Brenier Isotonic Regression

Han Bao[1] **Amirreza Eshraghi**[2] **Yutong Wang**[2]
[1]The Institute of Statistical Mathematics [2]Illinois Institute of Technology

Abstract

Isotonic regression (IR) is a shape-constrained regression to maintain a univariate fitting curve non-decreasing, which has numerous applications. When it comes to multivariate responses, IR is no longer applicable because monotonicity is not readily extendable. We consider a multi-output regression problem where a regression function is *cyclically monotone*. Roughly speaking, a cyclically monotone function is the gradient of some convex potential. Whereas enforcing cyclic monotonicity is apparently challenging, we leverage the fact that Kantorovich's optimal transport (OT) always yields a cyclically monotone coupling. This naturally allows us to interpret a regression function and the convex potential as a link function in generalized linear models and Brenier's potential in OT, respectively. We call this IR extension *Brenier isotonic regression*. We demonstrate applications to probability calibration and single-index models.

1 INTRODUCTION

Given a labeled dataset $\{(z_i, y_i)\}_{i\in[n]}$ with an input $z_i \in \mathbb{R}$ and response $y_i \in \mathbb{R}$, isotonic regression (IR) returns nonparametric estimates $\{\widehat{y}_i\} \subseteq \mathbb{R}$ to solve

$$\min_{\widehat{y}_1,\dots,\widehat{y}_n\in\mathbb{R}} \sum_{i\in[n]} (y_i - \widehat{y}_i)^2$$
$$\text{subject to} \quad (z_i - z_j)(\widehat{y}_i - \widehat{y}_j) \geq 0 \quad \forall (i,j) \in [n]^2.$$

IR is nonparametric, imposing no regression model but only enforcing monotonicity. This specific quadratic program can be solved efficiently by the Pool Adjacent Violaters (PAV) algorithm (Ayer et al., 1955), which achieves the optimal solution of IR provably (Robertson et al., 1988; Best and Chakravarti, 1990). IR has been extensively used in probability calibration of binary classifiers (Zadrozny and Elkan, 2002; Niculescu-Mizil and Caruana, 2005) and single-index models (Kalai and Sastry, 2009; Kakade et al., 2011). Moreover, IR has been recently used for the Bayes error estimation (Ushio et al., 2025).

When both inputs and responses become multivariate, the extension of IR becomes scarce. Consider a labeled dataset $\{(\mathbf{z}_i, \mathbf{y}_i)\}_{i\in[n]}$ with $\mathbf{z}_i, \mathbf{y}_i \in \mathbb{R}^d$, then what is a natural extension of univariate monotonicity?

Such a multi-output "IR" is essential in probability calibration. In binary classification, a covariate $\mathbf{x}_i$ is classified based on a univariate probability $\widehat{p}_i \in [0,1]$ representing the confidence of a binary label $y_i \in \{0,1\}$. IR applied to $\{(\widehat{p}_i, y_i)\}_{i\in[n]}$ typically ameliorates the quality of the confidence. Yet, multiclass classification deals with multinomial probability $\widehat{\mathbf{p}}_i \in \triangle^{d-1}$ representing the confidence of a one-hot label $\mathbf{y}_i \in \{0,1\}^d$, where $\triangle^{d-1}$ is the probability simplex. Since IR is limited to 1D responses, the one-vs-rest (OvR) formulation has been applied in this case (Zadrozny and Elkan, 2002; Niculescu-Mizil and Caruana, 2005). The OvR formulation requires an additional normalization step after learning a regressor for each class, and what is worse, each regressor is independent without taking care of class correlation (Arad and Rosset, 2025).

To extend IR to the multi-class case, let us first consider generalized linear models (GLMs) for multinomial outputs (Agarwal et al., 2014; Lam et al., 2023):

$$\mathbb{E}[Y|X = \mathbf{x}] = \nabla\Phi(\mathbf{W}^*\mathbf{x}) := \varphi(\mathbf{W}^*\mathbf{x}), \qquad (1)$$

where $\mathbf{W}^* \in \mathbb{R}^{d\times D}$ is the weight matrix, $\Phi : \mathbb{R}^d \to \mathbb{R}$ is a proper convex lower-semicontinuous (l.s.c.) function, $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ is the inverse link function, and $Y \in \{0,1\}^d$ is a one-hot class vector. The choice $\varphi = \nabla\Phi$ is standard in literature of canonical proper losses (Buja et al., 2005; Gneiting and Raftery, 2007; Williamson et al., 2016; Bao and Takatsu, 2024). In view of Fenchel–Young losses (Blondel et al., 2020; Bao and Sugiyama, 2021), the convex potential Φ can be regarded as a generalized log-partition function or free energy. The virtue of multinomial GLMs (1) is rooted in the following fact: Φ is proper l.s.c. convex if and

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

only if φ has a *cyclically monotone* graph (Rockafellar, 1966). Therefore, we consider regression with an unknown but cyclically monotone link, which we call *cyclically monotonic isotonic regression (CMIR)*:

$$\min_{\widehat{\mathbf{y}}_1,\dots,\widehat{\mathbf{y}}_n \in \triangle^{d-1}} \sum_{i\in[n]} \|\mathbf{y}_i - \widehat{\mathbf{y}}_i\|_2^2 \quad \text{subject to} \\ \widehat{\mathbf{y}}_i = \varphi(\mathbf{z}_i) \text{ for some cyclically monotone } \varphi. \tag{2}$$

We propose a *nonparametric* solution to CMIR (2). Thus far, Agarwal et al. (2014) models the inverse link φ with a finite number of bases, and Lam et al. (2023) models the convex potential Φ with input-convex neural networks (Amos et al., 2017), but both are *parametric* and may be subject to the approximation error. To nonparametrically solve CMIR (2), we leverage an elegant connection between convexity and *optimal transport (OT)*. Specifically, the celebrated Brenier's theorem states that an optimal transport map can be represented by the gradient of a convex potential (Brenier, 1991). Thus, we can consider first transporting $\{\mathbf{z}_i\}$ to $\{\widehat{\mathbf{y}}_i\}$ and then minimizing the regression error—reformulating CMIR into a bi-level optimization with OT being the inner problem. Inspired by Brenier's theorem, we call our approach *Brenier isotonic regression* (BrenierIR). We implement this bi-level program with `scipy` in an end-to-end fashion, which estimates the implicit gradient with the finite difference method (Virtanen et al., 2020).

Experimentally, we test BrenierIR on calibration and single-index model tasks. In particular, it performs surprisingly well and stably on the calibration task. BrenierIR is more principled with the connection to GLMs, and does not require complicated hyperparameter tuning. Hence, we suggest practitioners to test BrenierIR when calibrating multiclass classifiers.

1.1 Additional related work

Han et al. (2019) and Fang et al. (2021) considered IR with multivariate covariates, but the response remains real-valued and coordinate-wise monotonicity is considered. Sasabuchi et al. (1983) considered multi-output IR, but again coordinate-wise monotonicity is considered. The main drawback of coordinate-wise monotonicity is that it does not capture GLMs. Even in the simplest case of multinomial logistic regression, the softmax is not coordinate-wise monotone but *is* cyclically monotone (Gao and Pavel, 2017).

The connection between monotonicity and OT has been the basis of conditional vector quantile (Carlier et al., 2016; Chernozhukov et al., 2017). In particular, vector quantile estimation has been actively studied in terms of scalability (Rosenberg et al., 2023), statistical property (Hallin et al., 2021), and neural network modeling (Kondratyev et al., 2025). While they focus on *estimating* conditional vector quantiles, we optimize the regression error by *fitting* vector quantiles. To our knowledge, Dawkins et al. (2022) is the only previous work focusing on this monotone fitting problem with an application in economics, while they mainly focus on the PAC learnability analysis. The connection to OT is absent in their work, and establishing this connection is a central contribution of ours.

The connection between OT and regression might be apparently elusive because OT deals with two "uncoupled" sets but regression cares about "coupled" in-out relations. In fact, OT has been applied only to uncoupled isotonic regression (Rigollet and Weed, 2019; Slawski and Sen, 2024)—the unused result Rigollet and Weed (2019, Proposition 2.3) interestingly demonstrates that the pushforward is an isometry between the L_p- and p-Wasserstein distances. Paty et al. (2020) implied a connection between (coupled) isotonic regression and OT but only in 1D case. Berta et al. (2024) extended IR to the multiclass case by preserving AUC. However, the multiclass extension of AUC lacks a canonical definition, which limits the approach. Thus, our work fills the missing piece in these lines of work.

