MEMORIZED LOGITS REVERSE POST-HOC CORRECTION IN LONG-TAILED RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Two-stage methods for long-tailed recognition freeze a trained backbone and refit the classifier on the backbone's own training images. When the first stage is already class-balanced, this refit favors the head classes it is meant to rein in. The cause is memorization. A model inflates the true-class logit of its training images, most for the rarest classes, and its training-set outputs therefore overstate the tail. In a Gaussian response model we prove that a fitted correction takes its direction from the center of the logits it is fit on. Memorized logits send it toward the head, out-of-fold logits toward the tail, and the optimal strength has a closed form. Cross-fitted logit remapping therefore scores each training image with a fold model trained without it, fits a matrix-scaling correction on those logits, and folds it into the classifier head. It ranks first among all compared methods on CIFAR-LT with further gains on ImageNet-LT and iNaturalist2018, whereas the same correction fit on memorized logits underperforms logit adjustment. Where a correction is fit matters more than which correction is fit. Code is available at `https://anonymous.4open.science/r/crossfit_remap-BEC8`.

1 INTRODUCTION

Two-stage methods are the common paradigm in long-tailed recognition, the first stage learning a representation and the second correcting the decision boundary of the classifier (Kang et al., 2020). A model trained on a long-tailed training set is biased toward the head classes, which reduces its accuracy on the tail classes. The model is then evaluated on a class-balanced test set, where the tail counts as much as the head. The second stage's question is how much of that head bias to remove before deployment.

Prior-based rules answer it with a hyperparameter fixed in advance. Logit adjustment subtracts a multiple of the log class prior from each logit, moving the decision boundary off the head by an amount the counts fix and a single hyperparameter scales (Menon et al., 2021). τ-normalization divides each row of the classifier weights by its own norm raised to a power, shrinking the head rows that long-tailed training has grown, again under one hyperparameter (Kang et al., 2020). Because both are computed without any fitting, neither can see how much head bias a particular first stage has actually left behind, and neither can check the hyperparameter against a validation split, since the long-tailed training set is too sparse in the tail to set one aside.

Data-fitted rules answer it from the model itself. cRT discards the long-tailed classifier and re-trains it on class-balanced batches of the frozen features, LWS keeps that classifier and learns one scale per class (Kang et al., 2020), MiSLAS smooths the re-training labels by class frequency (Zhong et al., 2021), and MARC fits a class-specific affine margin on top of the frozen logits (Wang et al., 2023a). Each of them fits on the frozen model's outputs on its own training images. A model memorizes some of its training images, and memorization concentrates on rare and atypical examples (Feldman, 2020; Feldman & Zhang, 2020). We measure its logit-level form under class imbalance: the true-class logit of a training image is raised more the rarer its class. The model's training-set outputs thus understate the deployed model's head bias, and a correction fit on them undershoots it (Figure 1(a)). On a class-balanced first stage little head bias is left, and the same correction overshoots and reverses sign, favoring the head classes it should suppress (Figure 1(b)).

In a head–tail response model we prove that the center of the fitting responses fixes the direction of the fitted boundary. Memorization drags the center of reused responses toward the head, and the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

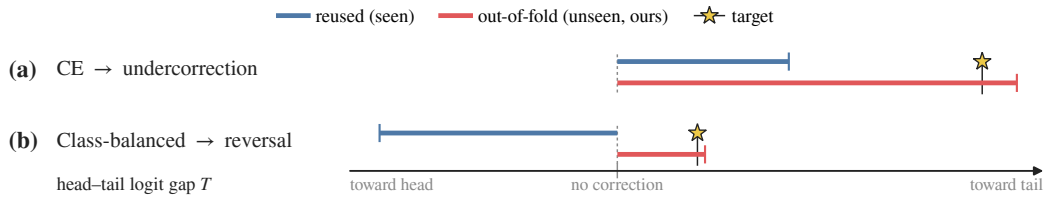

Figure 1: **A fit on memorized logits misses the target, and an out-of-fold fit reaches it.** Each track spans what the corrected rule can reach at any correction strength. A fit on responses the model has already seen misses the target both ways, moving too little on a cross-entropy base (a) and in the wrong direction on a class-balanced base (b).

correction reverses sign once that drag exceeds the head bias the first stage left behind. Out-of-fold responses keep their center on the tail side, and there the optimal strength of the correction has a closed form. The sign of the reused center is measurable before any deployment, and it predicts which of the two regimes a given first stage falls into (Figure 1). Based on this analysis, we propose *cross-fitted logit remapping*, which replaces memorized logits with out-of-fold logits as the fitting source. Changing only this source turns the same correction from a loss into a gain. The method ranks first on CIFAR-LT and transfers to ImageNet-LT and iNaturalist2018.

Our key contributions are as follows: 1) We identify the fitting source as the failure point of data-fitted second stages, which fit their correction on the model's responses to its own training images. We show that memorization understates the head bias the first stage leaves, which makes the correction undershoot and, once the understatement exceeds that bias, reverse. 2) In a head–tail response model we prove that the direction of the fitted correction is set by the center of the responses it is fit on, establish the condition for reversal, and derive a closed form for the optimal strength of the out-of-fold correction. 3) We propose cross-fitted logit remapping, a second stage that fits the correction on out-of-fold logits instead of memorized ones. It ranks first among all compared methods on CIFAR-LT, and adding it to GLMC (Du et al., 2023) improves accuracy on ImageNet-LT and iNaturalist2018.

2 Related Work

Two-stage correction in long-tailed recognition. Two-stage methods freeze the backbone learned in the first stage and correct the decision boundary. They fall into two categories depending on whether the correction is fit on data. Prior-based rules take it from the class counts or the classifier weights. Logit adjustment shifts logits by the log class prior and τ-normalization rescales the weight norms (Menon et al., 2021; Kang et al., 2020), and both tune only the strength. Data-fitted rules instead fit on the first-stage model's outputs on its own training images. cRT and LWS retrain or rescale the classifier head (Kang et al., 2020), and DisAlign aligns the corrected distribution to a reference (Zhang et al., 2021a). MiSLAS adds label-aware smoothing to the re-training labels (Zhong et al., 2021) and LOS over-smooths them (Sun et al., 2025). MARC fits a class-specific affine margin (Wang et al., 2023a), Prior2Posterior re-estimates the model's implicit prior (Bhat et al., 2025), and SAMN learns per-class scale factors under a monotonic constraint on the weight norms (Zhang et al., 2026). Each rule works on the first stage it was designed for, while the data-fitted ones share the premise that the fitting responses come from images the model has seen. We show that the sign of the gain follows that source rather than the rule, and that it reverses on class-balanced first stages (Theorem 1, Figure 3 and Table 5).

Out-of-fold fitting in calibration. Affine logit calibration in the multi-class case is matrix scaling (Guo et al., 2017), to which Kull et al. (2019) added off-diagonal and intercept regularization (ODIR), with its strength set by an internal cross-validation. Platt scaling fits its sigmoid on held-out cross-validation outputs rather than on the training outputs (Platt, 1999). Stacked generalization fits a second-level model on out-of-fold predictions and applies it to the model trained on all folds (Wolpert, 1992). Each fits the correction on the model's responses to images it has not seen. Double machine learning cross-fits its nuisance estimates on the same principle (Chernozhukov et al., 2018). Stratified K-fold predictions appear in long-tailed recognition as a way to estimate class and sample

difficulty for first-stage training (Sun et al., 2026). Their second stages still fit the correction on training-set outputs. We fit the second stage on out-of-fold responses and analyze how that choice moves the deployed decision boundary.

Memorization and the train–test response gap. Feldman (2020); Feldman & Zhang (2020) define memorization as the leave-one-out drop in the probability that the trained classifier predicts an example's own label, and show that this drop concentrates on rare and atypical examples. Under class imbalance, Ye et al. (2020) identify a feature deviation between training and test data. We measure the logit-level counterpart of this gap as a per-class reuse increment and derive its effect on the direction of the second-stage fit.

3 Cross-fitted logit remapping

3.1 Setting and correction family

We adopt the label-shift setting with C classes, in which training and test share the class-conditional distributions $P_c = P(x \mid y = c)$ and differ only in the class prior, proportional to the class counts $n_1 \geq \cdots \geq n_C$ in training and uniform at test. Under this shift, the target is the class-balanced risk (Menon et al., 2021). The training algorithm A trains on the construction sample $S_{\boldsymbol{n}}$ with class counts $\boldsymbol{n} = (n_1, \ldots, n_C)$ and returns the logit predictor $L_{\mathcal{C}}$ of the full model, where $\mathcal{C}$ collects the construction data and the training randomness. A two-stage method learns a correction on top of this frozen predictor.

The **remapper** is an affine map of the logits with parameters $G \in \mathbb{R}^{C\times C}$ and $g \in \mathbb{R}^C$. For any logit vector $z \in \mathbb{R}^C$, the deployment family and its interpolated parameters are

$$M_{\lambda,\tau}(z) = G_\lambda z + g_\lambda - \tau \log \pi, \qquad G_\lambda = (1-\lambda) I_C + \lambda G, \quad g_\lambda = \lambda g, \qquad \lambda \in [0,1],\ \tau \geq 0, \tag{1}$$

where λ is the shrinkage coefficient, τ the logit-adjustment coefficient, and π the training class prior, $\pi_c = n_c / \sum_{c'} n_{c'}$, with $\log \pi$ taken coordinatewise. Setting $\lambda = 0$ keeps the base model and the logit-adjustment shift, and $\lambda = 1$ applies the full remapper, which we call the **endpoint**.

Let $\ell_{01}(z, c) = \mathbf{1}\{\mathrm{top}(z) \neq c\}$ be the top-1 zero-one loss, with $\mathrm{top}(z)$ the top-1 class of z under a fixed tie order, and ℓ_{CE} the cross-entropy loss. For a loss ℓ and a logit table $\mathcal{Z}$ with N_c rows per class, whose i-th row carries the logit vector z_i and the label y_i, the class-balanced empirical risk of a remapper M on $\mathcal{Z}$ and the deployment risk of M are

$$\widehat{R}^{\ell}_{\mathcal{Z}}(M) = \frac{1}{C} \sum_c \frac{1}{N_c} \sum_{i: y_i = c} \ell(M(z_i), c), \qquad R^{\ell_{01}}_f(M) = \frac{1}{C} \sum_c \int \ell_{01}\big(M(L_{\mathcal{C}}(x)), c\big)\, P_c(dx), \tag{2}$$

where the subscript f marks the response source as fresh. The question is which responses (G, g) should be fit on to make $R^{\ell_{01}}_f$ small.

3.2 Reused and fresh responses

The responses a correction is fit on determine which classification errors it can correct. Existing second-stage rules fit on the model's outputs on its own training images, while at deployment the model faces unseen images. Under long-tailed training, these two sets of outputs differ. To compare them, we call the output of a model on an image it was trained on the **reused logit** of that image, and the output of a model that was not trained on the image its **fresh logit**. Both are measured by the true-class contrast of a logit vector, the true-class logit minus the largest of the remaining logits. The **reuse increment** κ_c of class c is then the expected amount by which the reused true-class contrast exceeds the fresh one.

In Figure 2(a), the reused and fresh true-class contrasts nearly coincide on the many classes and separate as the class count falls. Figure 2(b) shows the same order under all three settings, with the increment κ_c growing from many-shot to few-shot classes in each. A correction fit on reused responses therefore faces a spread across the classes, not a constant offset that would leave the head–tail order intact. Figure 2(a) also shows the **out-of-fold** contrast, that of a fold model on a training image it was not trained on, tracking the test contrast of the full model across the class counts. Section 3.3 builds the fitting table from such models.

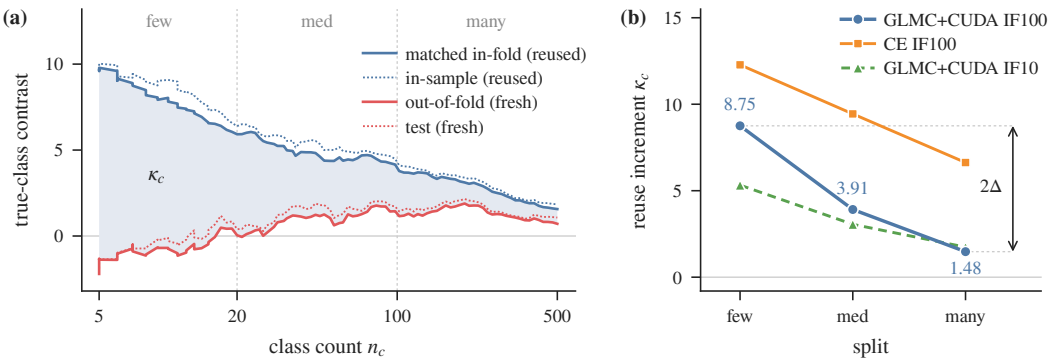

Figure 2: **True-class contrast of reused and fresh logits.** (a) Per-class contrast against n_c on CIFAR-100-LT under GLMC+CUDA, as a 9-class moving average. Blue marks the reused logits (in-sample, matched in-fold), red the fresh logits (out-of-fold, test), and the band between the solid lines is κ_c. (b) Split means of κ_c under three settings, the bracket spanning $2\Delta = \kappa_{\text{few}} - \kappa_{\text{many}}$.

3.3 Procedure

First, split the training set into $K \geq 2$ class-stratified folds, whose held-out counts of each class differ by at most one. The k-th fold model L_k and the out-of-fold logit z_i^o of sample i are

$$L_k = A(S_{\boldsymbol{n}} \setminus S_k), \qquad z_i^o = L_{k(i)}(x_i), \tag{3}$$

where S_k is the k-th fold and $k(i)$ the fold of sample i.

Next, the remapper is fit on the out-of-fold table $\mathcal{Z}^o$, whose rows are (z_i^o, y_i), by class-balanced matrix scaling with ODIR (Guo et al., 2017; Kull et al., 2019). The finite-sample fitting objective J is

$$J(G, g) = \widehat{R}^{\ell_{\text{CE}}}_{\mathcal{Z}^o}(z \mapsto Gz + g) + \xi_{\text{od}}\|G - \text{diag}(G)\|_F^2 + \xi_b\|g\|_2^2, \tag{4}$$

where ξ_{od} and ξ_b are the off-diagonal and intercept penalty coefficients. The pair $(\widehat{G}, \widehat{g})$ is the output of finite-step first-order optimization on this objective.

