Calibration Where It Counts: Cost- and Data-Informed Isotonic Regression[*]

Gaurav Sood[†]

October 26, 2025

Abstract

Thresholded decisions turn probability errors into utility losses. Pooling-based monotone calibrators such as isotonic regression are flexible and reliable but can collapse distinct scores into the same probability on wide *plateaus*. We adopt a decision-economic view: choose and tune a calibrator that reduces deployment cost by improving reliability where it affects decisions and by preserving discrimination where it matters (Elkan, 2001; Vickers and Elkin, 2006).

We contribute three pieces. First, a practical *diagnostic suite*—global bootstrap tie stability, within-plateau concordance, minimum detectable difference, and progressive-sampling diversity—with a decision rule that labels plateaus as *supported*, *limited-data*, or *inconclusive*. Second, two adaptive calibrators: **Relaxed PAVA**, which allows bounded local slack and reduces via a cumulative shift to one weighted isotonic projection; and **density-aware smoothed isotonic**, which smooths locally then projects back onto the monotone cone. Third, a formal, convex **cost- and data-informed isotonic** calibrator (CDI-ISO) that encodes economically weighted, variance-aware *local minimum-slope* constraints and solves in $O(n)$ time through a single isotonic pass (Barlow et al., 1972; Robertson et al., 1988). We implement all methods in a scikit-learn–compatible library and evaluate with proper scores and decision curves.

1 Introduction

A probabilistic predictor is *calibrated* when predicted probabilities match empirical frequencies (DeGroot and Fienberg, 1983). Calibration matters because real deployments threshold

[*] https://github.com/finite-sample/calibre

[†] Gaurav can be reached at contact@gsood.com

probabilities: a small bias around the operating point can swing treatment, triage, or approval decisions (Niculescu-Mizil and Caruana, 2005; Jiang et al., 2012). Isotonic regression is a popular post-hoc calibrator because it is flexible and enforces monotonicity (Zadrozny and Elkan, 2002). Its strength is also a liability: by pooling adjacent violators into constant blocks, isotonic often produces broad *plateaus* that erase within-block discrimination.

Economic lens. Thresholds encode asymmetric costs (Elkan, 2001). Decision-curve analysis (DCA) summarizes utility as net benefit versus threshold, with threshold odds $t/(1-t)$ weighting false positives relative to true positives (Vickers and Elkin, 2006). From this perspective, calibration is economically instrumental: better reliability near the operating point reduces expected loss, while avoidable plateaus can undercut useful ranking.

Aim. We seek to (i) diagnose whether a plateau is *supported* by data or merely *limited* by sample size, and (ii) adapt the calibrator locally when data and operating costs justify recovering discrimination.

1.1 Contributions

Our organizing principle is simple: *reduce decision cost without gratuitously destroying discrimination.*

- **Diagnostics → action.** A practical suite—global bootstrap tie stability, within-plateau concordance (WPC; a Wilcoxon/Mann–Whitney–style concordance inside the tied region), minimum detectable difference (two-proportion power near plateau boundaries), and progressive-sampling diversity—feeds a decision rule that labels plateaus as *supported*, *limited-data*, or *inconclusive*. Each label maps to an action (keep, relax, or smooth).

- **Algorithms with linear-time solvers.** (a) *Relaxed PAVA* allows bounded local slack and reduces to a single weighted isotonic projection via a cumulative shift (inherits $O(n)$ complexity on a total order (Barlow et al., 1972; Robertson et al., 1988)); (b) *density-aware smoothed isotonic* uses local windows, smooths, and reprojects to the monotone cone (Ramsay, 1998; Meyer, 2008; Jiang et al., 2011).

- **Formal cost- and data-informed calibration.** *CDI-ISO* is a convex projection with *local minimum-slope* constraints that are evidence-gated and economically weighted, solved exactly by one isotonic pass. This unifies standard isotone (Zadrozny and Elkan, 2002), relaxed isotone (cf. penalty-based Tibshirani et al., 2011 and ENIR Pakdaman Naeini and Cooper, 2016), and minimum-slope isotone in a single formulation.

2 Background and Related Work

We summarize how calibration is defined and used, how isotonic regression produces plateaus, and how prior work relaxes or smooths monotone fits. We close with decision-aware perspectives and the openings they leave.

