remaining activities. Volunteers are ordered by their first recording timestamp, with the earliest six forming the source and the remaining ten partitioned into four target groups. The first recording time is an ordering rule, not proof of recruitment chronology. The final target group contains 53 all-negative windows and is excluded from the primary analysis under the archived eligibility rule.

Chicago restaurant inspections use source records from 2016–2018 and target groups from 2019–2020, 2021, 2022, and 2023 [12]. Pass and pass-with-conditions are grouped against fail. Features are inspection type, risk category, facility type, zip code, latitude, and longitude, with source-only vocabulary and imputation. Calendar differences may include policy and sampling changes; the experiment does not identify a causal pandemic effect.

6.2 Backbones and query budgets

The primary evaluation uses logistic regression, histogram gradient boosting, and a one-hidden-layer multilayer perceptron, with five seeds (101, 211, 307, 503, 701). Source observations are split into 60% training, 20% temperature fitting, and 20% reserved monitoring records. Each target domain is randomly divided into a query pool and a held-out test set of approximately equal size. Nested prefixes of one random query permutation provide the five budgets; the queried indices are identical across methods within each cell. Budgets exceeding one quarter of the entire target-domain size are skipped.

After the entire-domain single-class exclusion, budgets 8, 16, 32, 64, and 128 cover 9, 9, 8, 6, and 4 domains, respectively. At 128 labels only the four Chicago domains remain. The evaluation is offline recalibration within target periods and does not emulate forward-chained forecasting. Later Gaussian naive-Bayes and UCB-selection experiments used a different mixture of development and evaluation domains; they are not incorporated into the primary confirmation claims here.

6.3 Analysis plan and deviations

An internally frozen protocol dated 11 October 2026 specified source-centered gated RAS as the primary method. Its composite criterion required a positive paired domain-average gain over the constant anchor, no increase in harmful-cell frequency, and a minimum domain-average gain of at least -0.02 at budgets 8, 16, and 32. No public preregistration is claimed. BCS was included as a secondary comparison; its subsequent emphasis is exploratory rather than a replacement primary hypothesis.

The archived summary had not implemented its entire-domain single-class exclusion. The present analysis applies that predeclared rule to unchanged archived NLL records. Model training and temperature fitting have not been repeated. We retain the all-domain summaries as a sensitivity analysis and explicitly correct the descriptions of the Wilt source and the global residual scale.

We average backbones and seeds within domains and use 4,000 domain-bootstrap resamples for percentile 95% intervals. Paired comparisons use 4,000 two-sided domain sign-flip draws; exact sign-flip results are supplied as a sensitivity check. Family-equal summaries are also reported. Domains from one family are dependent, and only three families are represented, so these intervals and tests describe the small benchmark and should not be interpreted as strong evidence of cross-family generalization. Exploratory comparisons are not adjusted for multiple testing.

7 Results

7.1 Average gain and harm frequency

Table 1 and Fig. 1 report the corrected primary-domain summaries. At eight labels, target TS has average gain -0.264 nats, a minimum domain-average gain of -0.833 , and 63.0% harmful cells. BCS has average gain $+0.055$, compared with $+0.051$ for the constant anchor. Both have 41.5% harmful cells. Gated RAS has average gain $+0.034$ and 12.6% harmful cells. Its frequency benefit is accompanied by a lower mean and a worse minimum domain-average gain than BCS.

At budgets 16 and 32, BCS yields $+0.072$ and $+0.096$ nats over retaining the source calibration. Table 2 shows that its additional paired gains over the constant anchor, $+0.016$ and $+0.028$, have intervals spanning zero. The data therefore do not establish mean-optimality or significant superiority to the strongest constant anchor. At budget 32, the constant anchor has lower harmful-cell frequency than BCS (25.8% versus 34.2%), despite lower mean gain.