Substituting $(\widehat{G}, \widehat{g})$ for (G, g) in equation 1 gives one candidate remapper $M_{\lambda,\tau}$ for each value of (λ, τ). To select among the candidates, an inner K-fold split of the out-of-fold table holds rows out from the fit, and $(\widehat{\xi}_{\text{od}}, \widehat{\xi}_b, \widehat{\lambda}, \widehat{\tau})$ is chosen jointly on fixed grids, with $\lambda = 0$ and $\lambda = 1$ among the candidates, by the balanced error $\widehat{R}^{\ell_{01}}$ on those rows. The held-out rows come from the same fold models as the fitting rows, so both sides of the split follow the out-of-fold distribution. The remapper is then refit on all out-of-fold logits and folded into the linear head, leaving the deployed model with the parameter count of the base model.

4 Theoretical Analysis

A second-stage correction is fit on one set of responses and deployed on fresh ones, and we locate its deployed boundary in a head–tail Gaussian model. Lemma 1 shows that a fitted boundary takes its direction from the center of the responses it is fit on. Theorem 1 shows that memorization drags the center of reused responses toward the head and reverses the correction once the drag exceeds the head bias. Theorem 2 and Corollary 1 show that out-of-fold responses centered on the tail side restore the tailward direction and that shrinkage sets how far the deployed boundary shifts. All proofs are in Appendix B.

Consider a head–tail binary task with class 1 the head, class 2 the tail and $n_1 > n_2$. Let the full model output two class logits $\tilde{z}_1, \tilde{z}_2$ on an image, and write $T = \tilde{z}_1 - \tilde{z}_2$, whose sign decides the binary top-1 prediction. On the input $z = v(T)$ with $v(t) = (t/2, -t/2)^\top$, the full affine remapper of Section 3.1 acts as an affine map $aT + b$ of the contrast, which for $a > 0$ predicts the head exactly

when T exceeds the threshold $t = -b/a$.

$$\begin{array}{c|cc} & Y=1 & Y=2 \\ \hline T_f & \mathcal{N}(d_f+\mu_f,\sigma_f^2) & \mathcal{N}(d_f-\mu_f,\sigma_f^2) \\ T_{\text{in}} & \mathcal{N}(d_f+\mu_f+\kappa_1,\sigma_f^2) & \mathcal{N}(d_f-\mu_f-\kappa_2,\sigma_f^2) \\ T_o & \mathcal{N}(d_o+\mu_o,\sigma_o^2) & \mathcal{N}(d_o-\mu_o,\sigma_o^2) \end{array} \tag{5}$$

where Y is the true class, d_f is the common center of the two fresh conditional means, μ_f the distance from each mean to that center, σ_f the within-class standard deviation, and d_o, μ_o, σ_o their fold-model counterparts. Under a balanced test prior the Bayes classifier splits at d_f, and we call its balanced error the Bayes error of the fresh distribution. The uncorrected full model, the identity classifier, splits at 0, and the center d_f is therefore the head bias the correction must cancel.

We call $\Delta = (\kappa_2 - \kappa_1)/2$ the memorization drag, since it moves the center of the reused source from d_f to $d_f - \Delta$. The increment grows as the class count falls (Figure 2(b), whose bracket marks 2Δ). The smaller tail count thus gives $\kappa_2 > \kappa_1$ and a drag $\Delta > 0$ toward the head. A class-balanced ridge learner realizing all three rows of equation 5 is constructed in Appendix B.1, and its increment $\kappa_c = \beta/(1+\gamma n_c)$ with $\beta, \gamma > 0$ strictly increases Δ as n_2 falls at fixed n_1.

On any source distribution, (G, g) is fit to $z = v(T)$ by the population form of equation 4, which replaces the empirical risk with the expected cross-entropy under the source distribution and a uniform class prior.

Lemma 1 (Population minimizer)**.** *Let the source distribution have a uniform class prior with $T \mid Y = 1 \sim \mathcal{N}(m+\nu, \sigma^2)$ and $T \mid Y = 2 \sim \mathcal{N}(m-\nu, \sigma^2)$, where $\nu, \sigma > 0$, m is the source center, ν is half the between-class separation and σ the standard deviation, and let the input logits be $z = v(T)$. For finite $\xi_{\text{od}}, \xi_b \geq 0$, the population form of equation 4 on this source has a finite minimizer, all its minimizers induce the same contrast $a^*T + b^*$, and $a^* > 0$. Defining $t^* = -b^*/a^*$,*

$$\begin{cases} t^* = 0, & m = 0, \\ t^* = m, \quad a^* = 2\nu/\sigma^2, & \xi_b = 0, \\ \operatorname{sign}(t^*) = \operatorname{sign}(m), \quad 0 < |t^*| < |m|, & \xi_b > 0,\ m \neq 0. \end{cases} \tag{6}$$

A fitted correction therefore takes its direction from the center of the responses it is fit on.

Lemma 1 applies to the reused source with center $d_f - \Delta$ and half-separation $\mu_f + (\kappa_1 + \kappa_2)/2$.

Theorem 1 (Reuse-induced boundary reversal)**.** *If $\Delta > d_f$, then the reused-source minimizer satisfies $a^*_{\text{in}} > 0$ and $t^*_{\text{in}} < 0$, and, taking $\tau = 0$, the deployment risk $R_f^{\ell_{01}}(\lambda)$ of $M_{\lambda,0}$ strictly increases over $\lambda \in [0,1]$, with tail error rising and head error falling, and any $\lambda > 0$ is strictly worse than the identity. On the reused distribution itself, cross-entropy is non-increasing in λ and balanced error strictly decreasing, both optimal at $\lambda = 1$.*

The reused fit keeps improving on its own source while deployment worsens, which places the fault in the responses. In all nine class-balanced first stages we measure, the memorization drag exceeds the head bias and the reused fit favors the head (Section 5.3). When the drag is smaller than the head bias, as on the cross-entropy first stage of cRT and LWS, the reused fit shifts the boundary toward the tail but short of the Bayes threshold.

The out-of-fold source has no increment. By Lemma 1, a center $d_o > 0$ gives a minimizer $a_o^*T + b_o^*$ with $a_o^* > 0$ and threshold $t_o^* = -b_o^*/a_o^* > 0$, the tailward direction that counteracts a head bias $d_f > 0$.

Theorem 2 (Boundary interpolation and optimal shrinkage)**.** *Given any finite fitted contrast $aT + b$ with $a > 0$, fix τ, take $\pi_1, \pi_2 > 0$ and write $h_\tau = \tau \log(\pi_1/\pi_2)$. The deployment threshold t_λ of $M_{\lambda,\tau}$ interpolates between the thresholds t_0 and t_1 at $\lambda = 0$ and $\lambda = 1$ through the boundary coefficient w_λ as*

$$t_0 = h_\tau, \qquad t_1 = \frac{h_\tau - b}{a}, \qquad t_\lambda = (1-w_\lambda)t_0 + w_\lambda t_1, \quad w_\lambda = \frac{\lambda a}{1-\lambda+\lambda a}. \tag{7}$$

For the fresh distribution of equation 5 and $t_1 \neq t_0$, the balanced error has a unique minimizer λ^* over $\lambda \in [0,1]$, given through its boundary coefficient w^* by

$$w^* = \operatorname{clip}_{[0,1]}\left(\frac{d_f - t_0}{t_1 - t_0}\right), \qquad \lambda^* = \frac{w^*}{a + (1-a)w^*}, \tag{8}$$

with $\mathrm{clip}_{[0,1]}$ *the projection onto* $[0,1]$. *If the two thresholds coincide, the top-1 risk does not vary with* λ. *The same conclusions hold for continuous class-conditional densities* f_c *of* $T \mid Y = c$ *with* $f_1(t) < f_2(t)$ *for* $t < d_f$ *and* $f_1(t) > f_2(t)$ *for* $t > d_f$.

The strength λ thus slides the decision boundary monotonically from logit adjustment at $\lambda = 0$ to the endpoint at $\lambda = 1$, and the best strength stops it at the Bayes threshold whenever the threshold lies between the two.

Corollary 1 (Out-of-fold specialization)**.** *Let* $d_o > 0$ *and take* $\tau = 0$ *with the out-of-fold minimizer, so that* $t_0 = 0$ *and* $t_1 = t_o^* > 0$. *If* $d_f > 0$, *every* $\lambda > 0$ *with* $t_\lambda < 2d_f$ *gives a balanced error strictly below the identity, and such a* λ *always exists; if* $0 < d_f < t_o^*$, *then* $\lambda^* \in (0,1)$ *and* $M_{\lambda^*,0}$ *attains the Bayes error of the fresh distribution; if* $d_f \geq t_o^*$, *then* $\lambda^* = 1$. *When* $\xi_b = 0$, $t_o^* = d_o$ *and*

$$0 < d_f < d_o \quad \Longrightarrow \quad \lambda^* = \frac{d_f}{a_o^* d_o + (1 - a_o^*) d_f}. \tag{9}$$

On out-of-fold logits the correction beats the uncorrected model at any positive strength that moves the boundary by less than twice the head bias, whereas the reused fit of Theorem 1 loses at any positive strength. Fold models train on fewer images and retain a slightly larger head bias than the full model, which at $\xi_b = 0$ places the best strength strictly inside $(0,1)$, as measured in Section 5.3. The interpolation also covers the finite-step output of Section 3.3 whenever its contrast has positive slope. At $\tau = 0$, selecting the strength on out-of-fold logits needs no test label, and any overshoot of the Bayes threshold is at most the extra head bias of the fold models, $\max\{d_o - d_f, 0\}$. Both endpoints are optimal on the logits they are fit on, yet once the memorization drag exceeds the head bias the two fits move the boundary in opposite directions, the reused fit toward the head and the out-of-fold fit toward the tail. The logits a correction is fit on therefore set its direction, and its strength only sets how far it goes.

5 Experiments

We evaluate on CIFAR-100-LT and CIFAR-10-LT (IF $\in \{100, 50, 10\}$) and on ImageNet-LT and iNaturalist2018, against logit adjustment (LA) and the uncorrected model. We report top-1 accuracy in percent. Table 1 averages three seeds, and later analyses fix one seed's base model. In each column the best and second-best results are in **bold** and underline, $\uparrow$ marks the gain of the remapper over its first stage, and results above the midrule are taken from the original papers and Zhang et al. (2026). Implementation details are in Appendix C, component ablations are in Appendix D.4, and further methods, including multi-expert ones, are compared in Appendix D.1.

5.1 Comparison with state-of-the-art methods

We add the remapper to three first stages and compare against state-of-the-art methods (Tables 1 and 2). On CIFAR-LT the remapper improves the three first stages in each setting, by $+0.07$ to $+2.10$, and GLMC+CUDA with the remapper ranks first throughout. On ImageNet-LT and iNaturalist2018 it improves GLMC from 56.72 to 58.31 and from 72.20 to 72.98, while many-shot accuracy on ImageNet-LT drops by 3.0 points, the head accuracy a second stage trades to cancel head bias. The gains use hyperparameters selected on the training set alone, without a validation split or a test label.

5.2 Effect of the response source

Isolating the response source. To isolate the response source, we hold the remapper and the penalty grid fixed and change only the logits it is fit on (Table 3). Both reused fits select $\lambda = 1$, consistent with Theorem 1, under which this endpoint is optimal on the reused source. The out-of-fold gain at IF100 has a 95% bootstrap confidence interval of $[+0.45, +1.13]$, and the same ordering of the three fits holds at IF10 and for each seed (Appendix D.2). The same correction underperforms LA by 5.06 points when fit on reused logits and outperforms it by 0.79 when fit on out-of-fold logits.

Second-stage rules. We then run five second-stage rules, with 100 to 25600 fitted parameters, on each of the three response sources (Figure 3). They are the per-class scale of LWS, a per-class affine

Table 1: **Comparison with state-of-the-art methods on CIFAR-100-LT and CIFAR-10-LT.**

	CIFAR-100-LT			CIFAR-10-LT		
Method	100	50	10	100	50	10
CE	47.4	52.3	67.2	80.0	84.3	91.9
cRT (Kang et al., 2020)	42.3	46.8	58.1	75.7	80.4	88.3
MiSLAS (Zhong et al., 2021)	52.7	56.6	69.7	84.2	87.2	92.7
DiVE (He et al., 2021)	45.4	51.1	62.0	—	—	—
WD+MaxNorm (Alshammari et al., 2022)	53.4	57.7	68.7	—	—	—
H2T (Li et al., 2024)	48.9	53.8	—	82.4	85.4	—
IWB (Dang et al., 2024)	53.3	56.3	—	—	—	—
DiffuLT (Shao et al., 2024)	51.5	56.3	63.8	84.7	86.9	90.7
LOS (Sun et al., 2025)	54.9	58.8	69.7	—	—	—
SAMN (Zhang et al., 2026)	57.7	64.0	74.1	88.5	91.3	95.2
GLMC (Du et al., 2023)	57.51	62.49	73.54	88.06	90.70	94.92
+ ours	58.22 ↑0.71	63.40 ↑0.91	74.16 ↑0.62	88.90 ↑0.84	91.15 ↑0.45	94.99 ↑0.07
GLMC + MaxNorm	57.85	62.08	73.28	86.84	90.00	94.67
+ ours	58.31 ↑0.46	63.29 ↑1.21	74.18 ↑0.90	88.94 ↑2.10	91.20 ↑1.20	95.03 ↑0.36
GLMC+CUDA	59.77	64.34	75.14	88.05	91.64	95.35
+ ours	**61.53** ↑1.76	**65.90** ↑1.56	**76.35** ↑1.21	**89.49** ↑1.44	**92.26** ↑0.62	**95.45** ↑0.10

Table 2: **Comparison with state-of-the-art methods on ImageNet-LT and iNaturalist2018.**

	ImageNet-LT				iNaturalist2018			
Method	Many	Med.	Few	All	Many	Med.	Few	All
CE	65.9	37.5	7.7	44.4	72.2	63.0	57.2	61.7
cRT (Kang et al., 2020)	61.8	46.2	27.4	49.6	69.0	66.0	63.2	65.2
MiSLAS (Zhong et al., 2021)	61.7	51.3	35.8	52.7	**73.2**	72.4	70.4	71.6
DiVE (He et al., 2021)	64.1	50.4	31.5	53.1	70.6	70.0	67.6	69.1
WD+MaxNorm (Alshammari et al., 2022)	62.5	50.4	41.5	53.9	71.2	70.4	69.7	70.2
H2T (Li et al., 2024)	62.4	48.8	**52.2**	54.6	67.7	71.9	72.2	71.6
IWB (Dang et al., 2024)	64.2	52.2	40.2	55.2	72.3	70.6	72.5	71.5
DiffuLT (Shao et al., 2024)	63.3	55.6	39.4	56.4	—	—	—	—
LOS (Sun et al., 2025)	63.2	50.7	42.3	54.4	69.2	70.7	71.3	70.8
SAMN (Zhang et al., 2026)	**69.7**	56.1	31.1	57.7	64.4	73.7	73.6	72.7
GLMC (Du et al., 2023)	68.3	54.3	32.4	56.72	65.7	73.6	72.1	72.20
+ ours	65.3	**57.0**	43.1	**58.31** ↑1.59	65.1	**73.8**	**74.0**	**72.98** ↑0.78

map of the form MARC fits, our full remapper and the retrained head of cRT with and without the weight-norm cap of MaxNorm. Each reused fit underperforms LA, by 1.79 to 7.23 points, and each out-of-fold fit outperforms it, by 0.16 to 0.88 (Figure 3(a)). On few-shot classes the reused fits fall to between 19.0 and 32.2 and the out-of-fold fits reach 40.7 to 48.0, on opposite sides of LA (Figure 3(b)). The sign of the gain follows the response source and not the rule.