2.1 What calibration is and why it is used

A predictor is calibrated if among examples assigned probability p, a fraction p are positive (DeGroot and Fienberg, 1983). Calibration matters because thresholding converts probability error into decision losses (Niculescu-Mizil and Caruana, 2005; Jiang et al., 2012). Post-hoc methods fall into two families. *Parametric* approaches—Platt/logistic scaling (Platt, 1999), beta calibration (Kull et al., 2017), and temperature/matrix/vector scaling for deep nets (Guo et al., 2017)—fit low-parameter transformations. *Nonparametric* approaches—histogram binning (Zadrozny and Elkan, 2001) and isotonic regression (Zadrozny and Elkan, 2002)—trade smoothness for flexibility and monotonicity.

2.2 Isotonic regression and the origin of plateaus

Univariate isotonic fits a nondecreasing sequence by minimizing weighted squared error subject to order constraints. The Pool Adjacent Violators Algorithm (PAVA) solves this in $O(n)$ time on a total order (Ayer et al., 1955; Barlow et al., 1972; Robertson et al., 1988). PAVA's solution is a right-continuous, nondecreasing *step function*: adjacent violators are merged into contiguous *blocks* and each block is assigned its weighted mean, so all scores in a block receive the same $\hat{p}$ (Barlow et al., 1972; Robertson et al., 1988). These constant blocks are the *plateaus*. They can reflect true flatness in the underlying calibration function, or they can be artifacts of sparsity and noise that force pooling in finite samples; both behaviors are observed in practice (Zadrozny and Elkan, 2002).

2.3 Relaxing or smoothing monotone fits

Two strands try to mitigate over-coarse plateaus while retaining monotonicity. **Relaxations** penalize violations instead of forbidding them. Nearly-isotonic regression tunes a penalty to trade bias and variance (Tibshirani et al., 2011); ENIR calibrates by ensembling near-isotonic fits (Pakdaman Naeini and Cooper, 2016). **Smooth monotone models** impose smoothness alongside monotonicity via basis functions and constraints (Ramsay, 1998; Meyer, 2008); in calibration, "smooth isotonic" tempers stepwise fits with local smoothing before reimposing monotonicity (Jiang et al., 2011). These approaches choose *how much* to relax or smooth, but not *where* it is justified by data or *why* it matters economically.

2.4 Decision-aware perspectives

Cost-sensitive learning formalizes how false-positive/false-negative costs move the operating threshold (Elkan, 2001). DCA summarizes *net benefit* versus threshold using the odds $t/(1-t)$ to weight false positives (Vickers and Elkin, 2006). These tools clarify *which* thresholds matter, but they do not change the calibration mapping itself.

Openings. We identify three gaps: (i) practical diagnostics that separate genuine from data-limited plateaus; (ii) *local*, evidence-based mechanisms that break ties only where supported; and (iii) integration of decision costs into the calibration mapping, so recovering discrimination is done where it improves utility.

3 Problem Setup and Decision-Economic Framework

Let s denote model scores, $g(s) \in [0, 1]$ the calibrator, and $\pi(s) = \mathbb{P}(y{=}1 \mid s)$ the (unknown) truth. Thresholding at t yields a decision $a \in \{0, 1\}$. DCA defines net benefit

$$\mathrm{NB}(t) \;=\; \mathrm{TPR}(t)\,\pi \;-\; \mathrm{FPR}(t)\,(1-\pi)\,\frac{t}{1-t},$$

where $\frac{t}{1-t}$ are threshold odds (Vickers and Elkin, 2006). We report NB curves to connect calibration to utility.

For model selection we consider

$$\mathcal{J}(g; \lambda, \mathcal{T}) = \mathbb{E}[S(y, g(s))] + \lambda\,\mathbb{E}_{t\sim\mathcal{T}}[\mathrm{DiscLoss}_t(g)],$$

with S a proper score (Brier or NLL) (Gneiting and Raftery, 2007). The term $\mathrm{DiscLoss}_t$ penalizes ties/ranking loss localized near threshold t (e.g., $1 - \mathrm{WPC}$ inside a plateau that covers t). We use $\mathcal{J}$ to guide hyperparameters; we do not claim Bayes-risk optimality.

4 Plateau Diagnostics

Let (s_i, y_i) be calibration data sorted by s_i and $\hat{p}_i = g(s_i)$ the fitted probabilities.