2 BACKGROUND

Notation. $\mathscr{P}(\mathbb{R}^d)$ denotes the set of Borel probability measures. The support of $\mu \in \mathscr{P}(\mathbb{R}^d)$ is denoted by $\operatorname{supp}(\mu) \coloneqq \left\{\mathbf{z} \in \mathbb{R}^d \;\middle|\; \mu(\mathbf{z}) > 0\right\}$. We write the Dirac measure by $\delta_\bullet$. We write the all-ones vector by $\mathbb{1}_n \in \mathbb{R}^n$. The probability simplex is denoted by $\triangle^{d-1} \coloneqq \left\{\mathbf{p} \in \mathbb{R}^d_{\geq 0} \;\middle|\; \langle \mathbf{p}, \mathbb{1}_d\rangle = 1\right\}$. The set of permutations over $[n]$ is written by $\operatorname{Perm}(n)$. The Lebesgue space $L^p(\mu)$ is the space of measurable functions whose p-th moment with respect to a measure μ is finite.

Optimal transport. Given a cost $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and probability measures $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$, the Monge problem is an OT problem formulated as follows:

$$\inf_T \left\{ \int_{\mathbb{R}^d} c(\mathbf{z}, T(\mathbf{z}))\, \mathrm{d}\mu(\mathbf{z}) \;\middle|\; T_\sharp \mu = \nu \right\}, \tag{3}$$

where $T : \mathbb{R}^d \to \mathbb{R}^d$ is a measurable map and $T_\sharp\mu$ denotes the pushforward of μ by T, defined as $(T_\sharp\mu)(A) = \mu(T^{-1}(A))$ for every Borel $A \subseteq \mathbb{R}^d$. If T^* attains the infimum, T^* is called the optimal transport map.

We typically relax the challenging constraint in the Monge problem by the following Kantorovich problem:

$$\inf_\pi \left\{ \int_{\mathbb{R}^d\times\mathbb{R}^d} c(\mathbf{z}, \mathbf{u})\, \mathrm{d}\pi(\mathbf{z}, \mathbf{u}) \;\middle|\; \pi \in \mathcal{U}(\mu, \nu) \right\}, \\ \text{where} \quad \mathcal{U}(\mu,\nu) \coloneqq \{\pi \mid \Pi_{1\sharp}\pi = \mu,\ \Pi_{2\sharp}\pi = \nu\}. \tag{4}$$

and $\Pi_{1\sharp}$ and $\Pi_{2\sharp}$ are the pushforward by the projections $\Pi_1(\mathbf{z},\mathbf{u}) = \mathbf{z}$ and $\Pi_2(\mathbf{z},\mathbf{u}) = \mathbf{u}$, respectively. This is an infinite-dimensional linear program over $\mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d)$, and the solution π_* is called the optimal coupling of μ and ν (whenever it exists).

The relationship between the Monge and Kantorovich problems is clear for the discrete case. Consider the discrete measures $\mu = \frac{1}{n}\sum_{i\in[n]} \delta_{\mathbf{z}_i}$ and $\nu = \frac{1}{n}\sum_{j\in[n]} \delta_{\mathbf{u}_j}$. Let $\mathbf{C} \in \mathbb{R}^{n\times n}$ be the cost defined by $C_{ij} \coloneqq c(\mathbf{z}_i, \mathbf{u}_j)$. The discrete Monge problem is

$$\min_{\sigma\in\mathrm{Perm}(n)} \frac{1}{n}\sum_{i\in[n]} C_{i\sigma(i)}. \tag{5}$$

Further, let us introduce the (scaled) Birkhoff polytope

$$\mathcal{B}(m,n) \coloneqq \left\{ \mathbf{P} \in \mathbb{R}^{m\times n}_{\geq 0} \;\middle|\; m\mathbf{P}\mathbb{1}_n = \mathbb{1}_m, n\mathbf{P}^\top \mathbb{1}_m = \mathbb{1}_n \right\}.$$

The discrete Kantorovich problem is

$$\min_{\mathbf{P}\in\mathcal{B}(n,n)} \langle \mathbf{C}, \mathbf{P}\rangle, \tag{6}$$

where $\langle \mathbf{C}, \mathbf{P}\rangle \coloneqq \sum_{i,j} C_{ij}P_{ij}$ is the transport cost. The polytope $\mathcal{B}(n,n)$ enforces $n\mathbf{P}$ to be doubly stochastic. The following well-known/celebrated proposition states that an optimizer for the Kantorovich problem (6) is a permutation matrix. For a permutation $\sigma \in \mathrm{Perm}(n)$, we write $\mathbf{P}^\sigma$ for the permutation matrix:

$$\forall (i,j) \in [n]^2, \quad P^\sigma_{ij} = \begin{cases} 1/n & \text{if } j = \sigma_i, \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 1 (Panaretos and Zemel (2020, Proposition 1.3.1))**.** *There exists an optimal solution for the discrete Kantorovich problem (6), $\mathbf{P}^{\sigma^*}$, which is a permutation matrix associated to an optimal permutation $\sigma^* \in \mathrm{Perm}(n)$ to the discrete Monge problem (5). Moreover, if $\{\sigma^*_1, \dots, \sigma^*_M\} \subseteq \mathrm{Perm}(n)$ is the set of optimal permutations to the discrete Monge problem (5), the set of optimal solutions to the discrete Kantorovich problem (6) is the convex hull of $\left\{\mathbf{P}^{\sigma^*_1}, \dots, \mathbf{P}^{\sigma^*_M}\right\}$.*

Intuitively, Proposition 1 says that we can recover a transport map from the optimal coupling. This exact correspondence holds only when the source and target distributions are supported uniformly on the same number of points. This stems from the fact that the set of extremal points of the Birkhoff polytope matches the set of permutation matrices (Birkhoff, 1946).

Cyclic monotonicity. We need to recall additional mathematical machinery from convex analysis. We keep using a generic symmetric cost $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, though we are mostly interested in the squared L_2 cost. Refer to Villani (2008, Section 5) for more details.

Definition 1. *A relation $\Gamma \subseteq \mathbb{R}^d \times \mathbb{R}^d$ is said to be a* c-cyclically monotone *if, for any $m \in \mathbb{N}$ and any family $\{(\mathbf{z}_i, \mathbf{u}_i)\}_{i\in[m]} \subseteq \Gamma$, the following inequality holds:*

$$\sum_{i=1}^m c(\mathbf{z}_i, \mathbf{u}_i) \leq \sum_{i=1}^m c(\mathbf{z}_i, \mathbf{u}_{i+1}), \tag{7}$$

where we use the convention $\mathbf{u}_{m+1} = \mathbf{u}_1$. A coupling $\pi \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d)$ is said to be c-cyclically monotone *if* $\mathrm{supp}(\pi)$ *is c-cyclically monotone.*

Informally, a cyclically monotone coupling is locally optimal in the following sense: The incurred transport cost (i.e., the left-hand side of (7)) must grow if we locally perturb any of the coupled points.

Definition 2. *A function $\Phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is said to be* c-convex *if $\{\mathbf{z} \mid \Phi(\mathbf{z}) < \infty\} \neq \emptyset$ and there exists $\zeta : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ such that*

$$\forall \mathbf{z} \in \mathbb{R}^d, \quad \Phi(\mathbf{z}) = \inf_{\mathbf{u}\in\mathbb{R}^d} c(\mathbf{z},\mathbf{u}) - \zeta(\mathbf{u}). \tag{8}$$

The c-transform *of Φ defined as*

$$\forall \mathbf{u} \in \mathbb{R}^d, \quad \Phi^c(\mathbf{u}) = \inf_{\mathbf{z}\in\mathbb{R}^d} c(\mathbf{z},\mathbf{u}) - \Phi(\mathbf{z}),$$

and its c-subdifferential *is the following c-cyclically monotone relation:*

$$\partial_c\Phi \coloneqq \left\{(\mathbf{z},\mathbf{u}) \in \mathbb{R}^d \times \mathbb{R}^d \;\middle|\; \Phi^c(\mathbf{u}) + \Phi(\mathbf{z}) = c(\mathbf{z},\mathbf{u})\right\}.$$

The c-subdifferential of Φ at point $\mathbf{z}$ is

$$\partial_c\Phi(\mathbf{z}) = \left\{\mathbf{u} \in \mathbb{R}^d \;\middle|\; (\mathbf{z},\mathbf{u}) \in \partial_c\Phi\right\}.$$

Consider the cost $c_{\|\cdot\|}(\mathbf{z},\mathbf{u}) = \|\mathbf{z}-\mathbf{u}\|_2^2$ and note that $c_{\|\cdot\|}$-cyclic monotonicity is equivalent to cyclic monotonicity with respect to $c_{\langle\cdot,\cdot\rangle}(\mathbf{z},\mathbf{u}) \coloneqq -\langle \mathbf{z},\mathbf{u}\rangle$ (Smith and Knott, 1992). Then $c_{\langle\cdot,\cdot\rangle}$-subdifferential recover the usual subdifferential. Thus, below, when we say cyclic monotonicity without specifying the cost, the cost is implicitly assumed to be $c_{\langle\cdot,\cdot\rangle}$.