Class-wise direction. We next ask which classes the reused fit moves the boundary toward as we strengthen the correction from LA at $\lambda = 0$ to the fitted endpoint at $\lambda = 1$ (Figure 4). The test top-1 of the in-sample fit drops below LA as λ grows (Figure 4(a)), yet its cross-entropy on the in-sample table falls from 0.331 to 0.001 (Figure 4(b)), as Theorem 1 predicts on the reused source. The out-of-fold fit raises few-shot accuracy at any $\lambda > 0$ (Figure 4(c)), whereas the in-sample fit raises many-shot accuracy steadily to 8.8 points above LA and lowers few-shot accuracy to 19.4 points below it (Figure 4(d)). A reused fit moves the boundary toward the head, favoring the classes the second stage is meant to suppress.

Table 3: **Top-1 accuracy of the same correction fit on each logit table.** CIFAR-100-LT under GLMC+CUDA, and vs LA is the gain over the LA row. In-sample and matched in-fold logits come from the full model and a fold model on images they trained on.

IF	Logit table	Top-1	vs LA
100	LA	60.39	—
100	Out-of-fold	**61.18**	**+0.79**
100	In-sample	55.33	−5.06
100	Matched in-fold	55.54	−4.85
10	LA	75.99	—
10	Out-of-fold	**76.53**	**+0.54**
10	In-sample	73.07	−2.92
10	Matched in-fold	73.07	−2.92

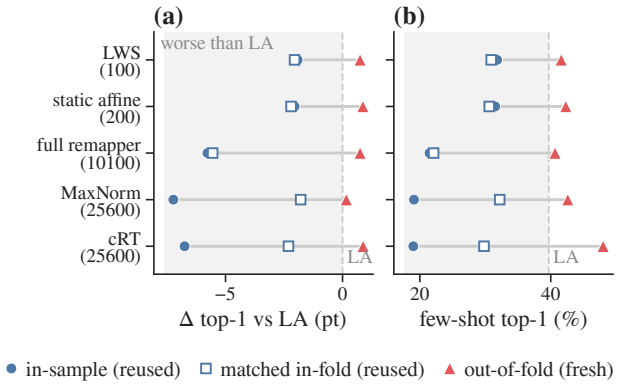

Figure 3: **Five second-stage rules fit on each logit table.** CIFAR-100-LT under GLMC+CUDA, fitted parameters in parentheses. (a) Top-1 gain over LA. (b) Few-shot top-1.

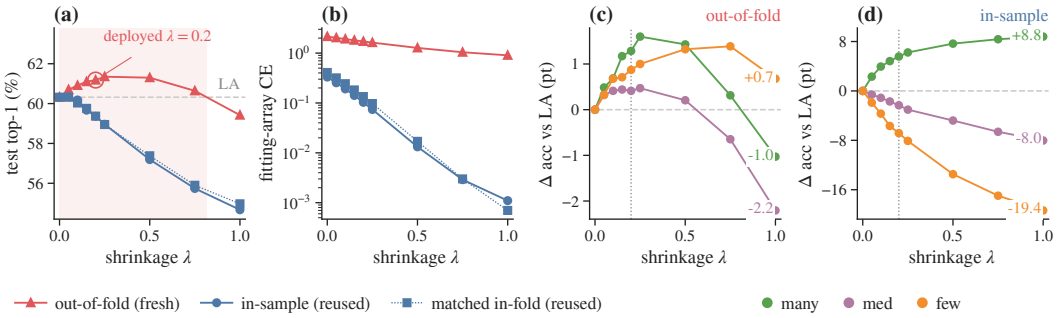

Figure 4: **Shrinkage paths of the three fits.** (a) Test top-1 against λ, with LA dashed and the deployed λ circled. (b) Cross-entropy of each fit on its own logit table. (c, d) Per-split accuracy against LA for the out-of-fold and the in-sample fit, with the value at $\lambda = 1$ at each line end and the dotted line marking the deployed λ.

5.3 Empirical validation of the theory

Inflation by response source. We first measure how much each response source overstates the true-class logit, split by split (Table 4). Down the four rows the few-shot true-class logit falls monotonically from 16.95 to 2.80, compared with 3.69 on the test set, and the gain over LA rises monotonically from -5.06 to $+0.79$. The fold model that retrains only its last block, whose frozen layers were trained on the held-out folds, lands in between. The two full-retrain rows come from the same fold models, so their difference isolates the effect of reuse, 12.8 on few-shot classes versus 0.9 on many-shot classes. The rarer the class, the more a model raises the true-class logit of its own training images, and the deployed gain rises as this inflation falls.

Source centers. By Lemma 1, a positive center of the fitting logits moves the boundary toward the tail and a negative one toward the head. Most out-of-fold tail images lose to another class, with center $d_o = 1.05$ near $d_f = 0.96$ (Figure 5(a)). The model classifies its tail training images correctly by a wide margin, and this memorization pulls the matched in-fold center to -2.59. The reused fit thus moves the boundary toward the head in all 1085 many-by-few pairs and predicts more tail images as head, while the out-of-fold fit places it near the optimum d_f (Figure 5(b)).

Reversal criterion. Theorem 1 predicts a reversal once half the spread Δ of the reuse increment exceeds the head bias d_f, that is, once the in-sample center $d_f - \Delta$ is negative, and this center is measured on the training set alone. We test the prediction against the test top-1 of the reused fit at $\lambda = 1$ and $\tau = 0$ (Figure 6(a)). All 27 class-balanced settings have a negative in-sample center,

Table 4: **True-class logit by response source.** CIFAR-100-LT under GLMC+CUDA. Many, Medium and Few are split means, and vs LA is the top-1 gain of the fit on each source over LA.

Response source	Fold retrain	Many	Medium	Few	vs LA
In-sample	—	4.51	9.23	16.95	−5.06
Matched in-fold	full	4.17	8.46	15.60	−4.85
Out-of-fold	last block	3.01	5.96	7.18	−1.32
Out-of-fold	full	3.30	4.20	2.80	**+0.79**
Test	—	3.69	4.86	3.69	—

Table 5: **cRT and LWS on a cross-entropy first stage.** LA uses the τ that is best on the test set and serves as an upper reference.

Method	Plain	+ remapper
CE	38.77	—
cRT	41.11	**42.47**
LWS	40.66	**42.19**
LA (upper ref.)	42.23	—



Figure 5: **Tail contrast and fitted thresholds.** (a) Distribution of the contrast T on tail images, the arrow joining the out-of-fold mean to the matched in-fold mean. (b) Fitted thresholds over many-by-few class pairs. The dashed line at zero is the split of the uncorrected model, and a threshold above zero favors the tail.

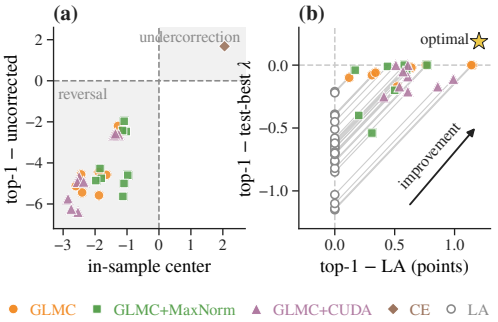

Figure 6: **Reversal criterion and shrinkage selection on CIFAR-100-LT.** (a) Top-1 change of the reused fit over the uncorrected model against the in-sample center, per first stage and seed. Shading marks predicted regimes. (b) Top-1 relative to LA and to the test-best λ per class-balanced first stage and seed.

and in each the reused fit lowers top-1 by 1.97 to 6.39 points, raising many-shot and lowering few-shot accuracy. The cross-entropy first stage has an in-sample center of $+2.05$ inside $(0, d_f)$ with $d_f = 4.37$, and its reused fit raises top-1 by 1.68 points, with few-shot accuracy up 4.29 and many-shot down 1.46. The reused fit thus corrects in the right direction but short of d_f, as cRT and LWS do relative to LA, and the out-of-fold map closes the remaining gap (Table 5, reused fits in Appendix D.3).

Optimal shrinkage and its selection. Theorem 2 gives $\lambda^* \in (0, 1)$ when d_f lies strictly between the endpoint thresholds t_0 and t_1, and the test-best λ is interior in 20 of the 27 class-balanced settings. We select λ on the out-of-fold source without a test label (Figure 6(b)). The selected strength trails the test-best one by 0.10 points on average and at most 0.54, and it outperforms LA in all 27 settings, by 0.56 on average.

6 Conclusion

Data-fitted second stages in long-tailed recognition fit on the first-stage model's outputs on its own training images, and memorization makes those outputs overstate the tail by an amount that grows with class rarity. We proved that the direction of the fitted correction is the sign of the center of the fitting responses, which sends a reused fit toward the head and an out-of-fold fit toward the tail, and we gave a closed form for the optimal shrinkage of the out-of-fold correction. Scoring each training image with a fold model that excludes it, fitting matrix scaling with ODIR on those logits and folding the result into the classifier head ranks first among all compared methods on CIFAR-LT and transfers to ImageNet-LT and iNaturalist2018. A data-fitted second stage should be fit on logits of images outside the training set of the model that produces them.

AI USE STATEMENT

In this work, we used generative AI tools to assist with searching for and identifying relevant literature, writing and refining software code, improving the readability and clarity of the manuscript, and retrieving and organizing information during the research and review process. We did not use generative AI tools to generate synthetic datasets, develop theoretical models, formulate or prove mathematical claims, design research methodology or experiments, interpret experimental results, or perform other research tasks requiring mandatory disclosure, where applicable.

All AI-assisted outputs were independently reviewed and verified by the authors. In particular, information obtained with the assistance of generative AI was checked against the original sources, AI-assisted code was reviewed and tested for correctness, and all manuscript edits were examined and revised by the authors as necessary. We take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendix B restates the assumptions and gives complete proofs of the results in Section 4. Appendix C lists the training schedules, remapper fitting and selection settings, and measurement protocols. Anonymous source code is available at `https://anonymous.4open.science/r/crossfit_remap-BEC8`.

REFERENCES

Sumyeong Ahn, Jongwoo Ko, and Se-Young Yun. CUDA: Curriculum of data augmentation for long-tailed recognition. In *International Conference on Learning Representations*, 2023.

Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

S Divakar Bhat, Amit More, Mudit Soni, and Surbhi Agrawal. Prior2posterior: Model prior correction for long-tailed learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1289–1298. IEEE, 2025.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.

Xiaohua Chen, Yucan Zhou, Dayan Wu, Chule Yang, Bo Li, Qinghua Hu, and Weiping Wang. AREA: Adaptive reweighting via effective area for long-tailed classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. doi: 10.1109/ICCV51070.2023.01766.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Wenqi Dang, Zhou Yang, Weisheng Dong, Xin Li, and Guangming Shi. Inverse weight-balancing for deep long-tailed learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. doi: 10.1609/aaai.v38i10.29055.

Yuan Dong, Zhe Zhao, Liheng Yu, Di Wu, and Pengkun Wang. GUIDE: Gated uncertainty-informed disentangled experts for long-tailed recognition. In *International Conference on Learning Representations*, 2026.

Fei Du, Peng Yang, Qi Jia, Fengtao Nan, Xiaoting Chen, and Yun Yang. Global and local mixture consistency cumulative learning for long-tailed visual recognitions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Vitaly Feldman. Does learning require memorization? A short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–959, 2020. doi: 10.1145/3357713.3384290.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems*, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Naoya Hasegawa and Issei Sato. Exploring weight balancing on long-tailed recognition problem. In *International Conference on Learning Representations*, 2024.

Yin-Yin He, Jianxin Wu, and Xiu-Shen Wei. Distilling virtual examples for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems*, 2019.

Mengke Li, Zhikai Hu, Yang Lu, Weichao Lan, Yiu-ming Cheung, and Hui Huang. Feature fusion from head to tail for long-tailed visual recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Gao Peifeng, Qianqian Xu, Peisong Wen, Zhiyong Yang, Huiyang Shao, and Qingming Huang. Feature directions matter: Long-tailed learning via rotated balanced representation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27542–27563, 2023.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans (eds.), *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, and R. Venkatesh Babu. Escaping saddle points for effective generalization on class-imbalanced data. In *Advances in Neural Information Processing Systems*, 2022.

Emanuel Sanchez Aimar, Arvi Jonnarth, Michael Felsberg, and Marco Kuhlmann. Balanced product of calibrated experts for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Jie Shao, Ke Zhu, Hanxiao Zhang, and Jianxin Wu. Diffult: Diffusion for long-tail recognition without external knowledge. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 123007–123031. Curran Associates, Inc., 2024. doi: 10.52202/079017-3909.

Jiang-Xin Shi, Tong Wei, Yuke Xiang, and Yu-Feng Li. How re-sampling helps for long-tail learning? In *Advances in Neural Information Processing Systems*, 2023.

Siyu Sun, Han Lu, Jiangtong Li, Yichen Xie, Tianjiao Li, Xiaokang Yang, Liqing Zhang, and Junchi Yan. Rethinking classifier re-training in long-tailed recognition: Label over-smooth can balance. In *International Conference on Learning Representations*, 2025.

Tianci Sun, Wanqiu He, Changbin Shao, Shang Zheng, and Hualong Yu. Acvm: An adaptive combination validation mechanism for long-tailed image recognition. *Information*, 17(5), 2026. ISSN 2078-2489. doi: 10.3390/info17050455.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X. Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021.

Yidong Wang, Bowen Zhang, Wenxin Hou, Zhen Wu, Jindong Wang, and Takahiro Shinozaki. Margin calibration for long-tailed visual recognition. In *Proceedings of the 14th Asian Conference on Machine Learning*, volume 189 of *Proceedings of Machine Learning Research*, pp. 1101–1116, 2023a.

Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. A unified generalization analysis of re-weighting and logit-adjustment for imbalanced learning. In *Advances in Neural Information Processing Systems*, 2023b.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.