7.2 The frozen primary criterion is not met

Gated RAS has paired mean differences of -0.0164 , -0.0044 , and $+0.0037$ nats against the constant anchor at budgets 8, 16, and 32. Although its harmful-cell frequency is lower at these budgets, the mean requirement fails at the first two. At budget 16, its minimum domain-average gain is -0.0396 , below the predeclared -0.02 threshold. The composite primary criterion is therefore not met. The empirical benefit is a reduction in negative-gain frequency under some conditions, rather than successful confirmation of all intended reliability goals.

Table 1 Eligible-domain NLL gain versus keep-source (95% domain-bootstrap interval), harmful-cell frequency, and minimum domain-average gain. Budgets 8, 16, and 32 have 9, 9, and 8 domains (135, 135, and 120 cells). Intervals do not account for all within-family dependence.

Method	B	Gain [95% CI]	Harm (%)	Minimum
Target TS	8	-0.264 [-0.473, -0.066]	63.0	-0.833
	16	+0.017 [-0.102, +0.158]	55.6	-0.213
	32	+0.063 [-0.045, +0.211]	53.3	-0.118
Anchor $\gamma = 0.5$	8	+0.051 [-0.002, +0.119]	41.5	-0.010
	16	+0.056 [+0.001, +0.128]	35.6	-0.003
	32	+0.068 [+0.002, +0.155]	25.8	+0.000
BCS	8	+0.055 [+0.001, +0.128]	41.5	-0.001
	16	+0.072 [+0.001, +0.165]	38.5	-0.004
	32	+0.096 [+0.001, +0.222]	34.2	-0.003
Gated RAS	8	+0.034 [-0.005, +0.083]	12.6	-0.019
	16	+0.052 [-0.010, +0.127]	14.1	-0.040
	32	+0.072 [-0.007, +0.169]	20.8	-0.017
Vector TS	8	+0.038 [+0.002, +0.085]	34.1	-0.001
	16	+0.041 [+0.002, +0.093]	28.9	-0.000
	32	+0.049 [+0.002, +0.114]	19.2	+0.000
EB-total	8	+0.013 [+0.001, +0.034]	39.3	-0.000
	16	+0.022 [+0.002, +0.052]	33.3	-0.001
	32	+0.035 [+0.001, +0.086]	25.0	+0.000

Table 2 Paired domain-average differences versus the constant anchor. Two-sided sign-flip tests use 4,000 draws. These small-benchmark comparisons are not multiplicity-adjusted.

Method	B	Difference [95% CI]	p	Domains
BCS	8	+0.005 [-0.001, +0.012]	0.250	9
	16	+0.016 [-0.001, +0.038]	0.300	9
	32	+0.028 [-0.001, +0.068]	0.328	8
Gated RAS	8	-0.016 [-0.041, -0.001]	0.115	9
	16	-0.004 [-0.018, +0.010]	0.593	9
	32	+0.004 [-0.010, +0.024]	0.912	8

7.3 Risk rankings depend on the endpoint

Table 3 reports the exploratory 0.01-nat threshold. At eight labels, BCS has 4.4% materially harmful cells, compared with 11.9% for gated RAS and 18.5% for the constant anchor. Gated RAS has fewer negative-gain cells than BCS, but more cells beyond this magnitude threshold. This is consistent with occasional accepted updates being more costly and explains why frequency, minimum domain-average gain, and