The following proposition is a fundamental characterization of convex functions. Therein, a relation Γ is said to be *maximally* monotone if no other monotone relation is strictly larger than it.

Proposition 2 (Rockafellar (1966, Theorem 3))**.** *A function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is convex and $\{\mathbf{z} \mid \Phi(\mathbf{z}) < \infty\} \neq \emptyset$ if and only if there exists a maximal cyclically monotone relation $\Gamma \subseteq \mathbb{R}^d \times \mathbb{R}^d$ such that $\partial\Phi = \Gamma$. Moreover, Φ is unique up to an arbitrary additive constant.*

Monotonicity of transport maps. We recall the optimality characterization of the Kantorovich problem (4) via cyclic monotonicity.

Proposition 3 (Villani (2008, Theorem 5.10)). *Let $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be l.s.c. such that $c(\mathbf{z}, \mathbf{u}) \geq a(\mathbf{z}) + b(\mathbf{u})$ $(\forall (\mathbf{z}, \mathbf{u}) \in \mathbb{R}^d \times \mathbb{R}^d)$ for some real-valued upper semicontinuous $a \in L^1(\mu)$ and $b \in L^1(\nu)$.[1] Assume that the optimal cost of (4) is finite. Then the optimal coupling $\pi_* \in \mathcal{U}(\mu, \nu)$ is c-cyclically monotone.*

This result is useful because we can obtain a cyclically monotone coupling only by solving the linear program—if we consider the discrete Kantorovich problem (6), standard solvers solve it with $O(n^3)$ time complexity (Bonneel et al., 2011), which is far cheaper than testing all $O(2^n)$ subsets of a graph as in Definition 1. Under standard cost c, the optimal coupling is readily cyclically monotone with any regular measures.

Barycentric map. In general, the optimal coupling π_* does not give a single-valued transport map. Hence, we need to define the following barycentric map (Ambrosio et al., 2005, Definition 5.4.2). When π admits the disintegration $\pi = \int \pi_{\mathbf{z}} \, \mathrm{d}\mu(\mathbf{z})$,

$$T_\pi(\mathbf{z}) = \int_{\mathbb{R}^d} \mathbf{u} \, \mathrm{d}\pi_{\mathbf{z}}(\mathbf{u}),$$

which maps each source point $\mathbf{z}$ to a weighted barycenter of its neighbors in the target $\mathrm{supp}(\nu)$.

Beyond Proposition 3, we can establish a stronger result to ensure the existence of a convex potential that induces the optimal transport map. This closes the loop to relate Kantorovich to Monge in general cases.

Proposition 4 (Brenier (1991)). *Consider the Monge problem (3) with $c(\mathbf{z}, \mathbf{u}) = \|\mathbf{z} - \mathbf{u}\|_2^2$. If μ has a density with respect to the Lebesgue measure, then there exists a unique optimal transport map T that satisfies $T = \nabla\Phi$ for some differentiable convex function $\Phi : \mathbb{R}^d \to \mathbb{R}$ such that $(\nabla\Phi)_\sharp \mu = \nu$.*

Make sure that this characterization is valid only when source μ is non-singular. Otherwise, we must allow a transport plan to become multivalued $T_\pi(\mathbf{z}) \coloneqq \{\mathbf{u} \mid (\mathbf{z}, \mathbf{u}) \in \mathrm{supp}(\pi)\}$. In parallel, T_π is a *sub*gradient of some convex potential (Peyré, 2024).

3 BRENIER IR

Consider the following CMIR (cyclically monotone isotonic regression) problem.

For a dataset $\{(\mathbf{z}_i, \mathbf{y}_i)\}_{i\in[n]}$, where $\mathbf{z}_i \in \mathbb{R}^d$ is an input and $\mathbf{y}_i \in \{0, 1\}^d$ is a one-hot encoded label, solve the following problem:

$$\min_{\widehat{\mathbf{y}}_1, \ldots, \widehat{\mathbf{y}}_n \in \mathbb{R}^d} \frac{1}{n} \sum_{i\in[n]} \|\mathbf{y}_i - \widehat{\mathbf{y}}_i\|_2^2 \tag{9}$$

such that $\{(\mathbf{z}_i, \widehat{\mathbf{y}}_i)\}_{i\in[n]}$ is cyclically monotone.

The cyclically monotone constraint in CMIR can be reformulated as a discrete Kantorovich problem (6), instead of directly unrolling Definition 1.

1. Instead of optimize in $\{\widehat{\mathbf{y}}_i\}_{i\in[n]}$, we solve the following BrenierIR problem.

$$\begin{aligned} \min_{\mathbf{u}_1, \ldots, \mathbf{u}_n \in \triangle^{d-1}} \; & \frac{1}{n} \|\mathbf{Y} - n\mathbf{P}\mathbf{U}\|_{\mathrm{F}}^2 \\ \text{subject to} \quad & \mathbf{P} \in \operatorname*{arg\,min}_{\mathbf{P}\in\mathcal{B}(n,n)} \langle \mathbf{C}, \mathbf{P} \rangle, \end{aligned} \tag{10}$$

   where $\mathbf{Y}, \mathbf{U} \in \mathbb{R}^{n\times d}$ store $\{\mathbf{y}_i\}$ and $\{\mathbf{u}_j\}$ in rows, respectively. We specify the inner problem below.

2. The inner problem of (10) is the discrete Kantorovich problem (6) with source $\mu = \frac{1}{n}\sum_{i\in[n]} \delta_{\mathbf{z}_i}$, target $\nu = \frac{1}{n}\sum_{j\in[n]} \delta_{\mathbf{u}_j}$, and the cost $C_{ij} = \|\mathbf{z}_i - \mathbf{u}_j\|_2^2$. Denote its solution by $\mathbf{P}^* \in \mathcal{B}(n, n)$.

3. With the barycentric map (Ferradans et al., 2014)

$$T_{\mathbf{P}^*}(\mathbf{z}_i) = \frac{\sum_{j\in[n]} P^*_{ij} \mathbf{u}_j}{\sum_{j\in[n]} P^*_{ij}} = n \sum_{j\in[n]} P^*_{ij} \mathbf{u}_j, \tag{11}$$

   the prediction is given by $\widehat{\mathbf{y}}_i = T_{\mathbf{P}^*}(\mathbf{z}_i)$.

Unlike the original CMIR, the inner problem of (10) can be solved efficiently in practice via the implementation in Fig. 2 (described later). The latent variable $\{\mathbf{u}_j\}$ can be viewed as the *vector quantiles* of the observed variable $\{\mathbf{z}_i\}$ (Carlier et al., 2016). Thus, the transport associated to the coupling can be interpreted as a transformation from the raw input to the vector quantiles. In most of the scenarios, the optimal coupling $\mathbf{P}^*$ is associated with a permutation σ^* (see Proposition 1). Then the barycentric map $T_{\mathbf{P}^*}$ is the optimal permutation from $\{\mathbf{z}_i\}$ to $\{\mathbf{u}_j\}$, and $\mathbf{z}_i \mapsto \mathbf{u}_{\sigma^*(i)}$ has a cyclically monotone graph, thanks to Proposition 3. Therefore, we have encoded the cyclic monotonicity constraint via the Kantorovich problem.