Han-Jia Ye, Hong-You Chen, De-Chuan Zhan, and Wei-Lun Chao. Identifying and compensating for feature deviation in imbalanced deep learning. *arXiv preprint arXiv:2001.01385*, 2020.

Shuo Zhang, Chenqi Li, and Tingting Zhu. Why not hyperparameter-friendly optimisation? a monotonic adaptive norm rescaling approach for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36947–36956, June 2026.

Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021a.

Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. In *Advances in Neural Information Processing Systems*, 2022.

Yongshun Zhang, Xiu-Shen Wei, Boyan Zhou, and Jianxin Wu. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021b. doi: 10.1609/aaai.v35i4.16458.

Qihao Zhao, Chen Jiang, Wei Hu, Fan Zhang, and Jun Liu. MDCS: More diverse experts with consistency self-distillation for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Appendix Contents

A Notation **13**

B Response reuse, boundary reversal and shrinkage **15**

B.1 A learner realizing class-dependent reuse 16

B.2 Population affine fitting with ODIR 17

B.3 Proof of Theorem 1 19

B.4 Proof of Theorem 2 and Corollary 1 21

C Experimental setup **22**

C.1 Training 22

C.2 Remapper fit and selection 23

C.3 Measurement protocols 23

D Additional experimental results **23**

D.1 Comparison with further methods 23

D.2 The reversal across seeds, images and rules 25

D.3 Mechanism analysis 26

D.4 Components 27

E Limitations **27**

A NOTATION

This appendix lists the notation. Table 6 gives each symbol with its meaning and the place in the paper where it is defined, grouped as the problem setting and data, the correction, its deployment and the risk targets, and the binary response model.

Table 6: Notation.

Symbol	Meaning	Defined in
Problem setting and data		
C	number of classes	§3.1
c, c'	class index; $c' \neq c$ is another class	§3.1
n_c, $\boldsymbol{n}$	training class count and count vector, $n_1 \geq \cdots \geq n_C$	§3.1
π, π_1, π_2	training class prior and its two binary components	§3.1, §4
x, y	input and label	§3.1
P_c	class-conditional distribution	§3.1
$S_{\boldsymbol{n}}$	construction sample	§3.1
A	training algorithm	§3.1
K, k, S_k	number of folds, fold index, k-th fold ($S_{-k} = S_{\boldsymbol{n}} \setminus S_k$)	§3.3
$k(i)$	fold containing sample i	§3.3
L, $L_{\mathcal{C}}$, L_k	logit predictor: generic, full model, fold model	§3.1, §3.3
z, z_i, y_i	logit vector; logit and label of row i	§3.1
z_i^o	out-of-fold logit of sample i	§3.3
$\mathcal{Z}$, N_c	logit table, one row per image; N_c rows of class c	§3.1
$\mathcal{Z}^o$	out-of-fold logit table, rows (z_i^o, y_i)	§3.3
I_C, I_2	identity matrix	§3.1

Symbol	Meaning	Defined in
Correction, deployment and risk		
G, g	affine remapper parameters	§3.1
$\widehat{G}$, $\widehat{g}$	fitted output of the finite-step solver	§3.3
$M_{\lambda,\tau}$, G_λ, g_λ	deployment family and its interpolated parameters	§3.1
λ	shrinkage coefficient, $\lambda \in [0,1]$	§3.1
τ	logit-adjustment coefficient	§3.1
Λ, $\mathcal{T}$	shrinkage grid and logit-adjustment grid	App. C.2
$\widehat{\lambda}, \widehat{\tau}, \widehat{\xi}_{\mathrm{od}}, \widehat{\xi}_b$	values returned by selection	§3.3
ξ_{od}, ξ_b	ODIR off-diagonal and intercept penalty coefficients	§3.3
$J(G,g)$	finite-sample fitting objective	§3.3
$\mathrm{top}(z)$	top-1 class under a fixed tie order	§3.1
ℓ, ℓ_{01}, ℓ_{CE}	generic loss, top-1 zero-one loss, cross-entropy	§3.1
$\widehat{R}^{\ell}_{\mathcal{Z}}(M)$	class-balanced empirical risk on the logit table $\mathcal{Z}$	§3.1
$R_f^{\ell_{01}}(M)$, $R_f^{\ell_{01}}(\lambda,\tau)$	deployment risk: fresh class-balanced risk, as a function of the rule or of (λ,τ)	§3.1, §4
Binary response model and population minimizer		
$\tilde{z}_1, \tilde{z}_2, T$	the two logits of the full model and the contrast $T = \tilde{z}_1 - \tilde{z}_2$	§4
$v(t) = (t/2, -t/2)^\top$	maps a contrast back to a symmetric pair of logits	§4
T_f, T_{in}, T_o	fresh, reused and out-of-fold contrast	§4
d_f, μ_f, σ_f	common center, half-separation and within-class standard deviation of the fresh distribution	§4
d_o, μ_o, σ_o	the same quantities on the fold models	§4
κ_c	reuse increment of class c	§3.2, §4
$\Delta = (\kappa_2 - \kappa_1)/2$	half the difference of the two reuse increments	§4
β, γ	size and decay parameters of the reuse increment	§4, App. B.1
m, ν, σ	source center, half-separation and standard deviation in Lemma 1	§4
a, b, $aT + b$	slope and intercept of the fitted contrast	§4
a^*, b^*, t^*	population minimizer of Lemma 1 and its threshold $t^* = -b^*/a^*$	§4
$a^*_{\mathrm{in}}, b^*_{\mathrm{in}}, t^*_{\mathrm{in}}$	the minimizer on the reused source	§4
a^*_o, b^*_o, t^*_o	the minimizer on the out-of-fold source	§4
t_λ, w_λ	deployment threshold at shrinkage λ and its boundary coefficient	§4
t_0, t_1	thresholds at $\lambda = 0$ and $\lambda = 1$	§4
$h_\tau = \tau \log(\pi_1/\pi_2)$	action of the logit-adjustment shift on the contrast	§4
λ^*, w^*	optimal shrinkage coefficient on $[0,1]$ and its boundary coefficient	§4
$F(t; m, \nu, \sigma)$	balanced error of the threshold-t rule on $\mathcal{N}(m \pm \nu, \sigma^2)$	App. B
$F_f(t)$, $F_{\mathrm{in}}(t)$	that balanced error on the fresh and on the reused distribution	App. B
f_1, f_2	generic class densities	§4, App. B.4
(c, c'), $T = z_c - z_{c'}$	head–tail class pair and the contrast on that pair	§5.3, App. C.3

B RESPONSE REUSE, BOUNDARY REVERSAL AND SHRINKAGE

This appendix proves every result of Section 4. Appendix B.1 gives the learner that realizes the distributions of Section 4. Appendix B.2 proves Lemma 1, Appendix B.3 proves Theorem 1, and Appendix B.4 proves Theorem 2 and Corollary 1.

Setting. *Binary response model.* Class 1 is the head and class 2 the tail with $n_1 > n_2$, $Y \in \{1, 2\}$ is the true class and $Y_\pm = 2\mathbf{1}\{Y = 1\} - 1$. The model outputs two logits $\tilde{z}_1, \tilde{z}_2$ with contrast $T = \tilde{z}_1 - \tilde{z}_2$, and top-1 is decided by the sign of T. The map $v(t) = (t/2, -t/2)^\top$ sends a contrast back to a symmetric pair of logits and the remapper input is $z = v(T)$. The fresh contrast T_f is the output of the full model on images that took no part in its training, the reused contrast T_{in} its output on its own training images, and the out-of-fold contrast T_o the output of a fold model on training images it did not see. Their conditional distributions are

$$\begin{array}{c|cc} & Y = 1 & Y = 2 \\ \hline T_f & \mathcal{N}(d_f + \mu_f, \sigma_f^2) & \mathcal{N}(d_f - \mu_f, \sigma_f^2) \\ T_{\text{in}} & \mathcal{N}(d_f + \mu_f + \kappa_1, \sigma_f^2) & \mathcal{N}(d_f - \mu_f - \kappa_2, \sigma_f^2) \\ T_o & \mathcal{N}(d_o + \mu_o, \sigma_o^2) & \mathcal{N}(d_o - \mu_o, \sigma_o^2) \end{array} \qquad \mu_f, \sigma_f, \mu_o, \sigma_o > 0,\ d_f \geq 0,\ \kappa_c \geq 0, \tag{10}$$

where d_f, the Bayes threshold under a balanced prior, is the common center of the two conditional means, μ_f the distance from each mean to that center, σ_f the within-class standard deviation, κ_c the reuse increment of the true-class contrast of a class-c training image, and d_o, μ_o, σ_o the fold-model counterparts of d_f, μ_f, σ_f. Let $\mathcal{C}$ be the random element with respect to which the construction data, the training randomness and the fitted outputs are all measurable, and $\mathcal{F} = \sigma(\mathcal{C})$. The first row is conditional on $\mathcal{F}$ and the other two are marginal over the generation of the training sample and the within-class draw. Half the difference Δ of the two reuse increments and the center m_{in} and half-separation ν_{in} of the reused distribution are

$$\Delta = \frac{\kappa_2 - \kappa_1}{2}, \qquad m_{\text{in}} = d_f - \Delta, \quad \nu_{\text{in}} = \mu_f + \frac{\kappa_1 + \kappa_2}{2}, \tag{11}$$

so that the reused distribution is $\mathcal{N}(m_{\text{in}} \pm \nu_{\text{in}}, \sigma_f^2)$. Lemma 1 is stated for a general center m, half-separation $\nu > 0$ and standard deviation $\sigma > 0$, with $T \mid Y = 1 \sim \mathcal{N}(m + \nu, \sigma^2)$, $T \mid Y = 2 \sim \mathcal{N}(m - \nu, \sigma^2)$ and a uniform class prior.

Affine map and contrast. The correction has parameters $G \in \mathbb{R}^{2\times 2}$ and $g \in \mathbb{R}^2$ with output $Gz + g$, whose coordinate difference is again affine in T with slope a and intercept b:

$$(Gz + g)_1 - (Gz + g)_2 = aT + b, \qquad a = \frac{G_{11} - G_{12} - G_{21} + G_{22}}{2}, \qquad b = g_1 - g_2, \tag{12}$$

derived at equation 27. The corrected top-1 is the sign of $aT + b$. For $a > 0$ it predicts class 1 exactly when $T > t := -b/a$, and we call t the threshold. The identity $(I_2, 0)$ has $a = 1$, $b = 0$ and threshold 0.

Population objective. On any source distribution, (G, g) is fit by balanced population cross-entropy with the ODIR penalty Ω,

$$\min_{G,g}\ \mathbb{E}\,\ell_{\text{CE}}(Gv(T) + g, Y) + \Omega(G, g), \qquad \Omega(G, g) = \xi_{\text{od}}\|G - \text{diag}(G)\|_F^2 + \xi_b\|g\|_2^2, \tag{13}$$

with $\xi_{\text{od}}, \xi_b \geq 0$ and $\ell_{\text{CE}}(z, y) = -\log(e^{z_y}/\sum_c e^{z_c})$. Its minimizer is written (G^*, g^*), its contrast $a^*T + b^*$ and its threshold $t^* = -b^*/a^*$. Quantities on the reused and out-of-fold sources carry the subscripts in and o.

Deployment family and thresholds. With (G^*, g^*) as endpoint, the deployment family of Section 3.1 and its interpolated parameters are

$$M_{\lambda,\tau}(z) = G_\lambda z + g_\lambda - \tau \log \pi, \qquad G_\lambda = (1 - \lambda)I_2 + \lambda G^*, \quad g_\lambda = \lambda g^*, \tag{14}$$

with $\lambda \in [0, 1]$, $\tau \geq 0$, training prior $\pi = (\pi_1, \pi_2)$ and $h_\tau = \tau \log(\pi_1/\pi_2)$ the action of the logit-adjustment shift on the contrast. The output contrast of $M_{\lambda,\tau}$ is $(1 - \lambda + \lambda a^*)T + \lambda b^* - h_\tau$. Its threshold t_λ satisfies

$$t_\lambda = (1 - w_\lambda)t_0 + w_\lambda t_1, \qquad t_0 = h_\tau, \quad t_1 = \frac{h_\tau - b^*}{a^*}, \qquad w_\lambda = \frac{\lambda a^*}{1 - \lambda + \lambda a^*}, \tag{15}$$

where t_0 and t_1 are the thresholds at $\lambda = 0$ and $\lambda = 1$ and w_λ is the boundary coefficient. Appendix B.4 derives this interpolation. At $\tau = 0$ this gives $t_0 = 0$, $t_1 = t^*$ and $t_\lambda = w_\lambda t^*$.

Balanced error. For the threshold rule "predict class 1 when $T > t$" on a general distribution $\mathcal{N}(m \pm \nu, \sigma^2)$, the balanced error is

$$F(t; m, \nu, \sigma) = \frac{1}{2}\left[\Phi\left(\frac{t - m - \nu}{\sigma}\right) + \Phi\left(\frac{m - \nu - t}{\sigma}\right)\right], \tag{16}$$

whose first term is the class-1 error $\Pr(T \le t \mid Y = 1)$ and second the class-2 error $\Pr(T > t \mid Y = 2)$, with Φ the standard Gaussian cdf and Φ' its density. The balanced error on the fresh full distribution is $F_f(t) = F(t; d_f, \mu_f, \sigma_f)$, the deployment risk at shrinkage coefficient λ is $R_f^{\ell_{01}}(\lambda) = R_f^{\ell_{01}}(M_{\lambda,0})$ as defined in equation 2, and Appendix B.3 shows $R_f^{\ell_{01}}(\lambda) = F_f(t_\lambda)$. The balanced error on the reused distribution is $F_{\text{in}}(t) = F(t; m_{\text{in}}, \nu_{\text{in}}, \sigma_f)$.