Definition 1 (Plateau) *A plateau is a maximal index interval $P = [i_s, i_e]$ with $\hat{p}_{i_s} = \cdots = \hat{p}_{i_e}$. We summarize P by its score span $[s_{\min}, s_{\max}]$, size $|P|$, and label variance.*

Global bootstrap tie stability. Resample the *entire* calibration set (optionally stratified by score quantiles), refit $g^{(b)}$, and evaluate $g^{(b)}$ at the observed scores in $[s_{\min}, s_{\max}]$. Stability τ is the fraction of bootstraps with empirical range $< \epsilon$ on that span; stable plateaus persist under global refits.

Within-plateau concordance (WPC). With P_+ and P_- the positives/negatives in P,

$$\mathrm{WPC}_P = \frac{1}{|P_+||P_-|} \sum_{i \in P_+} \sum_{j \in P_-} \mathbf{1}\{s_i > s_j\},$$

with a Wilcoxon/Mann–Whitney test against 0.5 (ties scored 0.5) to detect residual ranking.

Minimum detectable difference (MDD). At P's boundaries, a two-proportion power calculation estimates the minimum $\Delta = \mu_1 - \mu_0$ detectable at level α; large MDD implies low power to distinguish flatness from slope.

Progressive-sampling diversity. Fit isotonic on subsamples of size n and track a tie metric (or unique-value ratio). Increasing curves suggest additional data would reduce plateaus.

Decision rule. *Supported*: high τ, WPC ≈ 0.5, small MDD. *Limited-data*: low τ or WPC far from 0.5 or large MDD. *Inconclusive*: otherwise. Thresholds are tuned on held-out validation to avoid optimism.

5 Methods

5.1 Relaxed PAVA: local slack with $O(n)$ reduction

Standard PAVA solves $\min_{z_1 \leq \cdots \leq z_n} \sum_i w_i (y_i - z_i)^2$ in $O(n)$ time on a total order (Ayer et al., 1955; Barlow et al., 1972; Robertson et al., 1988). We allow bounded local violations by imposing adjacent-difference lower bounds

$$z_{i+1} - z_i \;\geq\; L_i, \qquad L_i \leq 0.$$

Data-adaptive L_i can come from adjacent *block-mean* differences (percentile rule) or from a variance-aware bound

$$L_i = -\, z_\alpha \sqrt{\hat{p}(1-\hat{p})\left(\tfrac{1}{n_i} + \tfrac{1}{n_{i+1}}\right)},$$

where $\hat{p}$ is the pooled rate of neighboring blocks and n_i their sizes. A cumulative shift $R_{i+1} = R_i + L_i$ reduces the problem to a single weighted isotonic projection on $y'_i = y_i - R_i$, with solution $z^\star_i = u^\star_i + R_i$; clipping to $[0, 1]$ preserves monotonicity because clipping is non-decreasing (Barlow et al., 1972; Robertson et al., 1988).

5.2 CDI-ISO: cost- and data-informed isotonic

We encode *economically weighted, variance-aware minimum slope* near operating thresholds and allow variance-aware relaxation elsewhere. Let $\Delta_i = z_{i+1} - z_i$. We solve the convex projection

$$\min_{z \in \mathbb{R}^n} \sum_{i=1}^{n} w_i (y_i - z_i)^2 \quad \text{s.t.} \quad \Delta_i \geq L_i \quad (i = 1, \ldots, n-1), \tag{1}$$

with $L_i = \phi_i - \varepsilon_i$. Here $\phi_i \geq 0$ is an economically weighted minimum slope near thresholds; $\varepsilon_i \geq 0$ is a variance-aware relaxation elsewhere.

Constructing L_i. Let $\bar{s}_i = (s_i + s_{i+1})/2$ and define economics weights

$$w_i^{\text{econ}} \;=\; \mathbb{E}_{t \sim \mathcal{T}}[\, K_h(|\bar{s}_i - t|)\,],$$

with triangular kernel K_h (half-width h) concentrating mass near thresholds of interest (from cost ratios or DCA Elkan, 2001; Vickers and Elkin, 2006). For data evidence, compute a lower confidence bound for adjacent block differences using the pooled standard error:

$$\Delta_i^{\text{LCB}} = \big(\hat{p}_{i+1} - \hat{p}_i\big) - z_\alpha \sqrt{\hat{p}(1 - \hat{p})\Big(\tfrac{1}{n_i} + \tfrac{1}{n_{i+1}}\Big)}.$$

Define

$$\phi_i = \gamma\, w_i^{\text{econ}}\, [\Delta_i^{\text{LCB}}]_+, \qquad \varepsilon_i = (1 - w_i^{\text{econ}})\, z_\alpha \sqrt{\hat{p}(1 - \hat{p})\Big(\tfrac{1}{n_i} + \tfrac{1}{n_{i+1}}\Big)},$$

with $\gamma \in [0, 1]$ setting a global slope budget; if $\Delta_i^{\text{LCB}} \leq 0$ we set $\phi_i = 0$ (no enforced slope without evidence).