Univariate case. Consider a 1D dataset $\{(z_i, y_i)\} \subseteq \mathbb{R} \times \mathbb{R}$ and latents $\mathbf{u} \coloneqq (u_i)_{i\in[n]} \subseteq \mathbb{R}^n$ instead. With the cost $C_{ij} = (z_i - u_j)^2$, the regression problem is

$$\min_{\substack{\mathbf{u}, \widehat{\mathbf{y}} \in \mathbb{R}^n \\ \widehat{\mathbf{y}} = n\mathbf{P}\mathbf{u}}} \frac{1}{n} \sum_{i\in[n]} (y_i - \widehat{y}_i)^2 \quad \text{subject to} \quad \mathbf{P} \in \mathcal{P}(\mathbf{u}), \tag{12}$$

where $\mathcal{P}(\mathbf{u}) \coloneqq \arg\min \{\langle \mathbf{C}, \mathbf{P} \rangle \mid \mathbf{P} \in \mathcal{B}(n, n)\}$. When $(z_i)_{i\in[n]}$ is increasing, we show that the set of $\widehat{\mathbf{y}}$ spans the whole space of nondecreasing sequences.

[1] The definition of lower semicontinuity can be found in Rockafellar (1970, Chapter 7).

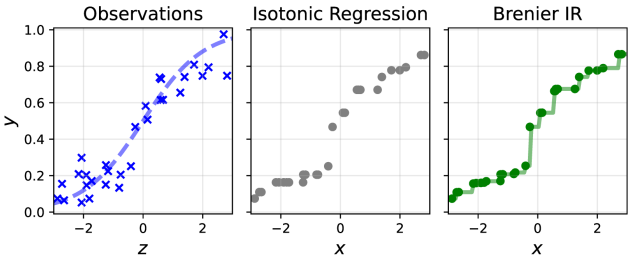

Figure 1: Comparison between IR and BIR. Observations are generated with $y = 1/(1+\exp(-z)) + \text{noise}$. The real line of BIR is computed via the Laguerre map over $\mathbb{R}$.

Theorem 1. *Suppose $z_1 < z_2 < \cdots < z_n$. Then the following equivalence holds:*

$$\{\widehat{\mathbf{y}} = n\mathbf{P}\mathbf{u} \mid \mathbf{P} \in \mathcal{P}(\mathbf{u}) \text{ for some } \mathbf{u} \in \mathbb{R}^n\} = \{\mathbf{v} \in \mathbb{R}^n \mid v_1 \le v_2 \le \cdots \le v_n\}.$$

As a corollary, (12) is exactly equivalent to the standard IR. Figure 1 demonstrates univariate BrenierIR (with the implementation described at the end of Section 3), which perfectly matches the PAV solution.

User-specified number of bins. The original bi-level program allows the vector quantiles $\{\mathbf{u}_j\}$ to be all distinct, i.e., it is possible that $|\operatorname{supp}(\nu)| = n$. Consequently, the inner problem in (10) is subject to $O(n^3)$ time complexity. To improve scalability in n, we implement an efficient variant with a hyperparameter $k \in \mathbb{N}$ to set the target measure $\nu := \frac{1}{k}\sum_{j\in[k]} \delta_{\mathbf{u}_j}$ such that $|\operatorname{supp}(\nu)| \le k (\ll n)$ is possible. The efficient version of (10) will be referred to as k-BrenierIR with k bins, which is defined as follows:

$$\begin{aligned} \min_{\mathbf{u}_1,\ldots,\mathbf{u}_k \in \triangle^{d-1}} \ & \frac{1}{n}\|\mathbf{Y} - n\mathbf{P}\mathbf{U}\|_{\mathrm{F}}^2 \\ \text{subject to} \ & \mathbf{P} \in \operatorname*{arg\,min}_{\mathbf{P}\in\mathcal{B}(n,k)} \langle \mathbf{C}, \mathbf{P}\rangle, \end{aligned} \tag{13}$$

where $\mathbf{U} \in \mathbb{R}^{k\times d}$ stores $\{\mathbf{u}_j\}_{j\in[k]}$ in each row. The hyperparameter $k \in \mathbb{N}$ represents the maximum number of bins and can be specified by a user. Usually, k must be at least greater than the number of classes d for a valid prediction. Given that classical IR trades off bias and variance adaptively (Zadrozny and Elkan, 2002), the choice of k may also affect the bias-variance tradeoff of BrenierIR.

Next, we formally state our first main theoretical contribution: BrenierIR yields a cyclically monotone regression function, regardless of whether the optimal coupling is a permutation matrix or not. All proofs can be found in Section A.

Theorem 2 (Cyclic monotonicity of $T_{\mathbf{P}^*}$)**.** *For fixed $\{\mathbf{u}_j\}_{j\in[k]}$ and the squared L_2 cost, let $\mathbf{P}^* \in \mathcal{B}(n,k)$ be the optimal transport from the observed inputs $\mu = \frac{1}{n}\sum_{i\in[n]}\delta_{\mathbf{z}_i}$ to $\nu = \frac{1}{k}\sum_{j\in[k]}\delta_{\mathbf{u}_j}$. Then the barycentric map $T_{\mathbf{P}^*}$ (11) has a cyclically monotone graph.*

Test prediction. By solving (13), we can obtain the barycentric map $\mathbf{z}_i \mapsto n\sum_{j\in[n]} P^*_{ij}\mathbf{u}_j$, but this is applicable to training inputs only. One way for test prediction is to re-compute (13) again after including test points, which is computational expensive. A more principled approach is to leverage Brenier's theorem (see Proposition 4). To do this, we need a non-singular source measure. Consider the kernel density estimator

$$\widetilde{\mu}(\mathbf{z}) := \frac{1}{nh}\sum_{i\in[n]} K\left(\frac{\mathbf{z}-\mathbf{z}_i}{h}\right)\mathscr{L}^d(\mathbf{z}), \tag{14}$$

where $\mathscr{L}^d$ is the Lebesgue measure in $\mathbb{R}^d$ and K is the standard normal density. We can apply Brenier's theorem to non-singular $\widetilde{\mu}$.

When source $\widetilde{\mu}$ is continuous and target ν is discrete, the transport problem is called semi-discrete. We introduce its solution based on Peyré and Cuturi (2019, Section 5). Given $\widetilde{\mu}$ and $\nu = k^{-1}\sum_j \delta_{\mathbf{u}_j}$ with the squared L_2 cost, the Kantorovich problem (4) admits the dual formula:

$$\sup_{\mathbf{g}\in\mathbb{R}^k}\left\{\int_{\mathbb{R}^d} g^c \,\mathrm{d}\widetilde{\mu} + \frac{1}{k}\langle \mathbf{g}, \mathbb{1}_k\rangle\right\}, \tag{15}$$

where $\mathbf{g} := [g(\mathbf{u}_1),\ldots,g(\mathbf{u}_k)]^\top$ is the Lagrangian multiplier with respect to ν, $g : \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ is the Kantorovich dual potential, and g^c is the c-transform of g. Once (15) is solved, consider the *Laguerre cells* associated to the Kantorovich potential $\mathbf{g}$:

$$\mathbb{L}_{\mathbf{g}}(\mathbf{u}_j) := \{\mathbf{z} \mid \forall j' \ne j,\ c(\mathbf{z},\mathbf{u}_j) - g_j \le c(\mathbf{z},\mathbf{u}_{j'}) - g_{j'}\}.$$

The Laguerre cells induce a disjoint partition such that $\mathbb{R}^d = \bigsqcup_j \mathbb{L}_{\mathbf{g}}(\mathbf{u}_j)$, which is a weighted extension of Voronoi tessellations (Aurenhammer, 1987). Given $\mathbf{z} \in \mathbb{R}^d$ and an optimizer g^* for (15), define the Laguerre map $T_{\mathbf{g}^*}$ by $T_{\mathbf{g}^*}(\mathbf{z}) = \mathbf{u}_j$, where $\mathbb{L}_{\mathbf{g}^*}(\mathbf{u}_j)$ is the unique partition such that $\mathbf{z} \in \mathbb{L}_{\mathbf{g}^*}(\mathbf{u}_j)$. We note that $T_{\mathbf{g}^*}$ is piecewise constant by construction. Therefore, the Laguerre map naturally extends the idea of the binning estimator to the multinomial case, and each quantile $\mathbf{u}_j$ serves as an individual bin level. The map obtained by a measurable selection from Laguerre cells also exhibits the cyclically monotone property.