B.1 A LEARNER REALIZING CLASS-DEPENDENT REUSE

Proposition 1 (Realizability of the logit distributions)**.** *Give every training sample a sample-identifying feature that only it activates, let the learner fit a unit residual on those features by class-balanced ridge regression, and add β times the fitted value to the contrast. A training observation activates its own feature and has its true-class contrast raised, while fresh and out-of-fold observations activate no feature and keep their contrast. The reuse increment of this learner is $\kappa_c = \beta/(1 + \gamma n_c)$ with $\beta, \gamma > 0$, and with $\mu_o = \mu_f$ and $\sigma_o = \sigma_f$ it satisfies equation 10 exactly.*

Proof. *Part (a): data model and learner.* Let $X = (T, V)$ with $T \mid Y = 1 \sim \mathcal{N}(\mu, \sigma^2)$, $T \mid Y = 2 \sim \mathcal{N}(-\mu, \sigma^2)$, $\mu, \sigma > 0$, and $V \sim \text{Unif}[0, 1]$ independent of (T, Y). The feature V only gives each sample an identity and carries no class information. Given training counts $\boldsymbol{r} = (r_1, r_2)$, the training samples are $\{(T_i, V_i, Y_i)\}_i$ with $Y_{i,\pm} = 2\mathbf{1}\{Y_i = 1\} - 1$. Since V is continuous, the V_i are almost surely pairwise distinct. The i-th sample-identifying feature is $\mathbf{1}\{V = V_i\}$, activated by sample i alone. The learner fits a unit residual on these features with class-balanced weights $1/r_{y_i}$ and a ridge penalty that does not scale with the class count, giving the ridge fit

$$\widehat{u} = \arg\min_u \ \frac{1}{2}\sum_i \frac{(u_i - 1)^2}{r_{y_i}} + \frac{\gamma}{2}\sum_i u_i^2, \qquad \gamma > 0. \tag{17}$$

For $\beta > 0$ the contrast and logit output of the predictor are

$$T_{S_{\boldsymbol{r}}}(T, V) = T + d_{\boldsymbol{r}} + \beta \sum_i Y_{i,\pm}\, \widehat{u}_i\, \mathbf{1}\{V = V_i\}, \qquad L_{S_{\boldsymbol{r}}}(T, V) = v\big(T_{S_{\boldsymbol{r}}}(T, V)\big), \tag{18}$$

where $d_{\boldsymbol{r}} \ge 0$ is the head bias of the learner at counts $\boldsymbol{r}$, a given function of the counts. Both equation 17 and the sum in equation 18 are symmetric in the sample index, so the learner is invariant to permutations of the training sample.

Part (b): the ridge solution and κ_c. The objective equation 17 separates coordinatewise and the second derivative of its i-th term in u_i is $1/r_{y_i} + \gamma > 0$, so it is strictly convex and its unique minimizer follows from the first-order condition:

$$\frac{u_i - 1}{r_{y_i}} + \gamma u_i = 0 \iff u_i\Big(\frac{1}{r_{y_i}} + \gamma\Big) = \frac{1}{r_{y_i}} \iff \widehat{u}_i = \frac{1}{1 + \gamma r_{y_i}}. \tag{19}$$

Thus $\widehat{u}_i$ depends on sample i only through its class count, and the reuse increment of class c, the absolute value of the third term of equation 18 when a training sample activates its own feature, is

$$\kappa_c(\boldsymbol{r}) = \beta\, \widehat{u}_i\big|_{y_i = c} = \frac{\beta}{1 + \gamma r_c}. \tag{20}$$

Part (c): the three response sources. We substitute into equation 18 in turn.

(c1) Reused (in-sample). The full model trains at counts $\boldsymbol{n}$ and scores its own training sample i of class c. Then $V = V_i$ activates the i-th feature and, by the pairwise distinctness of part (a), no other, so the sum keeps only the i-th term:

$$T_{\text{in}} = T_i + d_{\boldsymbol{n}} + \beta Y_{i,\pm}\widehat{u}_i = \begin{cases} T_i + d_{\boldsymbol{n}} + \kappa_1, & Y_i = 1, \\ T_i + d_{\boldsymbol{n}} - \kappa_2, & Y_i = 2. \end{cases} \tag{21}$$

Substituting $T_i \mid Y_i = 1 \sim \mathcal{N}(\mu, \sigma^2)$ and $T_i \mid Y_i = 2 \sim \mathcal{N}(-\mu, \sigma^2)$ and writing $d_f := d_{\boldsymbol{n}}$ gives $T_{\text{in}} \mid Y = 1 \sim \mathcal{N}(d_f + \mu + \kappa_1, \sigma^2)$ and $T_{\text{in}} \mid Y = 2 \sim \mathcal{N}(d_f - \mu - \kappa_2, \sigma^2)$, which is the second row of equation 10 with $\mu_f = \mu$ and $\sigma_f = \sigma$. The increment enters as $+\kappa_1$ for class 1 and $-\kappa_2$ for class 2 because T is defined head minus tail, so raising the true-class contrast of a tail sample lowers T.

(c2) Fresh. The full model scores an image $X^0 = (T^0, V^0)$ that took no part in training, with V^0 independent of all V_i and continuous, so $\Pr(V^0 = V_i \text{ for some } i) = 0$ and the third term of equation 18 vanishes almost surely:

$$T_f = T^0 + d_{\boldsymbol{n}}, \quad T_f \mid Y = 1 \sim \mathcal{N}(d_f + \mu, \sigma^2), \quad T_f \mid Y = 2 \sim \mathcal{N}(d_f - \mu, \sigma^2), \tag{22}$$

which is the first row of equation 10.

(c3) Out-of-fold. The fold model trains on S_{-k} and scores a training sample inside fold k. That sample's V is not among the $\{V_j\}$ of S_{-k}, so almost surely it activates none of their features and the third term vanishes again. Under exact stratification S_{-k} has $n'_c = n_c(1 - 1/K)$ samples per class, and writing $d_o := d_{\boldsymbol{n}'}$,

$$T_o = T + d_{\boldsymbol{n}'}, \quad T_o \mid Y = 1 \sim \mathcal{N}(d_o + \mu, \sigma^2), \quad T_o \mid Y = 2 \sim \mathcal{N}(d_o - \mu, \sigma^2), \tag{23}$$

the third row of equation 10 with $\mu_o = \mu$ and $\sigma_o = \sigma$. For unequal folds, replace $\boldsymbol{n}'$ by the actual count vector of that fold. Together equation 21–equation 23 give the claim.

Part (d): monotonicity of κ_c and Δ. From equation 20, κ_c is strictly decreasing in r_c:

$$\frac{\partial \kappa_c}{\partial r_c} = -\frac{\beta\gamma}{(1 + \gamma r_c)^2} < 0. \tag{24}$$

Hence $n_1 > n_2$ gives $\kappa_2 > \kappa_1$ and $\Delta = (\kappa_2 - \kappa_1)/2 > 0$, and fixing n_1 while decreasing n_2 strictly increases κ_2 and therefore Δ. The class dependence comes from the ratio in equation 17 between the class-balanced weight $1/r_{y_i}$ of the data term and the ridge penalty that does not scale with the class count: the larger the class, the smaller its data-term weight in relative terms and the more its residual is shrunk. □

B.2 Population affine fitting with ODIR

Proof of Lemma 1. Write $\ell_{\log}(x) = \log(1 + e^{-x})$ for the logistic loss and $\operatorname{expit}(x) = 1/(1 + e^{-x})$. Differentiating,

$$\ell'_{\log}(x) = \operatorname{expit}(x) - 1 = -\operatorname{expit}(-x), \qquad \ell''_{\log}(x) = \operatorname{expit}'(x) = \operatorname{expit}(x)\big(1 - \operatorname{expit}(x)\big) > 0, \tag{25}$$

and expit' is even. The proof has five parts:

Part (a): reduction to the contrast. The full problem has variables $G \in \mathbb{R}^{2\times 2}$ and $g \in \mathbb{R}^2$ with output $Gz + g \in \mathbb{R}^2$. Binary softmax cross-entropy depends only on the difference of the two output coordinates: for any $o \in \mathbb{R}^2$, $\ell_{\text{CE}}(o, 1) = -\log \frac{e^{o_1}}{e^{o_1} + e^{o_2}} = \ell_{\log}(o_1 - o_2)$ and $\ell_{\text{CE}}(o, 2) = \ell_{\log}(o_2 - o_1)$, which combine into $\ell_{\text{CE}}(o, Y) = \ell_{\log}(Y_{\pm}(o_1 - o_2))$. In the 2×2 case the ODIR penalty of equation 13 is $\xi_{\text{od}}(G_{12}^2 + G_{21}^2) + \xi_b(g_1^2 + g_2^2)$, so the full problem and its objective $\mathcal{J}_{\text{full}}$ are

$$\min_{G,g} \mathcal{J}_{\text{full}}(G, g) = \mathbb{E}\, \ell_{\log}\big(Y_{\pm}[(Gz + g)_1 - (Gz + g)_2]\big) + \xi_{\text{od}}\big(G_{12}^2 + G_{21}^2\big) + \xi_b\big(g_1^2 + g_2^2\big), \tag{26}$$

with $z = v(T)$. Substituting $z = v(T) = (T/2, -T/2)^\top$ gives $(Gz + g)_1 = (G_{11} - G_{12})T/2 + g_1$ and $(Gz + g)_2 = (G_{21} - G_{22})T/2 + g_2$, whose difference is

$$(Gz + g)_1 - (Gz + g)_2 = aT + b, \qquad a = \frac{G_{11} - G_{12} - G_{21} + G_{22}}{2}, \qquad b = g_1 - g_2. \tag{27}$$

Both a and b are linear in (G, g), and the identity $(I_2, 0)$ gives $a = 1$ and $b = 0$. Substituting equation 27 into equation 26, the cross-entropy term depends on (G, g) only through (a, b). The off-diagonal term is non-negative, and $(g_1 - g_2)^2 \le 2(g_1^2 + g_2^2)$ bounds the intercept penalty below by $\xi_b b^2/2$, so

$$\mathcal{J}_{\text{full}}(G, g) \;\ge\; \mathcal{J}(a, b) := \mathbb{E}\, \ell_{\log}\big(Y_{\pm}(aT + b)\big) + \frac{\xi_b}{2} b^2. \tag{28}$$

Conversely, for any (a, b) the canonical map $G = aI_2$, $g = (b/2, -b/2)^\top$ has contrast exactly (a, b) by equation 27, zero off-diagonal and $g_1^2 + g_2^2 = b^2/2$, so both bounds are tight:

$$\mathcal{J}_{\text{full}}\big(aI_2, (b/2, -b/2)^\top\big) = \mathcal{J}(a, b). \tag{29}$$

Together, $\inf \mathcal{J}_{\text{full}} = \inf \mathcal{J}$. If $\mathcal{J}$ has a unique minimizer (a^*, b^*), the canonical map is a finite minimizer of the full problem and the contrast of any full minimizer (G, g) must be (a^*, b^*), since otherwise equation 28 would give $\mathcal{J}_{\text{full}}(G, g) \geq \mathcal{J}(a, b) > \mathcal{J}(a^*, b^*) = \mathcal{J}_{\text{full}}(\text{canonical})$, contradicting minimality. The full parameters themselves need not be unique: a enters the contrast only through $G_{11} + G_{22}$ and the off-diagonal, so $G_{11} \mapsto G_{11} + \epsilon$ together with $G_{22} \mapsto G_{22} - \epsilon$ leaves both the contrast and Ω unchanged. This is the sense in which all minimizers induce the same contrast. It remains to study $\mathcal{J}$.

Part (b): $\mathcal{J}$ is strictly convex and coercive. Writing $x = Y_\pm(aT + b)$, the gradient of x in (a, b) is $Y_\pm(T, 1)^\top$ and the Hessian of the cross-entropy term is

$$\nabla^2_{(a,b)}\, \mathbb{E}\, \ell_{\log}(x) = \mathbb{E}\big[\ell''_{\log}(x)\, Y_\pm^2\, (T, 1)^\top (T, 1)\big] = \mathbb{E}\big[\operatorname{expit}'(aT + b)\, (T, 1)^\top (T, 1)\big], \tag{30}$$

using $Y_\pm^2 = 1$ and the evenness of expit'. For any $q = (q_1, q_2) \neq 0$ we have $q^\top \nabla^2 q = \mathbb{E}[\operatorname{expit}'(aT+b)(q_1 T+q_2)^2]$. Since $\operatorname{expit}' > 0$ and the marginal of T is a mixture of two Gaussians with everywhere positive density, $q_1 T + q_2 \neq 0$ almost everywhere and the expectation is strictly positive. The Hessian is positive definite, so the cross-entropy term is strictly convex, and $\frac{\xi_b}{2} b^2$ is convex, hence $\mathcal{J}$ is strictly convex. For coercivity, $\ell_{\log}(x) \geq \max(0, -x) = [-x]_+$ gives, for a unit vector q and $\varrho > 0$,

$$\mathcal{J}(\varrho q) \;\geq\; \mathbb{E}\,[-Y_\pm \varrho(q_1 T + q_2)]_+ \;=\; \varrho\, \chi(q), \qquad \chi(q) := \mathbb{E}\,[-Y_\pm(q_1 T + q_2)]_+. \tag{31}$$

Here $\chi(q) > 0$: for $q \neq 0$ the variable $q_1 T + q_2$ is not identically zero. Given $Y = 1$ the integrand requires $q_1 T + q_2 < 0$ and given $Y = 2$ it requires $q_1 T + q_2 > 0$, each class has probability $1/2$ and the conditional density of T is everywhere positive, so at least one of the two events has positive probability and the integrand is strictly positive on it. Since $\mathbb{E}|T| < \infty$, dominated convergence makes χ continuous on the unit circle, which is compact, so $\chi_{\min} = \min_q \chi(q) > 0$ and $\mathcal{J}(a, b) \geq \|(a, b)\| \chi_{\min} \to \infty$. Being continuous and coercive, $\mathcal{J}$ has a minimizer, and being strictly convex it is unique. We denote it by (a^*, b^*).

Part (c): $a^* > 0$. Take the gradient at $(a, b) = (0, 0)$. By equation 25, $\ell'_{\log}(0) = -1/2$, so

$$\partial_a \mathcal{J}(0, 0) = -\tfrac{1}{2}\, \mathbb{E}[Y_\pm T] = -\tfrac{1}{2}\Big[\tfrac{1}{2}(m + \nu) - \tfrac{1}{2}(m - \nu)\Big] = -\frac{\nu}{2}, \qquad \partial_b \mathcal{J}(0, 0) = -\tfrac{1}{2}\, \mathbb{E}[Y_\pm] = 0, \tag{32}$$

the last equality by the uniform class prior. A convex function lies above its supporting line, $\mathcal{J}(a, b) \geq \mathcal{J}(0, 0) - a\nu/2$, whose right side is at least $\mathcal{J}(0, 0)$ whenever $a \leq 0$. On the other hand $\mathcal{J}(\varepsilon, 0) = \mathcal{J}(0, 0) - \varepsilon\nu/2 + o(\varepsilon) < \mathcal{J}(0, 0)$ for small $\varepsilon > 0$, so the minimum is strictly below $\mathcal{J}(0, 0)$, no point with $a \leq 0$ can be a minimizer, and $a^* > 0$.