Linear-time solver. Let $R_1{=}0$, $R_{i+1}{=}R_i{+}L_i$, and set $u_i = z_i - R_i$, $y'_i = y_i - R_i$. Then (1) reduces to

$$u^\star = \arg\min_{u_1 \leq \cdots \leq u_n} \sum_i w_i (y'_i - u_i)^2, \quad z^\star_i = u^\star_i + R_i,$$

which is a single weighted isotonic projection solvable in $O(n)$ time on a total order (Barlow et al., 1972; Robertson et al., 1988). Clipping $z^\star$ to $[0, 1]$ preserves monotonicity.

5.3 Density-aware smoothed isotonic

We first smooth locally, then project to the monotone cone: (1) choose a kNN/quantile window W_i around s_i; (2) apply local regression to obtain $\tilde{y}_i$ (clip to $[0,1]$); (3) run a monotone projection (PAVA) on $\tilde{y}$ to obtain $\hat{y}^{\text{smooth}} = \Pi_{\text{mono}}(\tilde{y})$ (Ramsay, 1998; Meyer, 2008; Jiang et al., 2011). CDI-ISO can replace the final projection when minimum-slope constraints are desired.

6 Theory: scope and guarantees

Feasibility and complexity. Weighted PAVA on a total order is $O(n)$ (Barlow et al., 1972; Robertson et al., 1988). The shift-to-PAVA reduction preserves this complexity and ensures feasibility for any real L_i; clipping to $[0,1]$ preserves monotonicity.

Power near plateaus. Let boundary means be μ_0, μ_1 with $\Delta = \mu_1 - \mu_0$. A Hoeffding-style back-of-the-envelope suggests that detecting $\Delta > 0$ with error $\leq \delta$ requires local effective sample size $O(\log(1/\delta)/\Delta^2)$. This is a heuristic for interpreting diagnostic power, not a minimax rate.

7 Conclusion

Economics-guided diagnostics and CDI-ISO help practitioners keep calibration honest without sacrificing discrimination where it affects decisions. By encoding evidence- and cost-informed local constraints in a convex projection with a linear-time solver, we make it practical to calibrate where it counts.

References

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26(4):641–647.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression.* John Wiley & Sons, New York.

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

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 973–978.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pages 1321–1330. PMLR.

Jiang, X., Osl, M., Kim, J., and Ohno-Machado, L. (2012). Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274.

Jiang, X., Osl, M., and Ohno-Machado, L. (2011). Smooth isotonic regression: A new method to calibrate predictive models. *Journal of the American Medical Informatics Association*, 18(3):277–282.

Kull, M., Silva Filho, T. M., and Flach, P. (2017). Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pages 623–631. PMLR.

Meyer, M. C. (2008). Inference using shape-restricted regression splines. *The Annals of Applied Statistics*, 2(3):1013–1033.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pages 625–632. ACM.

Pakdaman Naeini, M. and Cooper, G. F. (2016). Binary classifier calibration using an ensemble of near isotonic regression models. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 360–369.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Smola, A. J., Bartlett, P., Schölkopf, B., and Schuurmans, D., editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

Ramsay, J. O. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):365–375.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. John Wiley & Sons, Chichester, UK.

Tibshirani, R. J., Höfling, H., and Tibshirani, R. (2011). Nearly-isotonic regression. *Technometrics*, 53(1):54–61.

Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574.

Zadrozny, B. and Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pages 609–616. Morgan Kaufmann.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699. ACM.

A Optional note on binomial intervals

For constructing Δ_i^{LCB} one can replace the normal-approximation SE with Wilson-style intervals, which have better finite-sample behavior; see Brown, Cai, and DasGupta (2001). If you prefer to avoid extra dependencies, the normal approximation used in the main text is consistent with standard practice.