Theorem 3 (Cyclic monotonicity of Laguerre map)**.** *Let μ be an absolutely continuous probability measure on $\mathbb{R}^d$ and $\nu = \sum_{j=1}^k b_j\delta_{\mathbf{u}_j}$ with $\mathbf{b} \in \triangle^{k-1}$. Let $\mathbf{g}^* \in \mathbb{R}^k$ be an optimal dual potential to semi-discrete*

```python
import numpy as np
import ot
from ot.utils.unif
from scipy.optimize import minimize

def obj(u_):
    u = u_.reshape(k, d)
    C = ot.dist(z, u)
    P = ot.emd(unif(n), unif(k), C)
    return np.sum((y - n*P@u)**2)/n

res = minimize(obj, [..] method='
   SLSQP')
u_opt = res.x.reshape(k, d)
```

Figure 2: `scipy` implementation of BrenierIR. Full code at `levelfour/Brenier_Isotonic_Regression`.

OT (15) *with the squared* L_2 *cost. If* $T_{\mathbf{g}^*}$ *is a measurable selection such that* $T_{\mathbf{g}^*}(\mathbf{z}) \in \{\mathbf{u}_j \mid \mathbf{z} \in \mathbb{L}_{\mathbf{g}^*}(\mathbf{u}_j)\}$*, then* $T_{\mathbf{g}^*}$ *has a cyclically monotone graph.*

As the Laguerre map $T_{\mathbf{g}^*}$ ends up with the gradient of Brenier's potential, we call our approach to CMIR as *Brenier isotonic regression*. The piecewise-constant nature is favorable for probability calibration, while the smoothness of T should be improved for specific applications (Paty et al., 2020; Peyré, 2024).

Implementation. We describe a practical implementation of (13). A solution to (10) is recovered as its special case. Generally speaking, bi-level programs including (13) is under a family of non-convex programs, called *mathematical programming with equilibrium constraints* (MPEC) (Luo et al., 1996), which has been an active research topic in the operations research community. In our early experiments, an MPEC-related approach (Hillbrecht, 2024) did not perform ideally, while theoretically grounded. Instead, we resort to simply estimating the derivative of the whole bi-level objective (13) in $\mathbf{U}$ by the finite difference method. This practically works surprisingly well, even though the objective (13) is not differentiable.[2] We summarize the implementation with `scipy` (Virtanen et al., 2020) and `pot` (Python OT) (Flamary et al., 2021) in Fig. 2, where the main procedure consists of only a few lines. With `jac=None` option, `scipy` uses the finite difference method to estimate the Jacobian. For the optimization method, we used Sequential Quadratic Programming (SQP) (Nocedal and Wright, 2006), specified by `method='SLSQP'`, which is an efficient second-order method and can deal with the linear constraints $\mathbf{u}_j \in \triangle^{d-1}$.

[2]This is because a tiny perturbation to $\mathbf{U}$ may bring the linear cost $\mathbf{C}$ from one normal cone of $\mathcal{B}(n,k)$ to another, where differentiability breaks down.

To compute the Laguerre cells, we need to solve semi-discrete OT (15). It is typically solved by the averaged SGD (Genevay et al., 2016) over fresh samples from $\widetilde{\mu}$ (14). While the time complexity is comparable to the primal Kantorovich problem, we adopt an additional heuristic to use the dual potential $\mathbf{g}^*$ returned by the primal Kantorovich problem without sampling fresh points. This simple procedure suffices practically. The Kantorovich potential can be obtained with `ot.emd`, which solves the primal Kantorovich problem, also known as the Earth Mover's Distance (EMD).

4 APPLICATION 1: PROBABILITY CALIBRATION

As the first application, we use BrenierIR for calibrating multiclass classifiers. Consider a probabilistic classifier $\widehat{\mathbf{p}} : \mathcal{X} \to \triangle^{d-1}$ that outputs class probabilities over d classes. It is called *calibrated* if, for every prediction vector $\mathbf{q} \in \triangle^{d-1}$, the conditional class proportions match the prediction:

$$\forall i \in [d], \quad \Pr(Y = i \mid \widehat{\mathbf{p}}(X) = \mathbf{q}) = q_i. \tag{16}$$

When a decision maker acts in response to a forecaster, calibration ensures "they speak the same language" (Foster and Hart, 2021). The measure equivalent to (16) is the L_1 calibration error (CE) (Murphy, 1973), while weaker measurements such as class-wise calibration error (cwCE) and expected calibration error (ECE) (Naeini et al., 2015) are also commonly used (Gruber and Buettner, 2022).

We are interested in recalibration (also known as post-hoc calibration): Given a classifier $\widehat{\mathbf{p}}$ and a small calibration set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in[n]} \subseteq \mathcal{X} \times \{0,1\}^d$, estimate the canonical calibration map (Vaicenavicius et al., 2019):

$$\eta(\mathbf{q}) = [\Pr(Y = i \mid \widehat{\mathbf{p}}(X) = \mathbf{q})]_{i\in[d]}.$$

We estimate η by BrenierIR applied on the pushforwarded calibration set $\{(\widehat{\mathbf{p}}(\mathbf{x}_i), \mathbf{y}_i)\}_{i\in[n]}$.

Benchmark. Table 1 compares BrenierIR (shortened as BIR_k) with baseline recalibrators: Bin (OvR binning) (Zadrozny and Elkan, 2001), IR (OvR isotonic regression) (Zadrozny and Elkan, 2002), MS (matrix scaling, extending Platt's scaling) (Guo et al., 2017), TS (temperature scaling) (Guo et al., 2017), Dir (Dirichlet calibration) (Kull et al., 2019), OI (order-invariant network) (Rahimi et al., 2020), and IRP (iterative recursive partitioning) (Berta et al., 2024). We recalibrated MLP (D-ReLU-100-ReLU-d with the L_2 regularization strength 10^{-4}, trained by Adam with 200 epochs) and linear SVM (with the L_2 regularization strength 40) and compared the recalibration performance by the L_1 calibration error (CE), which is the

Table 1: Recalibration results for MLP **(upper table)** and linear SVM **(lower table)**. Each number indicates the L_1 calibration error (lower is better) with averaging 10 trials, and bold-faced if the recalibrator achieves the best or second best performance or statistically indistinguishable from them by the Mann–Whitney U test (significance level: 5%).

Dataset \ Recalibrator	—	Bin	Dir	IRP	IR	MS	OI	TS	BrenierIR (Ours) $k=15$	BrenierIR (Ours) $k=30$	BrenierIR (Ours) $k=50$
balance-scale	0.244	0.184	0.108	**0.068**	0.139	0.171	0.140	0.177	**0.061**	0.070	0.084
car	0.063	0.050	**0.030**	**0.031**	0.034	0.037	0.132	0.036	0.045	0.040	0.042
cleveland	0.914	0.921	0.828	**0.224**	0.853	0.774	0.938	1.066	**0.519**	0.655	0.759
dermatology	0.178	0.187	0.153	0.204	**0.139**	0.167	0.798	0.163	**0.122**	0.159	0.170
glass	0.859	0.843	0.856	**0.574**	0.652	0.753	0.951	0.884	**0.579**	0.635	0.671
vehicle	0.294	0.300	0.208	**0.103**	0.199	0.310	0.474	0.298	0.177	**0.145**	0.202

Dataset \ Recalibrator	—	Bin	Dir	IRP	IR	MS	OI	TS	BrenierIR (Ours) $k=15$	BrenierIR (Ours) $k=30$	BrenierIR (Ours) $k=50$
balance-scale	0.110	0.236	0.283	**0.012**	0.268	0.160	0.698	0.160	**0.088**	0.096	0.106
car	**0.106**	0.284	0.332	0.558	0.266	0.413	0.717	0.589	**0.121**	0.179	0.173
cleveland	0.784	0.855	0.732	**0.255**	0.905	0.655	0.746	0.896	**0.573**	0.725	0.730
dermatology	0.253	0.289	0.192	0.314	**0.082**	0.144	0.436	0.635	0.139	0.128	**0.126**
glass	0.831	0.795	0.846	**0.044**	0.649	0.711	0.780	0.905	**0.647**	0.685	0.799
vehicle	0.553	0.444	0.536	**0.009**	0.456	0.515	0.600	0.567	**0.308**	0.402	0.450