Part (d): sign and range of t^*. Fix $a > 0$ and treat b as the only variable. By equation 25, $\ell'_{\log}(Y_\pm x) Y_\pm$ equals $\operatorname{expit}(x) - 1$ when $Y_\pm = 1$ and $\operatorname{expit}(x)$ when $Y_\pm = -1$. Adding the two with weight $1/2$ each,

$$\partial_b \mathcal{J}(a, b) = \mathbb{E}\, \operatorname{expit}(aT + b) - \frac{1}{2} + \xi_b b, \tag{33}$$

the expectation taken over the marginal of T. Since expit is strictly increasing and $\xi_b \geq 0$, equation 33 is strictly increasing in b, so b^* is its unique zero and $t^* = -b^*/a$. Let $\tilde{T} = T - m$, whose marginal $\frac{1}{2}\mathcal{N}(\nu, \sigma^2) + \frac{1}{2}\mathcal{N}(-\nu, \sigma^2)$ is symmetric about 0. Evaluate equation 33 at two candidate points. At $b = -am$ we have $aT + b = a\tilde{T}$. Since $\operatorname{expit}(x) - \frac{1}{2}$ is odd and $\tilde{T}$ symmetric, $\mathbb{E}\operatorname{expit}(a\tilde{T}) = \frac{1}{2}$ and

$$\partial_b \mathcal{J}(a, -am) = -\xi_b\, a\, m. \tag{34}$$

At $b = 0$ we have $aT = a\tilde{T} + am$. For $m > 0$ the inequality $\operatorname{expit}(a\tilde{T} + am) > \operatorname{expit}(a\tilde{T})$ holds pointwise, so taking expectations $\partial_b \mathcal{J}(a, 0) > 0$, and symmetrically < 0 for $m < 0$ and $= 0$ for $m = 0$:

$$\operatorname{sign}\, \partial_b \mathcal{J}(a, 0) = \operatorname{sign}(m). \tag{35}$$

Three cases follow. For $m = 0$, equation 35 makes $b^* = 0$ and $t^* = 0$. For $\xi_b > 0$ and $m > 0$, equation 34 is negative at $b = -am$ and equation 35 positive at $b = 0$, so the unique zero of a strictly increasing function lies in $(-am, 0)$ and $t^* = -b^*/a \in (0, m)$; the case $m < 0$ is symmetric with $t^* \in (m, 0)$. Together these give $\operatorname{sign}(t^*) = \operatorname{sign}(m)$ and $0 < |t^*| < |m|$. For $\xi_b = 0$, equation 34 vanishes, so $b^* = -am$ and $t^* = m$.

Part (e): the slope at $\xi_b = 0$. Here $\mathcal{J}(a, b) = \mathbb{E}\,\ell_{\log}(Y_{\pm}(aT + b))$ is unregularized population logistic cross-entropy. The log-likelihood ratio of the two Gaussians is

$$\log \frac{\Phi'\big((T - m - \nu)/\sigma\big)}{\Phi'\big((T - m + \nu)/\sigma\big)} = \frac{(T - m + \nu)^2 - (T - m - \nu)^2}{2\sigma^2} = \frac{2\nu(T - m)}{\sigma^2}, \tag{36}$$

so under a uniform prior $\Pr(Y = 1 \mid T) = \operatorname{expit}(a_{\mathrm{B}}T + b_{\mathrm{B}})$ with $a_{\mathrm{B}} = 2\nu/\sigma^2$ and $b_{\mathrm{B}} = -2\nu m/\sigma^2$. Population cross-entropy equals $\mathbb{E}[\mathrm{KL}(\Pr(Y \mid T) \,\|\, \operatorname{expit}(aT + b))]$ plus a conditional entropy that does not depend on (a, b), and the KL term vanishes at $(a, b) = (a_{\mathrm{B}}, b_{\mathrm{B}})$, which therefore attains the minimum. By the uniqueness of part (b), $(a^*, b^*) = (a_{\mathrm{B}}, b_{\mathrm{B}})$. Hence $a^* = 2\nu/\sigma^2$ and $t^* = -b_{\mathrm{B}}/a_{\mathrm{B}} = m$, consistent with the $\xi_b = 0$ case of part (d). □

B.3 Proof of Theorem 1

Preliminary: a comparison principle for balanced error. For center m, half-separation $\nu > 0$, standard deviation $\sigma > 0$ and a uniform class prior, consider the rule "predict class 1 when $T > t$". The class-1 error is $\Pr(T \le t \mid Y = 1) = \Phi((t - m - \nu)/\sigma)$ and the class-2 error is $\Pr(T > t \mid Y = 2) = \Phi((m - \nu - t)/\sigma)$, so the balanced error is equation 16. The two terms exchange under $t \mapsto 2m - t$, so F is symmetric about $t = m$. Differentiating,

$$\partial_t F = \frac{1}{2\sigma}\left[\Phi'\left(\frac{t - m - \nu}{\sigma}\right) - \Phi'\left(\frac{t - m + \nu}{\sigma}\right)\right]. \tag{37}$$

For $t > m$ we have $|t - m - \nu| < |t - m + \nu|$ and Φ' is strictly decreasing in $|x|$, so $\partial_t F > 0$; for $t < m$ the same argument gives $\partial_t F < 0$. Being symmetric about $t = m$ and strictly increasing on $t > m$, F grows with the distance from the threshold to the center:

$$|t - m| < |t' - m| \quad \Longrightarrow \quad F(t; m, \nu, \sigma) < F(t'; m, \nu, \sigma). \tag{38}$$

□

Proof of Theorem 1. Throughout, Φ' is the standard Gaussian density and $\Phi''(x) = -x\Phi'(x)$.

Part (a): the sign of the reused-source minimizer. Writing the reused distribution of equation 10 in the form $\mathcal{N}(m \pm \nu, \sigma^2)$ of Lemma 1 requires $m + \nu = d_f + \mu_f + \kappa_1$ and $m - \nu = d_f - \mu_f - \kappa_2$. Adding and subtracting the two,

$$m_{\mathrm{in}} = d_f - \frac{\kappa_2 - \kappa_1}{2} = d_f - \Delta, \qquad \nu_{\mathrm{in}} = \mu_f + \frac{\kappa_1 + \kappa_2}{2} > 0, \qquad \sigma = \sigma_f. \tag{39}$$

Every hypothesis of Lemma 1 holds: the class prior is uniform under balanced population cross-entropy, the two conditional distributions are Gaussian with a common variance, $\nu_{\mathrm{in}} > 0$, $\sigma_f > 0$, and ξ_{od}, ξ_b are finite. The lemma gives $a^*_{\mathrm{in}} > 0$. The hypothesis $\Delta > d_f$ is exactly $m_{\mathrm{in}} < 0$, so by the branches of the lemma $t^*_{\mathrm{in}} = m_{\mathrm{in}} < 0$ when $\xi_b = 0$, and $\operatorname{sign}(t^*_{\mathrm{in}}) = \operatorname{sign}(m_{\mathrm{in}}) < 0$ with $|t^*_{\mathrm{in}}| < |m_{\mathrm{in}}|$ when $\xi_b > 0$. The two branches combine into

$$m_{\mathrm{in}} \le t^*_{\mathrm{in}} < 0. \tag{40}$$

Part (b): the deployment threshold and the threshold form of the deployment risk. Take any minimizer (G^*, g^*) of Lemma 1 as the endpoint of equation 14. At $\tau = 0$, $M_{\lambda,0}(z) = (1 - \lambda)z + \lambda(G^*z + g^*)$. By equation 27 the contrast (a, b) is linear in (G, g) and the identity has contrast $a = 1$, $b = 0$, namely T itself, so the output contrast of $M_{\lambda,0}$ is the convex combination of the two endpoint contrasts,

$$(1 - \lambda)\,T + \lambda\,(a^*_{\mathrm{in}}T + b^*_{\mathrm{in}}) = (1 - \lambda + \lambda a^*_{\mathrm{in}})\,T + \lambda b^*_{\mathrm{in}}. \tag{41}$$

Top-1 predicts class 1 exactly when equation 41 is positive. It is affine in T with slope $1 - \lambda + \lambda a^*_{\mathrm{in}}$, a convex combination of 1 and $a^*_{\mathrm{in}} > 0$ and therefore strictly positive for $\lambda \in [0, 1]$, so equation 41

is positive to the right of its zero and negative to the left. Call that zero the threshold t_λ of $M_{\lambda,0}$, so that its top-1 decision is the rule "predict class 1 when $T > t_\lambda$". Solving equation 41 for zero gives $t_\lambda = -\lambda b^*_{\text{in}}/(1-\lambda+\lambda a^*_{\text{in}})$, and the endpoint $M_{1,0}$ has threshold $t^*_{\text{in}} = -b^*_{\text{in}}/a^*_{\text{in}}$. Substituting $b^*_{\text{in}} = -a^*_{\text{in}} t^*_{\text{in}}$,

$$t_\lambda = \frac{\lambda a^*_{\text{in}}}{1-\lambda+\lambda a^*_{\text{in}}}\, t^*_{\text{in}} = w_\lambda\, t^*_{\text{in}}. \tag{42}$$

Direct substitution gives $w_0 = 0$ and $w_1 = 1$, and differentiating in λ, whose numerator is $a^*_{\text{in}}(1-\lambda+\lambda a^*_{\text{in}}) - \lambda a^*_{\text{in}}(a^*_{\text{in}}-1) = a^*_{\text{in}}$,

$$w'_\lambda = \frac{a^*_{\text{in}}}{(1-\lambda+\lambda a^*_{\text{in}})^2} > 0, \tag{43}$$

so w_λ is strictly increasing on $[0,1]$. Since the decision of $M_{\lambda,0}$ is a threshold rule, the preliminary gives its balanced error on a distribution whose contrast follows $\mathcal{N}(m \pm \nu, \sigma^2)$ as $F(t_\lambda; m, \nu, \sigma)$, and the same t_λ evaluated at the parameters of different sources gives different values. Taking the fresh parameters (d_f, μ_f, σ_f) of the first row of equation 10 and writing $F_f = F(\cdot\,; d_f, \mu_f, \sigma_f)$, the deployment risk of equation 2 is the balanced error of $M_{\lambda,0}$ on the fresh distribution, so $R^{\ell_{01}}_f(\lambda) = F_f(t_\lambda)$. Taking the reused parameters equation 39 and writing $F_{\text{in}} = F(\cdot\,; m_{\text{in}}, \nu_{\text{in}}, \sigma_f)$, the balanced error of the same t_λ on the reused distribution is $F_{\text{in}}(t_\lambda)$.

Part (c): balanced error and cross-entropy on the reused distribution.

(c1) Balanced error strictly decreases. By equation 40 and $w_\lambda \in [0,1]$ we have $t_\lambda = w_\lambda t^*_{\text{in}} \in [t^*_{\text{in}}, 0]$, hence $t_\lambda \geq m_{\text{in}}$ and

$$|t_\lambda - m_{\text{in}}| = t_\lambda - m_{\text{in}} = |m_{\text{in}}| - w_\lambda |t^*_{\text{in}}|, \tag{44}$$

which decreases strictly in λ since w_λ increases strictly and $|t^*_{\text{in}}| > 0$. By the comparison principle equation 38, $F_{\text{in}}(t_\lambda)$ decreases strictly in λ and is minimized at $\lambda = 1$.

(c2) Cross-entropy is non-increasing. Write $R^{\ell_{\text{CE}}}_{\text{in}}(G, g) = \mathbb{E}\,\ell_{\text{CE}}(Gv(T_{\text{in}}) + g, Y)$ and $R^{\ell_{\text{CE}}}_{\text{in}}(\lambda) = R^{\ell_{\text{CE}}}_{\text{in}}(G_\lambda, g_\lambda)$, with Ω the ODIR penalty of equation 13. The pair (G^*, g^*) is the global minimizer of $R^{\ell_{\text{CE}}}_{\text{in}} + \Omega$ over all (G, g). The identity has zero off-diagonal and zero bias, so the off-diagonal of G_λ is λ times that of G^* and $g_\lambda = \lambda g^*$, giving

$$\Omega(G_\lambda, g_\lambda) = \lambda^2\, \Omega(G^*, g^*). \tag{45}$$

Applying global optimality at (G_λ, g_λ),

$$\begin{aligned} & R^{\ell_{\text{CE}}}_{\text{in}}(1) + \Omega(G^*, g^*) \leq R^{\ell_{\text{CE}}}_{\text{in}}(\lambda) + \lambda^2 \Omega(G^*, g^*) \\ \Longrightarrow \quad & R^{\ell_{\text{CE}}}_{\text{in}}(1) - R^{\ell_{\text{CE}}}_{\text{in}}(\lambda) \leq (\lambda^2 - 1)\, \Omega(G^*, g^*) \leq 0, \end{aligned} \tag{46}$$

the last step by $\lambda \in [0,1]$ and $\Omega \geq 0$. Moreover $R^{\ell_{\text{CE}}}_{\text{in}}(\lambda)$ is convex in λ: $\ell_{\text{CE}}(\cdot, y)$ is convex in the logit vector, $M_{\lambda,0}(z)$ is affine in λ, the composition of a convex function with an affine map is convex, and taking expectations preserves convexity. A convex function attaining its minimum at the right endpoint of an interval is non-increasing on the whole interval: for $\lambda' < \lambda'' \leq 1$, writing $\lambda'' = \vartheta\lambda' + (1-\vartheta)\cdot 1$ with $\vartheta = (1-\lambda'')/(1-\lambda') \in [0,1)$ gives $R^{\ell_{\text{CE}}}_{\text{in}}(\lambda'') \leq \vartheta R^{\ell_{\text{CE}}}_{\text{in}}(\lambda') + (1-\vartheta) R^{\ell_{\text{CE}}}_{\text{in}}(1) \leq R^{\ell_{\text{CE}}}_{\text{in}}(\lambda')$. Inequality equation 46 uses global optimality, so it covers the population minimizer. Appendix D.2 reports the endpoint property of the finite-step solver output.