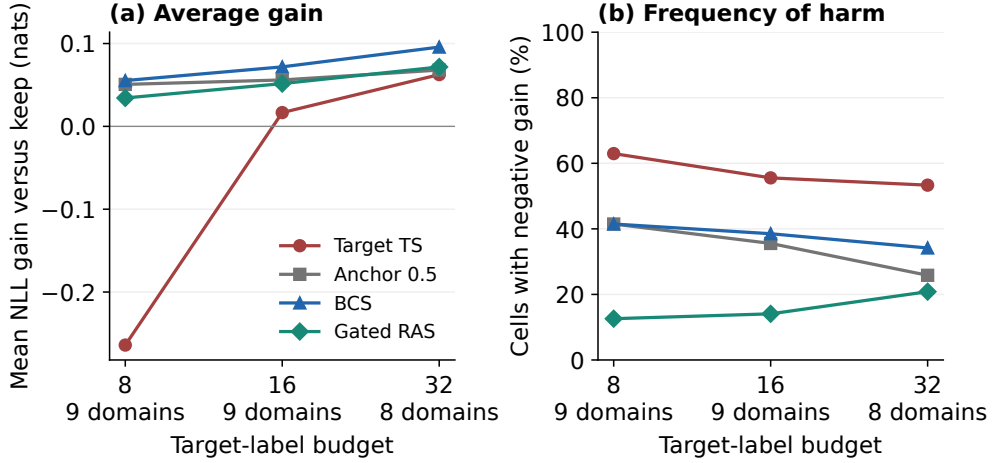


Fig. 1 Average NLL gain and harmful-cell frequency after the single-class exclusion. Curves use shared query trajectories but the eligible domain set changes at budget 32. Points are descriptive summaries; intervals and minimum domain-average gains appear in Table 1. Lower harmful-cell frequency does not by itself imply lower loss severity.

Table 3 Exploratory sensitivity: cells with gain below -0.01 nats (%). This magnitude threshold is not the predeclared primary harm endpoint.

Method	$B = 8$	$B = 16$	$B = 32$
Target TS	54.1	39.3	32.5
Anchor $\gamma = 0.5$	18.5	7.4	5.0
BCS	4.4	5.2	7.5
Gated RAS	11.9	11.1	11.7

magnitude-threshold harm must be reported together. The comparison does not isolate gating because the methods also use different curvature locations and scales.

7.4 Family dependence and sensitivity

At eight labels, the BCS mean gains within Wilt, activity recognition, and Chicago are approximately $+0.244$, $+0.0042$, and -0.00053 nats, respectively. The overall improvement is concentrated in the two Wilt blocks. Removing Wilt leaves seven eligible domains with a BCS mean gain of approximately $+0.0015$ nats. This limited improvement prevents a claim of uniform success across shift types.

The larger-budget curves also change population because of the cap. Online Resource 1 reports a fixed four-domain Chicago comparison across all budgets, family-equal summaries, the all-domain sensitivity, and secondary probabilistic metrics. Curvature-quartile patterns in the archive are exploratory associations; pooling budgets and families does not independently validate the risk propositions. Recomputing

the implemented weight to floating-point precision is an implementation consistency check, not a test of a statistical guarantee.

8 Discussion

The experiments support a cautious practical interpretation. An unconstrained target-temperature fit can be unreliable when the queried labels are few. A strong constant anchor is a competitive, low-complexity baseline. Curvature shrinkage sometimes provides higher average gain, while a gate can avoid many small negative updates. Which rule is preferable depends on the user’s loss tolerance and on whether the objective is average gain, harmful-update frequency, or downside magnitude.

Several limitations constrain the conclusions. First, the mathematical interpretation assumes a quadratic or locally well-behaved loss and correctly specified variance relationships; the implemented estimators use noisy curvature, fixed scales, clipping, and bounded fits. Second, the evaluation contains only three new families and is dominated by a remote-sensing sampling difference. Third, population windows and calendar groups may be dependent, making domain-level intervals optimistic for broad generalization. Fourth, the archived ablations change multiple factors and do not establish a causal contribution for variance conditioning or gating. Fifth, no valid distribution-free risk-control baseline is evaluated in the primary comparison. Finally, the backbones and binary target tasks do not establish behavior for large pretrained models, multiclass tasks, or online sequential deployment.

A stronger follow-up should freeze BCS as a new primary hypothesis, obtain genuinely new external domains, control center, scale, curvature point, and gate separately, and use a valid risk-control comparator. These experiments could determine whether the small average advantage over a constant anchor persists under independent family-level variation. Until then, the present findings support an empirical trade-off and an explicit risk-reporting protocol rather than a universal safe-update claim.

9 Conclusion

We studied curvature-based source-temperature shrinkage for recalibration with limited target labels. A conditional quadratic account explains its bias–variance motivation, while the corrected archived evaluation shows both benefits and limits. Curvature shrinkage mitigates the substantial small-budget failures of target TS, but its additional gain over a constant anchor is statistically inconclusive. Gated updating reduces harmful-cell frequency at a cost that is visible in average gain and loss severity. Reporting all of these endpoints is necessary for a meaningful deployment decision.

