Rank-Preserving Calibration for Multiclass Classification

Gaurav Sood*

Abstract

We study post-hoc calibration for multiclass classifiers when two desiderata collide: (i) match known class totals (label marginals) and (ii) preserve the within-class ranking of instances induced by the original model. Classical prior-correction methods (e.g., EM under label shift) and per-class logit shifts achieve (i) but can scramble rankings via instance-specific softmax renormalization; popular multiclass calibrators (temperature, matrix/vector scaling, Dirichlet) address calibration error but do not enforce (ii). We formulate calibration as Euclidean projection onto the intersection of two closed convex sets: a row-simplex set and a columnwise isotonic+sum set, where isotonic order within each class is defined by the model's pre-calibration scores. We give a Dykstra alternating-projection solver and an ADMM splitting; for the column subproblem we provide an exact one-pass PAV solution followed by a scalar shift to satisfy the column sum constraint. We also provide nearly-isotonic variants that trade tiny, controlled violations for calibration gains. Experiments show we can match target totals while preserving within-class orderings and maintaining competitive ECE/Brier scores across neural and tabular benchmarks. Our Python implementation is open source.[1]

1 Introduction

Calibration of probabilistic classifiers is essential for decision-making where predicted probabilities should reflect empirical frequencies. In many applications (health, finance, survey weighting), practitioners must both (a) adjust predictions to match known class totals and (b) maintain the model's

*Corresponding author: contact@gsood.com

[1] `https://github.com/finite-sample/rank_preserving_calibration`

within-class discrimination—the relative ordering of instances within each class.

In the multiclass setting ($J \geq 3$), naive per-class logit shifts (or label-prior corrections) can change instance orderings within a class because the softmax renormalization term varies by instance. Standard multiclass calibrators such as temperature scaling [Guo et al., 2017] and Dirichlet calibration [Kull et al., 2019] improve calibration but do not guarantee order preservation. Label-shift/ prior-correction methods [Saerens et al., 2002, Lipton et al., 2018] match marginals, but can induce the same ranking problem. We address this gap.

1.1 The Rank-Scrambling Problem

Given original probabilities p_{ij} and per-class multiplicative weights w_j, the adjusted probabilities

$$q_{ij} \;=\; \frac{w_j p_{ij}}{\sum_{k=1}^{J} w_k p_{ik}}$$

depend on the instance-specific denominator. Hence $p_{i_1 j} > p_{i_2 j}$ need not imply $q_{i_1 j} > q_{i_2 j}$.

Worked example (3 classes, two instances).

$$p_{1,\cdot} = (0.51,\, 0.48,\, 0.01), \quad p_{2,\cdot} = (0.49,\, 0.02,\, 0.49), \quad w = (1,\, 100,\, 1).$$

Then $q_{11} = 0.51/(0.51 + 48 + 0.01) \approx 0.0105$ while $q_{21} = 0.49/(0.49 + 2 + 0.49) \approx 0.1644$, reversing the original class-1 order.

1.2 Contributions

1. **Rank-preserving calibration as projection:** We project the original probability matrix onto the intersection of (a) the row-simplex and (b) per-column isotonic+sum sets, where isotonic order is the total preorder induced by the original scores for that class.

2. **Algorithms:** We develop (i) a Dykstra alternating-projection method and (ii) an ADMM splitting. The column projection is solved *exactly* by a single PAV pass followed by a scalar shift to match the target column sum.

3. **Nearly isotonic extensions:** We provide ϵ-slack and hinge-penalty variants to tune the calibration–discrimination trade-off.

4. **Empirical validation:** Across neural and tabular classifiers, we match targets while preserving within-class weak orders (ties respected) and maintaining competitive calibration metrics.

2 Related Work

Label/prior shift and matching marginals. When deployment priors differ from training priors, the EM update of Saerens et al. [2002] and later black-box shift estimation [Lipton et al., 2018] adjust predicted posteriors to match new class totals without refitting. These methods are effective for marginals but apply instance-specific renormalization that can reverse within-class orderings in $J \geq 3$.

Multiclass calibration. Temperature scaling and class-wise linear maps (vector/matrix scaling) [Guo et al., 2017] and Dirichlet calibration [Kull et al., 2019] are widely used but do not enforce within-class ranking constraints. Recent work also explores ranking-aware criteria [Ma and Blaschko, 2021] and normalization-aware isotonic approaches for multiclass settings [Arad and Rosset, 2025], which are complementary; neither simultaneously enforces rank preservation *and* matches *given* class totals.

3 Problem Formulation

Let $P \in \mathbb{R}_+^{N \times J}$ be the matrix of predicted probabilities, with rows on the probability simplex. We seek $Q \in \mathbb{R}_+^{N \times J}$ such that:

1. **Row-simplex:** $\sum_{j=1}^{J} Q_{ij} = 1$ for all i and $Q_{ij} \geq 0$.
2. **Column sums:** $\sum_{i=1}^{N} Q_{ij} = T_j$ for targets $T_1, \ldots, T_J$.
3. **Isotonic columns:** For each class j, sorting rows by the original P_{ij} (stable, ties respected), the sequence $(Q_{1j}, \ldots, Q_{Nj})$ is nondecreasing.