Part (d): balanced error on the fresh full distribution, and the closed forms. By equation 40, $t^*_{\text{in}} < 0 \leq d_f$, and with $w_\lambda \in [0,1]$ we get $t_\lambda = w_\lambda t^*_{\text{in}} \in [t^*_{\text{in}}, 0]$ and $t_\lambda \leq 0 \leq d_f$, so

$$|t_\lambda - d_f| = d_f - t_\lambda = d_f + w_\lambda |t^*_{\text{in}}|, \tag{47}$$

which increases strictly in λ. By equation 38 and $R^{\ell_{01}}_f(\lambda) = F_f(t_\lambda)$ from part (b), the deployment risk increases strictly in λ. At $\lambda = 0$ we have $w_0 = 0$ and $t_0 = 0$, so $R^{\ell_{01}}_f(0) = F_f(0)$ is the deployment risk of the identity and every $\lambda > 0$ is strictly worse. Per class, t_λ decreases strictly in λ, so the class-2 (tail) error $\Phi((d_f - \mu_f - t_\lambda)/\sigma_f)$ increases strictly and the class-1 (head) error $\Phi((t_\lambda - d_f - \mu_f)/\sigma_f)$ decreases strictly. When $\xi_b = 0$, Lemma 1 gives $a^* = 2\nu/\sigma^2$ and $t^* = m$, and substituting equation 39,

$$a^*_{\text{in}} = \frac{2\nu_{\text{in}}}{\sigma_f^2} = \frac{2\mu_f + \kappa_1 + \kappa_2}{\sigma_f^2}, \qquad t^*_{\text{in}} = m_{\text{in}} = d_f - \Delta, \qquad R^{\ell_{01}}_f(1) = F_f(d_f - \Delta). \tag{48}$$

□

On the reused distribution both cross-entropy and balanced error are optimal at $\lambda = 1$, and it is precisely at that endpoint that the fresh balanced error is largest: the scoring objective and the deployment risk point in opposite directions. When $\Delta < d_f$ we have $m_{\text{in}} > 0$ and Lemma 1 gives $t^*_{\text{in}} \in (0,\, d_f - \Delta]$, the correct direction stopping short of the Bayes threshold, which is undercorrection. The conflict is decided by the source of the logits, and Appendix B.4 gives the optimal strength once that source is correct. The same mechanism can be extended to multiclass settings.

B.4 Proof of Theorem 2 and Corollary 1

Proof of Theorem 2. *Part (a): the interpolation formula.* With (a, b) as endpoint contrast, the output contrast of $M_{\lambda,\tau}$ is the convex combination of the two endpoint contrasts minus the logit-adjustment shift, $(1-\lambda)T + \lambda(aT+b) - h_\tau = (1-\lambda+\lambda a)T + \lambda b - h_\tau$. Its slope $1-\lambda+\lambda a$ is a convex combination of 1 and $a > 0$ and therefore strictly positive, so top-1 predicts class 1 exactly when

$$T > t_\lambda := \frac{h_\tau - \lambda b}{1-\lambda+\lambda a}. \tag{49}$$

Substituting $\lambda = 0$ gives $t_0 = h_\tau$ and $\lambda = 1$ gives $t_1 = (h_\tau - b)/a$. For the interpolation, the definition of w_λ gives $1 - w_\lambda = (1-\lambda)/(1-\lambda+\lambda a)$ and

$$(1-w_\lambda)t_0 + w_\lambda t_1 = \frac{(1-\lambda)h_\tau + \lambda(h_\tau - b)}{1-\lambda+\lambda a} = \frac{h_\tau - \lambda b}{1-\lambda+\lambda a} = t_\lambda. \tag{50}$$

The properties of w_λ are those of equation 43 with a^*_{in} replaced by a: $w_0 = 0$, $w_1 = 1$ and $w'_\lambda = a/(1-\lambda+\lambda a)^2 > 0$, so w_λ is continuous, strictly increasing and onto $[0,1]$, and t_λ sweeps the entire segment between t_0 and t_1.

Part (b): the optimal w^* and λ^*. The balanced error of the fresh full distribution is $F_f(t) = F(t; d_f, \mu_f, \sigma_f)$ of equation 16. By equation 37, F_f decreases strictly on $t < d_f$ and increases strictly on $t > d_f$, so d_f is its unique minimizer. Parameterize the points of the segment by $w \in [0,1]$ as $t_0 + w(t_1 - t_0)$. For $t_1 \neq t_0$ this point is strictly monotone in w, so $F_f(t_0 + w(t_1 - t_0))$ is strictly monotone on each side of d_f and its minimizer is the projection of d_f onto the segment. Solving $t_0 + w(t_1 - t_0) = d_f$ and projecting onto $[0,1]$,

$$w^* = \text{clip}_{[0,1]}\Big(\frac{d_f - t_0}{t_1 - t_0}\Big), \tag{51}$$

unique by the strict monotonicity on each side. Inverting $w = \lambda a/(1-\lambda+\lambda a)$ for λ, namely $w(1-\lambda+\lambda a) = \lambda a \iff w = \lambda(a + (1-a)w)$, whose denominator $a + (1-a)w = (1-w)a + w > 0$ is a convex combination of $a > 0$ and 1,

$$\lambda = \frac{w}{a + (1-a)w}. \tag{52}$$

Since w_λ is a strictly increasing bijection of $[0,1]$ onto itself by part (a), the unique w^* gives a unique λ^*.

Part (c): $t_0 = t_1$. By equation 50, $t_\lambda \equiv t_0$, so the rule "predict class 1 when $T > t_0$" does not vary with λ and the top-1 risk is constant.

Part (d): general continuous class densities. For class-conditional densities f_1, f_2 and a uniform prior, the balanced error of the threshold-t rule and its derivative are

$$R(t) = \frac{1}{2}\Big[\int_{-\infty}^{t} f_1(u)\,du + \int_{t}^{\infty} f_2(u)\,du\Big], \qquad R'(t) = \frac{f_1(t) - f_2(t)}{2}. \tag{53}$$

The hypotheses $f_1 < f_2$ on $t < d_f$ and $f_1 > f_2$ on $t > d_f$ give $R' < 0$ and $R' > 0$ on the two sides, so R has the same unimodal structure as F_f and the projection argument of parts (b) and (c) applies verbatim. The Gaussian case is a special case: by equation 37 the sign of F'_f is exactly that of the difference of the two Gaussian densities. □

Proof of Corollary 1. *Part (a): the threshold segment of the out-of-fold endpoint.* The out-of-fold distribution is the third row of equation 10 with center $m = d_o$ and half-separation $\nu = \mu_o$. Lemma 1

at $m = d_o > 0$ gives $a_o^* > 0$ and $\text{sign}(t_o^*) = \text{sign}(d_o)$, so $t_o^* > 0$. At $\tau = 0$ we have $h_\tau = 0$, and Theorem 2(a) gives $t_0 = 0$ and $t_1 = -b_o^*/a_o^* = t_o^*$, so

$$t_\lambda = w_\lambda \, t_o^* \in (0, t_o^*] \qquad (\lambda > 0), \tag{54}$$

and the segment is $[0, t_o^*]$.

Part (b): strictly better than the identity. The identity is $\lambda = 0$ with threshold 0 and balanced error $F_f(0)$. By the comparison principle equation 38, for $d_f > 0$,

$$F_f(t_\lambda) < F_f(0) \iff |t_\lambda - d_f| < |0 - d_f| = d_f \iff 0 < t_\lambda < 2d_f. \tag{55}$$

For $\lambda > 0$, equation 54 already gives $t_\lambda > 0$, so the condition reduces to $t_\lambda < 2d_f$. Such a λ exists because w_λ is continuous with $w_0 = 0$, so $t_\lambda = w_\lambda t_o^* \to 0$ as $\lambda \downarrow 0$ and every small enough $\lambda > 0$ satisfies $0 < t_\lambda < 2d_f$.

Part (c): interior or endpoint. Substituting $t_0 = 0$ and $t_1 = t_o^*$ into equation 51 gives $w^* = \text{clip}_{[0,1]}(d_f/t_o^*)$. If $0 < d_f < t_o^*$ then $w^* = d_f/t_o^* \in (0, 1)$, the segment contains the Bayes threshold, $t_{\lambda^*} = d_f$, the balanced error attains $F_f(d_f)$, the Bayes error of the fresh distribution, and equation 52 gives $\lambda^* \in (0, 1)$. If $d_f \geq t_o^*$ then $w^* = 1$ and $\lambda^* = 1$. If $d_f = 0 < t_o^*$ then $w^* = 0$ and $\lambda^* = 0$, the identity already being the Bayes rule.

Part (d): the closed form at $\xi_b = 0$. The $\xi_b = 0$ branch of Lemma 1 gives $t_o^* = m = d_o$. For $0 < d_f < d_o$ we have $w^* = d_f/d_o$, and substituting into equation 52,

$$\lambda^* = \frac{d_f/d_o}{a_o^* + (1 - a_o^*)\, d_f/d_o} = \frac{d_f}{a_o^* d_o + (1 - a_o^*) d_f}. \tag{56}$$

□

Remark 1 (The case $d_o = 0$)**.** Lemma 1 then gives $t_o^* = 0 = t_0$, which falls under Theorem 2(c): at $\tau = 0$ the classification rule does not vary with λ, so the out-of-fold correction neither improves nor degrades the identity.

C Experimental setup

C.1 Training

Table 7: Datasets and first-stage recipes.

Dataset	Classes; IF	First-stage recipe
CIFAR-100-LT	100; $\{100, 50, 10\}$	GLMC+CUDA, 400 epochs or plain GLMC, 200 epochs
CIFAR-10-LT	10; $\{100, 50, 10\}$	GLMC+CUDA, 400 epochs or plain GLMC, 200 epochs
ImageNet-LT	1,000; IF 256	GLMC
iNaturalist2018	8,142; IF 500	GLMC

Large-scale runs. For ImageNet-LT and iNaturalist2018, fold models come from a 5-fold stratified split under the same GLMC recipe as the full model.

Class splits. The CIFAR-100-LT split thresholds by n_c are many $n_c > 100$, medium $20 < n_c \leq 100$ and few $n_c \leq 20$ at IF100 and IF50, and many $n_c > 200$, medium $100 < n_c \leq 200$ and few $n_c \leq 100$ at IF10. The CIFAR-10-LT tables report overall top-1, and its few-shot split is the three rarest classes. ImageNet-LT and iNaturalist2018 follow Kang et al. (2020): many $n_c > 100$, medium $20 \leq n_c \leq 100$, few $n_c < 20$.

Folds. Fold assignment is a per-class random permutation under a fixed seed, with images dealt to the K folds in turn. The rarest CIFAR-100-LT class has $n_c = 5$ with $K = 5$, so each held-out fold holds one image and its fold model trains on the remaining four. When n_c is not divisible by K, the class counts of the held-out folds differ by at most one, and so do the training counts of the fold models. Every fold model's training complement retains every class.

Baselines. The GLMC and GLMC+CUDA rows of Table 1 deploy the class-balanced classifier head of the first stage. GLMC + MaxNorm applies the second stage of Alshammari et al. (2022) to that head: class-balanced cross-entropy for 10 epochs from the first-stage head, projecting each class weight row onto a norm ball every epoch. GLMC + MaxNorm + ours remaps the logits of that head after MaxNorm, the full model and the five fold models each taking their own MaxNorm pass.

C.2 Remapper fit and selection

Table 8: Fitting and selection hyperparameters.

Item	Value
Fit	Adam, lr 0.05, 400 steps (200 inside the inner split), class-balanced CE, per-observation weight $1/(Cn_{y_i})$
Penalty grid	$\xi_{\mathrm{od}} \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$, $\xi_b \in \{10^{-3}, 10^{-2}, 10^{-1}\}$
Inner split	$K = 5$ inner folds of the logit table; fit on 4, score on 1
Shrinkage grid	$\Lambda = \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.5, 0.75, 1\}$
LA grid	$\mathcal{T} = \{0, 0.1, 0.2, 0.25, 0.3, 0.4, 0.5\}$

C.3 Measurement protocols

Contrast, center and reuse increment. In the multiclass setting the true-class contrast of an image is its true-class logit minus the largest of the remaining logits, the counterpart of T in the binary model of Section 4. The head–tail center of a response source is half the difference between its many-shot and few-shot per-class contrast means, and each pair of centers is taken from the same set of models: the full model gives the in-sample center and the test center d_f, and the fold models give the matched in-fold center and the out-of-fold center d_o. Their difference is the $d_o - d_f$ of Corollary 1, and $w^* = d_f/d_o$ is the boundary coefficient of Theorem 2 at $\tau = 0$ and $\xi_b = 0$. The reuse increment κ_c is the per-class average of the true-class contrast difference between a training image's matched in-fold logit and its out-of-fold logit.

Net flips and bootstrap. The net-flip counts compare the test predictions at $\lambda = 0$ and $\lambda = 1$. The paired bootstrap resamples the test indices 5000 times, applying the same index set to our fit and to LA. All measurements use only the existing fold models, the full model and the logits of the training and test images, and train no new model. The reliability diagrams of Figure 9 use 15 equal-width confidence bins, each drawn at its mean confidence with point area proportional to bin count. For LWS the map (G, g) acts on the logits before the per-class scale, with $\mathrm{diag}(G)$ replacing that scale.

D Additional experimental results

The first three subsections follow the three subsections of Section 5, and the last splits the remapper into its components.

D.1 Comparison with further methods

Tables 9 and 10 extend Tables 1 and 2 with further methods. The remapper leaves the architecture and training of the first stage unchanged and measures how much accuracy a correction fit on the long-tailed training set still recovers. On CIFAR-LT, GLMC+CUDA with the remapper outperforms every listed method, including the multi-expert ones. MaxNorm on the class-balanced head lowers five of the six columns of Table 1, and the remapper stacked after it returns all settings to within 0.11 of GLMC + ours. On ImageNet-LT and iNaturalist2018, where the first stage learns its features from far more training images, the remapper adds 1.59 and 0.78 points to GLMC, and the four multi-expert methods listed below it outperform it by up to 4.19 and 3.12 points.

Table 9: **Comparison on CIFAR-100-LT and CIFAR-10-LT with further methods.** Results of other methods are from the original papers, Zhang et al. (2022), Zhang et al. (2026) and Dong et al. (2026).