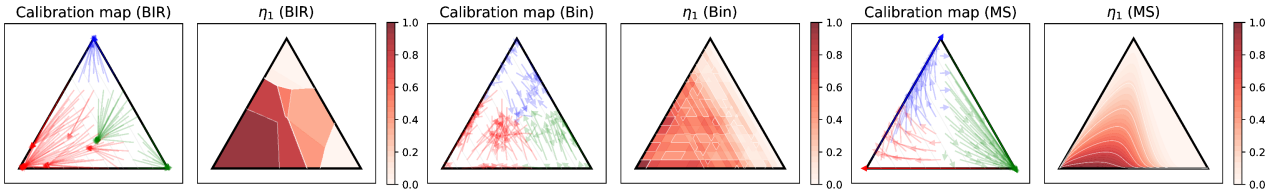

Figure 3: From left to right, BrenierIR ($k = 50$), Binning, and Matrix Scaling. We show their estimated calibration maps as vector fields (each left) and contour plots of their first coordinate (each right).

empirical estimate of $\mathbb{E}\,\|\Pr(Y|\widehat{\mathbf{p}}(X) = \mathbf{q}) - \mathbf{q}\|_1$ averaged across $\mathbf{q}$ in the same bin $B \in \triangle^{d-1}$. Here, the simplex binning is constructed by splitting each axis uniformly into 15 intervals. The further details of the baselines and implementations can be found in Section C.1. From our experimental results: (i) BrenierIR consistently outperforms many baselines and performs comparably with IRP. (ii) Even with relatively small k, BrenierIR matches IRP while scaling to larger numbers of classes, where IRP struggles. In Section C.2, we show the average running time of BrenierIR and IRP, demonstrating that BrenierIR scales better. Therefore, BrenierIR serves as a powerful recalibrator with a mild computational overhead.

We tested accuracy and classwise/confidence calibration error in Section C.4. For classwise/confidence calibration error, the overall trend remains similar to Table 1. For accuracy, BrenierIR performs on par with the other baselines, while IRP is no longer competitive.

Visualization. Figure 3 visualizes the estimated calibration maps, recalibrated on a linear SVM trained with the balance-scale dataset (Shultz et al., 1994). BrenierIR has a strong binning effect compared with Matrix Scaling. It also captures inter-class relations unlike the standard OvR Binning, whose contour levels are parallel to the simplex boundary. As expected from the test prediction by the Laguerre map, the calibration map of BrenierIR concentrates on only a few points. We visualize the calibration functions of the other baselines in Section C.3, where we can see qualitatively similar results between BrenierIR and IRP.

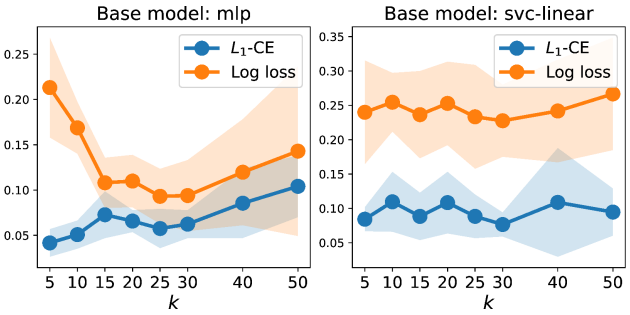

Figure 4: Each base model is recalibrated by BIR with different bin size k (shown with the standard deviation).

Tradeoff in k. We trained MLP and linear SVM with the same setup on the balance-scale dataset, and recalibrated with BIR with different bin sizes k. The trend shown in Fig. 4 is rather clear in the MLP case: The log loss has a sweet spot in k. CE exhibits a similar tradeoff for $k \in [15, 50]$, but CE decreases for

Table 2: Multiclass classification results. Each number indicates the accuracy and L_1 calibration error with averaging 10 trials, and bold-faced if a method achieves the best performance or statistically indistinguishable from the best method by the Mann–Whitney U test (significance level 5%).

Metric	Accuracy (↑)				L_1 calibration error (↓)			
Method / Dataset	Log	CLS	LT	BIR_{30} (Ours)	Log	CLS	LT	BIR_{30} (Ours)
balance-scale	0.860	0.682	**0.910**	**0.895**	0.227	0.507	0.170	**0.117**
car	0.657	0.509	**0.964**	0.808	0.831	0.899	**0.086**	0.120
cleveland	**0.585**	0.343	**0.582**	**0.585**	1.002	0.682	0.863	**0.470**
glass	0.630	0.328	**0.684**	0.595	1.030	0.682	0.793	**0.608**

Algorithm 1: Brenier Single Index Model

Input: $T_{\max}$ total number of iters, $\mathbf{U}_0$ randomly initialized
Output: $\mathbf{W}_{T_{\max}}$ learned index, $\mathbf{U}_{T_{\max}}$ learned target support

for $0 \le t \le T_{\max} - 1$ **do**
  ▷ `Update linear predictor (W-step)`
  $\mathbf{W}_{t+1} \leftarrow \arg\min_{\mathbf{W}} J(\mathbf{W}, \mathbf{U}_t) + \frac{\lambda_W}{2}\|\mathbf{W}\|_{\mathrm{F}}^2$
  ▷ `Update target measure (U-step)`
  $\mathbf{U}_{t+1} \leftarrow \arg\min_{\mathbf{U}} J(\mathbf{W}_{t+1}, \mathbf{U})$
end

very small k. This reflects predictions averaged over the simplex rather than genuine calibration. The trend is less clear in the SVM case, but it calibrates slightly better around $k \simeq 30$. Overall, generalization analysis of recalibration is still underrepresented (Fujisawa and Futami, 2025), and further analysis is left for future work.

5 APPLICATION 2: SINGLE-INDEX MODELS

Let us revisit single-index models with multinomial outputs. Given a dataset $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in[n]}$ for a covariate $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^D$ and a one-hot encoded label $\mathbf{y}_i \in \{0,1\}^d$, we posit the multiclass classification model

$$\mathbf{y}_i \mid \mathbf{x}_i \sim \mathrm{Categorical}(\varphi(\mathbf{W}^*\mathbf{x}_i)), \tag{17}$$

where both the true index $\mathbf{W}^* \in \mathbb{R}^{d\times D}$ and the link function $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ are known. We assume the graph of φ is cyclically monotone, as is usual to assume link monotonicity in SIMs.

Algorithm 1 describes our approach to learn the model (17) via BrenierIR ($k = 30$). We minimize the objective function

$$J(\mathbf{W}, \mathbf{U}) \coloneqq \frac{1}{n}\|\mathbf{Y} - n\mathbf{P}\mathbf{U}\|_{\mathrm{F}}^2 \quad \text{subject to } \mathbf{P} \in \operatorname*{arg\,min}_{\mathbf{P}\in\mathcal{B}(n,k)} \langle \mathbf{C}(\mathbf{W}), \mathbf{P}\rangle,$$

and the cost $\mathbf{C}(\mathbf{W})$ is $C_{ij} \coloneqq \|\mathbf{W}\mathbf{x}_i - \mathbf{u}_j\|_2^2$. Both $\mathbf{W}$- and $\mathbf{U}$-steps can be implemented via SQP with the finite difference method, as in Fig. 2. This alternating procedure is reminiscent of the calibrated least squares (Agarwal et al., 2014, Algorithm 2), which alternates the residual fitting and the link fitting steps.

We compared with the following baselines: Log is the standard multinomial logistic regression with the L_2 regularization strength 1.0 (optimized by L-BFGS); CLS is the calibrated least squares (Agarwal et al., 2014). We ran 100 and 2000 iters for BIR and CLS, respectively. CLS requires introducing bases of the monotone link, for which we used $\{z, z^2, z^3\}$ as the original paper suggests. LT is the LegendreTron (Lam et al., 2023), which learns the inverse link via ICNNs together with the linear predictor. The baseline details are again deferred to Section C.

Table 2 demonstrates the results of multiclass classification on each dataset. We measured the accuracy and L_1 calibration error of each learned SIM. BIR consistently outperforms CLS, which also models SIMs explicitly but parametrically. This admits how a nonparametric approach performs nicely in practice. While BIR greatly improves calibration over LT, classification performance of BIR is not ideal. Therefore, we recommend practitioners to use BIR only for recalibration for now and leave developing better classification algorithms for future work.

6 DISCUSSION

BrenierIR as adaptive binning. Binning and IR are the two most common recalibrators (Zadrozny and Elkan, 2002). Binning is a simple nonparametric approach yet yielding a nice bias-variance tradeoff, by averaging confidence prediction falling into the same bin (Zadrozny and Elkan, 2001). IR can be viewed as an adaptive binning such that the bin boundaries are determined depending on the base model (detailed in Guo et al. (2017)). While their OvR extensions are insensitive to inter-class correlation—in fact, all contour levels are parallel to the simplex boundaries in Fig. 3—BrenierIR captures inter-class correlation

and yields class-adaptive simplex binning, as seen in Fig. 3. Such class-adaptive simplex binning is possible by IRP (Berta et al., 2024), but it needs to sweep the uniform grid points over the whole probability simplex, which is computationally demanding for large d. Thus, BrenierIR is a tractable and class-adaptive binning.