We solve the projection

$$\min_{Q} \ \frac{1}{2}\|Q - P\|_F^2 \quad \text{s.t.} \quad Q \in S_{\text{row}} \cap I_{\text{cols}}. \tag{1}$$

3.1 Feasibility

If $\sum_{j=1}^{J} T_j = N$ and $0 \le T_j \le N$ for all j, the intersection is nonempty: the constant-column matrix $Q_{ij} = T_j/N$ lies on the row simplex, matches the column totals, and is isotone (constant) in each column.

4 Algorithms

4.1 Dykstra's Alternating Projection Algorithm

Algorithm 1 Dykstra's method for rank-preserving calibration

Require: $P \in \mathbb{R}^{N \times J}$, targets $T \in \mathbb{R}^J$, tolerance ε, max_iters

Initialize: $Q^{(0)} = P$, $U^{(0)} = 0$, $V^{(0)} = 0$
for $k = 1$ to max_iters **do**
  $Y^{(k)} = Q^{(k-1)} + U^{(k-1)}$ (add correction)
  $Z^{(k)} = \Pi_{\text{row}}(Y^{(k)})$ (project each row onto the simplex)
  $U^{(k)} = Y^{(k)} - Z^{(k)}$ (update correction)
  $W^{(k)} = Z^{(k)} + V^{(k-1)}$ (add correction)
  $Q^{(k)} = \Pi_{\text{iso,sum}}(W^{(k)}, T)$ (per-column isotone + fixed sum)
  $V^{(k)} = W^{(k)} - Q^{(k)}$ (update correction)
  if $\|Q^{(k)} - Q^{(k-1)}\|_F / \max(1, \|Q^{(k-1)}\|_F) < \varepsilon$ **then**
    break
  end if
end for
return $Q^{(k)}$

Row projection Π_{row}. We use a fast probability-simplex projection per row with $O(J \log J)$ time via sorting [Duchi et al., 2008] (or Condat's variant) [Condat, 2016].

4.2 Exact Column Projection: Isotone + Sum Constraint

For each column j, let $w \in \mathbb{R}^N$ be the current column vector arranged in the fixed order that sorts $P_{\cdot j}$ once and for all (stable, ties respected). We seek

$$\min_{x \in \mathbb{R}^N} \tfrac{1}{2}\|x - w\|_2^2 \quad \text{s.t.} \quad x_1 \le \cdots \le x_N, \ \sum_{i=1}^{N} x_i = T_j.$$

Solution. Compute $y = \mathrm{PAV}(w)$ (one pass). Then set the scalar shift

$$c \;=\; \frac{T_j - \sum_{i=1}^{N} y_i}{N} \quad \text{and} \quad x^\star \;=\; y + c \cdot \mathbf{1}.$$

Correctness. The isotone set is closed under addition of a constant vector, and the Euclidean projection onto a translation-invariant convex set commutes with that translation; the KKT system for the sum constraint introduces a single Lagrange multiplier that is exactly the constant shift above. Note: nonnegativity is *not* enforced in this subproblem; it is handled by the subsequent row-simplex projection.

Algorithm 2 Column projection $\Pi_{\mathrm{iso,sum}}$ (single PAV + scalar shift)

Require: Column $w \in \mathbb{R}^N$ in fixed score order, target sum T

1: $y \leftarrow \mathrm{PAV}(w)$
2: $c \leftarrow (T - \sum_i y_i)/N$
3: **return** $y + c \cdot \mathbf{1}$

Complexity. Per Dykstra iteration: row projections cost $O(NJ \log J)$; column projections are $O(NJ)$ given pre-sorted order. The orders per column are computed once upfront in $O(JN \log N)$.

4.3 ADMM Formulation

Introduce $Q = R = S$, with $R \in S_{\mathrm{row}}$ and $S \in I_{\mathrm{cols}}$:

$$\min_{Q,R,S} \; \frac{1}{2}\|Q - P\|_F^2 \quad \text{s.t. } R \in S_{\mathrm{row}},\; S \in I_{\mathrm{cols}},\; Q = R = S.$$

Standard ADMM updates apply [Boyd et al., 2011]. The R-update is the row-simplex projection; the S-update is Algorithm 2. Warm-starts reuse PAV block structure and previous shifts.

5 Nearly Isotonic Variants

Epsilon-slack constraints. Replace $x_{i+1} \geq x_i$ with $x_{i+1} \geq x_i - \epsilon$. Let $\tilde{x}_i = x_i + i\epsilon$. Then $\tilde{x}$ is standard isotone, and the sum constraint becomes $\sum_i \tilde{x}_i = T + \epsilon \sum_i i$. Apply Algorithm 2 to $\tilde{w}$ with target $\tilde{T}$, then recover $x_i = \tilde{x}_i - i\epsilon$.