	CIFAR-100-LT			CIFAR-10-LT		
Method	100	50	10	100	50	10
CE	47.4	52.3	67.2	80.0	84.3	91.9
Focal loss (Lin et al., 2017)	38.4	44.3	55.8	70.4	76.7	86.7
LDAM-DRW (Cao et al., 2019)	42.0	46.6	58.7	77.0	81.0	88.2
cRT (Kang et al., 2020)	42.3	46.8	58.1	75.7	80.4	88.3
BBN (Zhou et al., 2020)	42.6	47.0	59.1	79.8	82.2	88.3
MiSLAS (Zhong et al., 2021)	52.7	56.6	69.7	84.2	87.2	92.7
DiVE (He et al., 2021)	45.4	51.1	62.0	—	—	—
CAM-BS (Zhang et al., 2021b)	41.7	46.0	—	75.4	81.4	—
RIDE (3 experts) (Wang et al., 2021)	48.0	51.7	61.8	87.2	87.8	89.2
WD+MaxNorm (Alshammari et al., 2022)	53.4	57.7	68.7	—	—	—
SAM (Rangwani et al., 2022)	45.4	49.4	60.1	81.9	84.8	89.4
CMO (Park et al., 2022)	47.2	51.7	58.4	—	—	—
SADE (Zhang et al., 2022)	48.8	53.8	63.6	86.9	87.3	88.6
CSA (Shi et al., 2023)	46.6	51.9	62.6	82.5	86.0	90.8
ADRW (Wang et al., 2023b)	46.4	—	61.9	83.6	—	90.3
CUDA (Ahn et al., 2023)	47.6	51.1	58.4	—	—	—
BalPoE (Sanchez Aimar et al., 2023)	52.0	56.3	64.8	87.3	87.7	90.8
MDCS (Zhao et al., 2023)	53.2	57.2	—	85.8	89.4	—
H2T (Li et al., 2024)	48.9	53.8	—	82.4	85.4	—
IWB (Dang et al., 2024)	53.3	56.3	—	—	—	—
DiffuLT (Shao et al., 2024)	51.5	56.3	63.8	84.7	86.9	90.7
LOS (Sun et al., 2025)	54.9	58.8	69.7	—	—	—
GUIDE (Dong et al., 2026)	56.4	60.3	69.2	87.6	90.1	92.5
SAMN (Zhang et al., 2026)	57.7	64.0	74.1	88.5	91.3	95.2
GLMC (Du et al., 2023)	57.51	62.49	73.54	88.06	90.70	94.92
+ ours	58.22	63.40	74.16	88.90	91.15	94.99
	+0.71	+0.91	+0.62	+0.84	+0.45	+0.07
GLMC + MaxNorm	57.85	62.08	73.28	86.84	90.00	94.67
+ ours	58.31	63.29	74.18	88.94	91.20	95.03
	+0.46	+1.21	+0.90	+2.10	+1.20	+0.36
GLMC+CUDA (Du et al., 2023; Ahn et al., 2023)	59.77	64.34	75.14	88.05	91.64	95.35
+ ours	**61.53**	**65.90**	**76.35**	**89.49**	**92.26**	**95.45**
	+1.76	+1.56	+1.21	+1.44	+0.62	+0.10

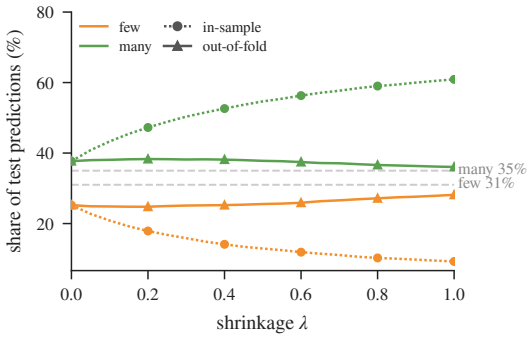

Figure 7: **Split shares of the test predictions along λ.** Dashed lines mark the true shares of the test set.

Table 10: **Comparison on ImageNet-LT and iNaturalist2018 with further methods.** Methods below the remapper rows outperform it in overall accuracy on at least one dataset. Results of other methods are from the original papers, Zhang et al. (2022), Zhang et al. (2026) and Dong et al. (2026).

	ImageNet-LT				iNaturalist2018			
Method	Many	Med.	Few	All	Many	Med.	Few	All
CE	65.9	37.5	7.7	44.4	72.2	63.0	57.2	61.7
cRT (Kang et al., 2020)	61.8	46.2	27.4	49.6	69.0	66.0	63.2	65.2
τ-norm (Kang et al., 2020)	59.1	46.9	30.7	49.4	65.6	65.3	65.9	65.6
LogitAdjust (Menon et al., 2021)	—	—	—	51.1	—	—	—	66.4
MiSLAS (Zhong et al., 2021)	61.7	51.3	35.8	52.7	73.2	72.4	70.4	71.6
DiVE (He et al., 2021)	64.1	50.4	31.5	53.1	70.6	70.0	67.6	69.1
RIDE (3 experts) (Wang et al., 2021)	68.0	52.9	35.1	56.3	71.5	70.0	71.6	71.8
WD+MaxNorm (Alshammari et al., 2022)	62.5	50.4	41.5	53.9	71.2	70.4	69.7	70.2
AREA (Chen et al., 2023)	—	—	—	49.5	—	—	—	68.4
MARC (Wang et al., 2023a)	60.4	50.3	36.6	52.3	—	—	—	70.4
RBL (Peifeng et al., 2023)	64.8	49.6	34.2	53.3	—	—	—	—
H2T (Li et al., 2024)	62.4	48.8	**52.2**	54.6	67.7	71.9	72.2	71.6
IWB (Dang et al., 2024)	64.2	52.2	40.2	55.2	72.3	70.6	72.5	71.5
DiffuLT (Shao et al., 2024)	63.3	55.6	39.4	56.4	—	—	—	—
EWB-FDR (Hasegawa & Sato, 2024)	63.4	50.0	35.1	53.2	—	—	—	—
LOS (Sun et al., 2025)	63.2	50.7	42.3	54.4	69.2	70.7	71.3	70.8
SAMN (Zhang et al., 2026)	**69.7**	56.1	31.1	57.7	64.4	73.7	73.6	72.7
GLMC (Du et al., 2023)	68.3	54.3	32.4	56.72	65.7	73.6	72.1	72.20
+ ours	65.3	57.0	43.1	58.31	65.1	73.8	74.0	72.98
				+1.59				+0.78
SADE (Zhang et al., 2022)	66.5	57.0	43.5	58.8	**74.5**	72.5	73.0	72.9
BalPoE (Sanchez Aimar et al., 2023)	68.2	**57.2**	44.9	59.8	73.2	**75.5**	**74.7**	75.0
MDCS (Zhao et al., 2023)	—	—	—	59.3	71.8	73.1	72.4	72.5
GUIDE (Dong et al., 2026)	—	—	47.6	**62.5**	—	—	—	**76.1**

Table 11: **Top-1 over three seeds and net flips of the in-sample endpoint.** Rows of Table 3 averaged over three seeds, and test images the in-sample endpoint flips against $\lambda = 0$.

	Top-1, mean of three seeds				In-sample endpoint		
IF	LA	Out-of-fold	In-sample	Matched	Correct→wrong	Wrong→correct	Balanced error
100	61.01	**61.53**	55.57	55.82	978	413	0.397 → 0.453
10	75.80	**76.35**	73.22	73.10	500	193	0.242 → 0.273

D.2 The reversal across seeds, images and rules

Seeds and individual images. We check that the reversal of Table 3 holds on every seed and on individual test images (Table 11). Averaged over three seeds the two reused fits underperform LA by 5.44 and 5.19 points at IF100 and by 2.58 and 2.70 at IF10, and every seed selects $\lambda = 1$ on both. At IF100 the in-sample endpoint turns 978 correct test images wrong and 413 wrong ones correct, and along λ it raises the many-shot share of test predictions from 37.7% to 60.9% (Figure 7). The reversal holds on every seed, and the reused endpoint turns more test images wrong than right.

Where each fit stops. We next ask why the reused fits stop at the endpoint $\lambda = 1$. On every fitting table the cross-entropy falls monotonically to $\lambda = 1$, from 2.145 to 0.903 on the out-of-fold table and from 0.331 to 0.001 on the in-sample table (Figure 4(b)), and the finite-step output of Section 3.3 attains its endpoint optimum on every fit's table, checked on the solver output. The balanced-error selector on held-out rows therefore decides where each fit stops.

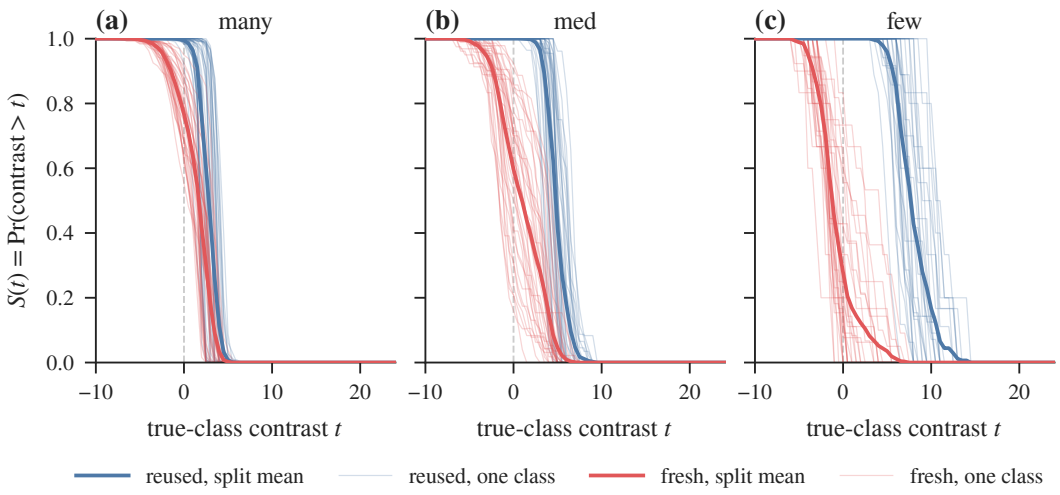

Figure 8: **Survival curves of the true-class contrast.** $S(t) = \Pr(\text{contrast} > t)$, one panel per split, thin lines per class and thick lines the split mean.

D.3 Mechanism analysis

Memorization in the contrast distribution. We trace the inflation of Table 4 to the whole contrast distribution (Figure 8). Beyond the mean gap κ_c of Figure 2, the reused contrast curve lies nowhere below the fresh one on 93 of the 100 classes. Memorization shifts the whole contrast distribution of a training image.

Table 12: **True-class contrast by response source.** Split means on CIFAR-100-LT under GLMC+CUDA.

Split	In-sample	Matched in-fold	Out-of-fold	Test
Many	3.00	2.78	1.30	1.55
Medium	5.32	4.89	0.97	1.34
Few	8.69	7.95	−0.80	−0.36

Centers. We give the contrast means from which the centers of Section 5.3 are computed (Table 12). Both reused sources sit above the test means in every split and the out-of-fold source below them, and the per-class distance of each mean vector to the test means is 2.31 for out-of-fold against 9.72 for in-sample and 8.15 for matched.

Table 13: **cRT and LWS with the remapper fit on each logit table.** Top-1 on the cross-entropy first stage, with ECE in parentheses.

Method	Plain	In-sample	Matched	Out-of-fold
cRT	41.11 (0.30)	39.90 (0.42)	39.94 (0.42)	**42.47** (0.03)
LWS	40.66 (0.39)	40.02 (0.42)	40.02 (0.42)	**42.19** (0.03)

The cross-entropy first stage. We run the three logit tables through cRT and LWS on the cross-entropy first stage of Table 5 (Table 13, Figure 9). On both methods the reused fits underperform the method's own second stage, at 39.90 to 40.02, and raise the expected calibration error (ECE) to about 0.42, while the out-of-fold fit reaches 42.47 and 42.19 with ECE 0.03. A fit on memorized logits inherits their overconfidence, and the out-of-fold map removes the calibration gap together with the accuracy gap.

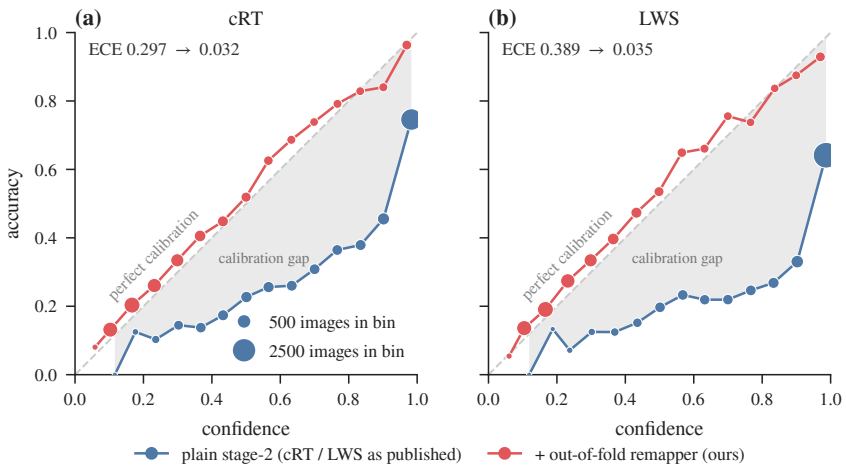

Figure 9: **Reliability diagrams of cRT and LWS.** Blue is each method's own second stage and red the out-of-fold (G, g) applied to its logits, with ECE in each panel.

D.4 Components

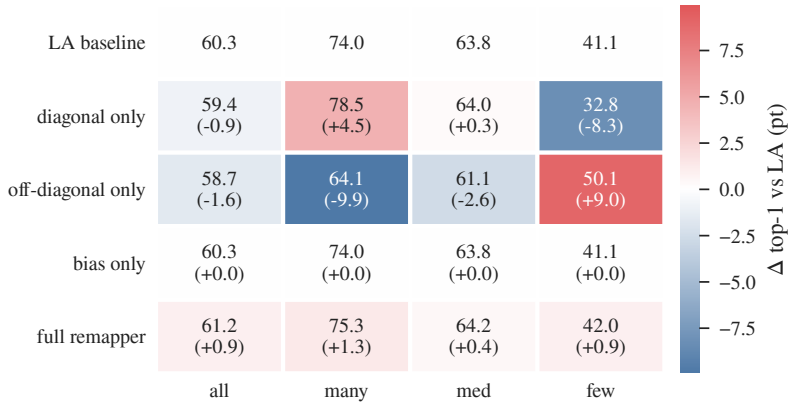

Figure 10: **Component ablation of $\widehat{G}$.** Rows are the deployed component and columns the split. Each cell gives the top-1 and its change against LA in parentheses, red above LA and blue below.

Components. We deploy the diagonal, the off-diagonal and the bias of $\widehat{G}$ separately at the deployed setting (Figure 10). The diagonal alone raises many-shot accuracy to 78.5 and lowers few-shot accuracy to 32.8, the direction of the reused fit, and the off-diagonal alone does the reverse. The bias alone reproduces LA, since the selected $\xi_b = 0.1$ leaves its largest component at 0.006. On the cross-entropy first stage the diagonal carries $+1.44$ of the LWS gain and the full G a further $+0.09$, while on cRT the full G adds $+0.79$ over the diagonal. The diagonal and the off-diagonal pull in opposite directions, and only together do they exceed LA in every split.

E Limitations

The construction trains K fold models beside the full model to produce the fitting logits, and the analysis of Section 4 applies to any cheaper source whose center sits on the tail side. On ImageNet-LT and iNaturalist2018 these K extra trainings return 1.59 and 0.78 points, a low accuracy gain per unit of training compute. The analysis covers the direction and strength of an affine correction fit on a frozen predictor. An affine remapper is not the optimal second stage, and the closed form for λ^* is optimal only within the deployment family of equation 1.