Statements and Declarations

Funding. The authors received no financial support for this research.

Competing interests. The authors declare no competing interests.

Author contributions. Jufeng Li: conceptualization, methodology, formal analysis, writing—original draft, and project administration. Wenjie Li: software, data curation, validation, visualization, and writing—review and editing.

Data availability. The underlying data are publicly available from the UCI Machine Learning Repository and the City of Chicago Data Portal [10–12]. Derived per-cell results, eligibility exclusions, and corrected analysis tables are supplied in the accompanying reproducibility materials (Online Resource 2). The exact archived source-data hashes and split rules are retained there.

Code availability. The statistical reanalysis script and archived calibration implementation are supplied in the accompanying reproducibility materials (Online Resource 2). The reanalysis does not retrain classifiers or change temperature fits.

Ethics statement. This study analyses previously released datasets and does not involve recruitment of new participants or collection of new personal data. Information about the original collection is available from the dataset providers.

Use of AI assistance. An OpenAI AI assistant was used to support manuscript restructuring, English revision, and consistency checks of the mathematical exposition and reported results. The authors retain responsibility for the final scientific content.

Supplementary information. Online Resource 1 provides conditional derivations, full corrected tables, eligibility and analysis details, and sensitivity results. Online Resource 2 contains the archived per-cell results, corrected analysis tables, and statistical reanalysis code.

References

- [1] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330 (2017). <https://proceedings.mlr.press/v70/guo17a.html>
- [2] Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., Flach, P.: Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In: Advances in Neural Information Processing Systems, vol. 32 (2019). <https://proceedings.neurips.cc/paper/2019/hash/8ca01ea920679a0fe3728441494041b9-Abstract.html>
- [3] Yu, Y., Bates, S., Ma, Y., Jordan, M.I.: Robust calibration with multi-domain temperature scaling. In: Advances in Neural Information Processing Systems, vol. 35, pp. 27510–27523 (2022). <https://doi.org/10.52202/068431-1995>
- [4] Roschewitz, M., Mehta, R., Ribeiro, F.d.S., Glocker, B.: Where are we with calibration under dataset shift in image classification? Transactions on Machine Learning Research (2025). Published October 2025

- [5] Alexandari, A., Kundaje, A., Shrikumar, A.: Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In: Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 222–232 (2020). <https://proceedings.mlr.press/v119/alexandari20a.html>
- [6] Casella, G.: An introduction to empirical bayes data analysis. *The American Statistician* **39**(2), 83–87 (1985) <https://doi.org/10.1080/00031305.1985.10479400>
- [7] El-Yaniv, R., Wiener, Y.: On the foundations of noise-free selective classification. *Journal of Machine Learning Research* **11**, 1605–1641 (2010)
- [8] Angelopoulos, A.N., Bates, S., Fisch, A., Lei, L., Schuster, T.: Conformal risk control. In: International Conference on Learning Representations (2024). <https://openreview.net/forum?id=33XGfHLtZg>
- [9] Johnson, B.A., Tateishi, R., Hoan, N.T.: A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees. *International Journal of Remote Sensing* **34**(20), 6969–6982 (2013) <https://doi.org/10.1080/01431161.2013.810825>
- [10] Johnson, B.: Wilt. UCI Machine Learning Repository. Dataset (2013). <https://doi.org/10.24432/C5KS4M> . <https://archive.ics.uci.edu/dataset/285/wilt>
- [11] Bruno, B., Mastrogiovanni, F., Sgorbissa, A.: Dataset for ADL Recognition with Wrist-worn Accelerometer. UCI Machine Learning Repository. Dataset (2012). <https://doi.org/10.24432/C5PC99> . <https://archive.ics.uci.edu/dataset/283>
- [12] City of Chicago: Food Inspections. City of Chicago Data Portal. Public dataset, accessed 13 October 2026. <https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5>