Future work. Last but not least, we recognize that the computational cost of BrenierIR remains a bottleneck. Even the inner OT requires $O(n^3)$ time complexity, and the total time complexity of the bi-level program is opaque. Given fairly practical performance of BrenierIR, further investigation from the optimization perspective is an interesting open direction.

ACKNOWLEDGMENT

HB is supported by JST PRESTO JPMJPR24K6. YW acknowledges support from NSF CISE CRII 2451714. The projected was initiated during YW's research visit to ISM.

References

Alekh Agarwal, Sham Kakade, Nikos Karampatziakis, Le Song, and Gregory Valiant. Least squares revisited: Scalable approaches for multi-class prediction. In *Proceedings of the 31st International Conference on Machine Learning*, pages 541–549, 2014.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer, 2005.

Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 146–155, 2017.

Alon Arad and Saharon Rosset. Improving multiclass calibration through normalization-aware isotonic techniques. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Franz Aurenhammer. Power diagrams: Properties, algorithms and applications. *SIAM Journal on Computing*, 16(1):78–96, 1987.

Miriam Ayer, H Daniel Brunk, George M Ewing, William T Reid, and Edward Silverman. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, pages 641–647, 1955.

Han Bao and Masashi Sugiyama. Fenchel-Young losses with skewed entropies for class-posterior probability estimation. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pages 1648–1656, 2021.

Han Bao and Asuka Takatsu. Proper losses regret at least 1/2-order. *arXiv preprint arXiv:2407.10417*, 2024.

Eugene Berta, Francis Bach, and Michael Jordan. Classifier calibration with ROC-regularized isotonic regression. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pages 1972–1980, 2024.

Michael J Best and Nilotpal Chakravarti. Active set algorithms for isotonic regression; a unifying framework. *Mathematical Programming*, 47(1):425–439, 1990.

Garrett Birkhoff. Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucuman, Ser. A*, 5:147–154, 1946.

Bernd Bischl, Giuseppe Casalicchio, Taniya Das, Matthias Feurer, Sebastian Fischer, Pieter Gijsbers, Subhaditya Mukherjee, Andreas C Müller, László Németh, Luis Oala, Lennart Purucker, Sahithya Ravi, Jan N van Rijn, Prabhant Singh, Joaquin Vanschoren, Jos van der Velde, and Marcel Wever. OpenML: Insights from 10 years and more than a thousand papers. *Patterns*, 6(7):101317, 2025.

Mathieu Blondel, André FT Martins, and Vlad Niculae. Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using Lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–12, 2011.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4): 375–417, 1991.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Technical Report*, 2005.

Guillaume Carlier, Victor Chernozhukov, and Alfred Galichon. Vector quantile regression: An optimal transport approach. *The Annals of Statistics*, 44 (3):1165–1192, 2016.

Victor Chernozhukov, Alfred Galichon, Marc Hallin, and Marc Henry. Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45 (1):223–256, 2017.

Quinlan Dawkins, Minbiao Han, and Haifeng Xu. First-order convex fitting and its application to economics and optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6480–6487, 2022.

BY Billy Fang, Adityanand Guntuboyina, and Bodhisattva Sen. Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation. *The Annals of Statistics*, 49(2):769–792, 2021.

Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7 (3):1853–1882, 2014.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(1), 2021.

Dean P Foster and Sergiu Hart. Forecast hedging and calibration. *Journal of Political Economy*, 129(12): 3447–3490, 2021.

Masahiro Fujisawa and Futoshi Futami. PAC-Bayes analysis for recalibration in classification. In *Proceedings of 42nd International Conference on Machine Learning*, 2025.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. *Advances in Neural Information Processing Systems*, 29, 2016.

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

Sebastian Gruber and Florian Buettner. Better uncertainty calibration via proper scores for classification and beyond. *Advances in Neural Information Processing Systems*, 35:8618–8632, 2022.

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

Marc Hallin, Eustasio del Barrio, Juan Cuesta-Albertos, and Carlos Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139–1165, 2021.

Qiyang Han, Tengyao Wang, Sabyasachi Chatterjee, and Richard J Samworth. Isotonic regression in general dimensions. *The Annals of Statistics*, 47(5): 2440–2471, 2019.

Sebastian Hillbrecht. Bilevel optimization of the kantorovich problem and its quadratic regularization part III: The finite-dimensional case. *arXiv preprint arXiv:2406.08992*, 2024.

Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. *Advances in Neural Information Processing Systems*, 24, 2011.

Adam Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

Vladimir Kondratyev, Alexander Fishkov, Nikita Kotelevskii, Mahmoud Hegazy, Remi Flamary, Maxim Panov, and Eric Moulines. Neural optimal transport meets multivariate conformal prediction. *arXiv preprint arXiv:2509.25444*, 2025.

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. *Advances in Neural Information Processing Systems*, 32, 2019.

Kevin H Lam, Christian Walder, Spiridon Penev, and Richard Nock. Legendretron: uprising proper multiclass loss learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 18454–18470. PMLR, 2023.

Zhi-Quan Luo, Jong-Shi Pang, and Daniel Ralph. *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press, 1996.

Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600, 1973.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632, 2005.

Jorge Nocedal and Stephen J Wright. *Numerical Optimization*. Springer, 2006.

Victor M. Panaretos and Yoav Zemel. *An Invitation to Statistics in Wasserstein Space*. Springer International Publishing, 2020.

François-Pierre Paty, Alexandre d'Aspremont, and Marco Cuturi. Regularity as regularization: Smooth and strongly convex brenier potentials in optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1222–1232, 2020.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Gabriel Peyré. Course notes on computational optimal transport, December 2024. URL https://mathematical-tours.github.io/book-sources/optimal-transport/CourseOT.pdf.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11 (5-6):355–607, 2019.

Amir Rahimi, Amirreza Shaban, Ching-An Cheng, Richard Hartley, and Byron Boots. Intra order-preserving functions for calibration of multi-class neural networks. *Advances in Neural Information Processing Systems*, 33:13456–13467, 2020.

Philippe Rigollet and Jonathan Weed. Uncoupled isotonic regression via minimum Wasserstein deconvolution. *Information and Inference: A Journal of the IMA*, 8(4):691–717, 2019.

T. Robertson, F.T. Wright, and R. Dykstra. *Order Restricted Statistical Inference*. Probability and Statistics Series. Wiley, 1988.

R. Tyrrell Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 1970.

Ralph Rockafellar. Characterization of the subdifferentials of convex functions. *Pacific Journal of Mathematics*, 17(3):497–510, 1966.

Aviv A Rosenberg, Sanketh Vedula, Yaniv Romano, and Alexander Bronstein. Fast nonlinear vector quantile regression. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

Syoichi Sasabuchi, Manabu Inutsuka, and DD Sarath Kulatunga. A multivariate version of isotonic regression. *Biometrika*, 70(2):465–472, 1983.

Thomas R Shultz, Denis Mareschal, and William C Schmidt. Modeling cognitive development on balance scale phenomena. *Machine Learning*, 16(1): 57–86, 1994.

Martin Slawski and Bodhisattva Sen. Permuted and unlinked monotone regression in $\mathbb{R}^d$: An approach based on mixture modeling and optimal transport. *Journal of Machine Learning Research*, 25(183):1–57, 2024.

Cyril Smith and Martin Knott. On hoeffding-fréchet bounds and cyclic monotone relations. *Journal of Multivariate Analysis*, 40(2):328–334, 1992.

Ryota Ushio, Takashi Ishida, and Masashi Sugiyama. Practical estimation of the optimal classification error with soft labels and calibration. *arXiv preprint arXiv:2505.20761*, 2025.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 3459–3467. PMLR, 2019.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2008.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Robert C Williamson, Elodie Vernet, and Mark D Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17:1–52, 2016.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the 18th International Conference on Machine Learning*, pages 609–616, 2001.