Hinge-penalty approach. Add $\lambda \sum_i \max(0, x_i - x_{i+1})$ to the objective. The proximal step can be implemented efficiently via splitting: a standard isotonic projection substep plus a shrinkage on violations; we provide a practical solver in the package docs.

6 Theoretical Notes

Convergence. For closed convex sets, Dykstra's method converges to the Euclidean projection onto the intersection [Bauschke and Borwein, 1994]. Quantitative rates depend on geometry: polyhedral regularity can yield linear or even finite convergence in special cases; in worst cases convergence can be arbitrarily slow. We report iteration counts and primal residuals in experiments. ADMM convergence is standard for strongly convex objectives with closed convex constraints [Boyd et al., 2011, Tibshirani, 2017].

Rank preservation. Because column isotonicity is enforced in the fixed order induced by $P_{\cdot j}$ with stable ties, the calibrated $Q_{\cdot j}$ preserves the original weak order (ties may remain ties).

7 Computational Considerations

- **One-time sorting:** For each class j, compute and cache the permutation that sorts $P_{\cdot j}$ once. All iterates reuse this order.
- **Row projection choice:** We use the $O(J \log J)$ sorter-based simplex projection [Duchi et al., 2008]; Condat's variant is a drop-in alternative with strong empirical performance [Condat, 2016].
- **Warm starts:** Dykstra and ADMM both benefit from reusing PAV block structure and previous scalar shifts.

8 Experiments

Baselines. We compare to (i) EM prior correction [Saerens et al., 2002], (ii) BBSE reweighting [Lipton et al., 2018], (iii) temperature scaling [Guo et al., 2017], (iv) Dirichlet calibration [Kull et al., 2019], and (v) normalization-aware multiclass isotonic calibrators [Arad and Rosset, 2025].

Metrics. We report ECE, classwise ECE, Brier score, and *rank-preservation rate* (fraction of within-class pairs whose order is unchanged), plus top-K stability (Jaccard of top-K instance sets per class), column-total deviation $\max_j |\sum_i Q_{ij} - T_j|$, and row-simplex deviation. Because ECE has binning pathologies, we follow recent guidance on binning and class-conditioning [Vaicenavicius et al., 2019, Nixon et al., 2019] and include reliability diagrams.

Stress tests. We evaluate on (a) CNNs (e.g., ResNet-50) on CIFAR-100, (b) gradient boosting and random forests on UCI tabular data, and (c) label-shifted holdouts to diagnose rank scrambling in prior-correction baselines.

Summary. Our method matches column totals to numerical tolerance, preserves within-class weak orderings by construction, and attains competitive calibration error. Nearly-isotonic variants yield smooth calibration–discrimination trade-offs.

9 Discussion and Limitations

The approach is suitable when relative ordering matters (regulation, triage, survey weighting). It is less suitable when the base model has poor discrimination (no ranks to preserve) or when computational budgets prohibit iterative projections on very large N (where stochastic variants may help). Extending to KL/Bregman geometries would connect to classical matrix scaling/IPF and is a promising direction.

10 Conclusion

We presented a projection-based multiclass calibrator that preserves within-class rankings while matching population totals. The method is simple (PAV+shift inside Dykstra/ADMM), efficient in practice, and plugs neatly into existing pipelines.

References

Adam Arad and Saharon Rosset. Improving multi-class calibration through normalization-aware isotonic techniques. In *International Conference on Machine Learning (ICML)*, 2025. URL `https://openreview.net/forum?id=PuVmGAggkU`. OpenReview preprint.

Heinz H Bauschke and Jonathan M Borwein. Dykstra's alternating projection algorithm for two sets. *Journal of Approximation Theory*, 79(3): 418–443, 1994.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Laurent Condat. Fast projection onto the simplex and the ℓ_1 ball. *Mathematical Programming*, 158(1-2):575–585, 2016. doi: 10.1007/s10107-015-0946-6.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, pages 272–279, 2008.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330, 2017.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems*, pages 12316–12326, 2019.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018. URL `https://arxiv.org/abs/1802.03916`.

Xingchen Ma and Matthew B. Blaschko. Meta-cal: Well-controlled post-hoc calibration by ranking. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. URL `https://arxiv.org/abs/2105.04290`.

Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, 2019. URL `https://openaccess.thecvf.com/content_CVPRW_2019/html/Uncertainty%20and%20Robustness%20in%20Deep%20Visual%20Learning/Nixon_Measuring_Calibration_in_Deep_Learning_CVPRW_2019_paper.html`.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002. doi: 10.1162/089976602753284446.

Ryan J Tibshirani. Dykstra's algorithm, ADMM, and coordinate descent: connections, insights, and extensions. *Advances in Neural Information Processing Systems*, 30, 2017.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B. Schön. Evaluating model calibration in classification. *Proceedings of Machine Learning Research (AISTATS)*, 89:3459–3467, 2019. URL `https://proceedings.mlr.press/v89/vaicenavicius19a.html`.