Bianca Zadrozny and Charles Elkan. 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, 2002.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, see Section 3.]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes, see Section 3.]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No, will be made public upon acceptance.]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes, see Section 3.]
   (b) Complete proofs of all theoretical results. [Yes, see Section A.]
   (c) Clear explanations of any assumptions. [Yes, see Section 3.]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes, see Section C.1.]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes, see Section C.1.]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes, see Sections 4 and 5.]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes, see Fig. 2.]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes, see Section C.1.]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes, see Section C.1.]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A PROOFS

A.1 Proof of Theorem 1

Before proceeding with the proof, we show the following claim: Under the assumption of $z_1 < z_2 < \cdots < z_n$, if a permutation matrix $\mathbf{P}^{\sigma^*}$ with a permutation $\sigma^* \in \mathrm{Perm}(n)$ is optimal to the Kantorovich problem $\mathcal{P}(\mathbf{u}) = \arg\min\{\langle \mathbf{C}, \mathbf{P}\rangle \mid \mathbf{P} \in \mathcal{B}(n,n)\}$ with the cost $C_{ij} = (z_i - u_j)^2$, then σ^* must be *order-preserving*, that is, $u_{\sigma^*(1)} \le u_{\sigma^*(2)} \le \cdots \le u_{\sigma^*(n)}$. This can be seen based on the following inequality:

$$\text{For any } i < k \text{ and } j < l \text{ such that } u_j \le u_l, \quad (C_{il} + C_{kj}) - (C_{ij} + C_{kl}) = 2(z_k - z_i)(u_l - u_j) \ge 0.$$

If a permutation $\sigma \in \mathrm{Perm}(n)$ has a "crossing" quadruple (i,j,k,l) such that $j = \sigma(i) > \sigma(k) = l$ but $i < k$, the "uncrossed" permutation

$$\sigma'(i_0) = \begin{cases} \sigma(i) & \text{if } i_0 = k, \\ \sigma(k) & \text{if } i_0 = i, \\ \sigma(i_0) & \text{otherwise} \end{cases}$$

yields a lower (or nonincreasing) transport cost than that of σ. Therefore, any permutation can be made order-preserving by uncrossing all crossing quadruples, and the Kantorovich problem is optimized at an order-preserving permutation.

The forward inclusion ($\subseteq$). Suppose that $\widehat{\mathbf{y}} \in \mathbb{R}^n$ can be written as $\widehat{\mathbf{y}} = n\mathbf{P}\mathbf{u}$ for some $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{P} \in \mathcal{P}(\mathbf{u})$. By Proposition 1, there exists $M \ge 1$ and $\{\sigma_1^*, \ldots, \sigma_M^*\} \subseteq \mathrm{Perm}(n)$ such that $\mathbf{P}$ can be represented as a convex combination of permutation matrices $\left\{\mathbf{P}^{\sigma_i^*}\right\}_{i\in[M]}$, where each $\mathbf{P}^{\sigma_i^*}$ is optimal to $\mathcal{P}(\mathbf{u})$. Let us write this convex combination by

$$\mathbf{P} = \sum_{i\in[M]} \lambda_i \mathbf{P}^{\sigma_i^*}$$

for some $(\lambda_i)_{i\in[M]} \in [0,1]^M$ such that $\lambda_1 + \cdots + \lambda_M = 1$. Now we have

$$\widehat{y}_j = n[\mathbf{P}\mathbf{u}]_j = n \sum_{i\in[M]} \lambda_i \cdot \frac{1}{n} \cdot u_{\sigma_i^*(j)} = \sum_{i\in[M]} \lambda_i u_{\sigma_i^*(j)} \quad \text{for any } j \in [n],$$

and for $j < j'$,

$$\widehat{y}_j = \sum_{i\in[M]} \lambda_i u_{\sigma_i^*(j)} \le \sum_{i\in[M]} \lambda_i u_{\sigma_i^*(j')} = \widehat{y}_{j'}$$

because the optimality of each $\mathbf{P}^{\sigma_i^*}$ indicates the order-preserving property of σ_i^*. Thus, $\widehat{y}_1 \le \widehat{y}_2 \le \cdots \le \widehat{y}_n$.

The reverse inclusion ($\supseteq$). Fix a fixed nondecreasing sequence $\mathbf{v} \in \mathbb{R}^n$. We choose $\mathbf{u} = \mathbf{v}$ and $\mathbf{P} = \frac{1}{n}\mathbf{I}_n$, where $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ is the identity matrix. This $\mathbf{P}$ corresponds to the identity permutation and is obviously order-preserving. Hence, $\mathbf{P}$ is optimal to the Kantorovich problem and

$$\widehat{y}_j = n[\mathbf{P}\mathbf{u}]_j = n \cdot \frac{1}{n} v_j = v_j.$$

This proves the reverse inclusion. □

A.2 Proof of Theorem 2

First, we recapitulate the setup. Let $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{z}_i}$ and $\nu = \frac{1}{k}\sum_{j=1}^k \delta_{\mathbf{u}_j}$, and $\mathbf{P}^* \in \mathcal{B}(n,k)$ be an optimal solution to the following problem:

$$\min_{\mathbf{P}\in\mathcal{B}(n,k)} \langle \mathbf{C}, \mathbf{P}\rangle, \quad \text{where } C_{ij} = \frac{1}{2}\|\mathbf{z}_i - \mathbf{u}_j\|_2^2 \text{ and } \mathcal{B}(n,k) = \left\{ \mathbf{P} \in \mathbb{R}^{n\times k}_{\ge 0} \;\middle|\; \mathbf{P}\mathbb{1}_k = \frac{1}{n}, \mathbf{P}^\top \mathbb{1}_n = \frac{1}{k} \right\}.$$

Let $\mathbf{f}^* \in \mathbb{R}^n$ and $\mathbf{g}^* \in \mathbb{R}^k$ be the optimal potentials to the Kantorovich dual problem, then we have

$$\begin{cases} f_i^* + g_j^* \le \frac{1}{2}\|\mathbf{z}_i - \mathbf{u}_j\|_2^2 \\ P_{ij}^* > 0 \implies f_i^* + g_j^* = \frac{1}{2}\|\mathbf{z}_i - \mathbf{u}_j\|_2^2 \end{cases} \tag{18}$$

for any $(i,j) \in [n] \times [k]$ due to the complementary slackness (see Peyré and Cuturi (2019, Section 2.5)). The barycentric map $T_{\mathbf{P}^*} : \{\mathbf{z}_1, \dots, \mathbf{z}_n\} \to \operatorname{conv}\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ is given by

$$T_{\mathbf{P}^*}(\mathbf{z}_i) = n \sum_{j=1}^{k} P^*_{ij} \mathbf{u}_j, \quad \text{for } i \in [n]$$

so that $n \sum_{j=1}^{k} P^*_{ij} = 1$ for each $i \in [n]$.

Now we show that there exists a proper convex function $\Phi : \mathbb{R}^d \to \mathbb{R}$ such that $T_{\mathbf{P}^*}(\mathbf{z}_i) \in \partial\Phi(\mathbf{z}_i)$ for every $i \in [n]$. Once we establish this fact, the graph of the barycentric map $\{(\mathbf{z}_i, T_{\mathbf{P}^*}(\mathbf{z}_i)) \mid i \in [n]\}$ immediately turns out to be cyclically monotone (thanks to Rockafellar (1970, Theorem 24.8)).[3] To show this, we construct the convex function as follows:

$$\Phi(\mathbf{z}) \coloneqq \max_{1 \le j \le k} \left\{ \langle \mathbf{z}, \mathbf{u}_j \rangle - \frac{1}{2}\|\mathbf{u}_j\|_2^2 + g^*_j \right\}. \tag{19}$$

First, the maximum of $\Phi(\mathbf{z}_i)$ is attained at the "active" $\mathbf{u}_j$, namely, any $\mathbf{u}_j$ with $P^*_{ij} > 0$. Indeed, for any $(i,j) \in [n] \times [k]$, the complementary slackness (18) implies

$$f^*_i + g^*_j \le \frac{1}{2}\|\mathbf{z}_i - \mathbf{u}_j\|_2^2 = \frac{1}{2}\|\mathbf{z}_i\|_2^2 + \frac{1}{2}\|\mathbf{u}_j\|_2^2 - \langle \mathbf{z}_i, \mathbf{u}_j \rangle,$$

which is rearranged as follows:

$$\langle \mathbf{z}_i, \mathbf{u}_j \rangle - \frac{1}{2}\|\mathbf{u}_j\|_2^2 + g^*_j \le \frac{1}{2}\|\mathbf{z}_i\|_2^2 - f^*_i.$$

By taking the maximum over $j \in [k]$, this gives

$$\Phi(\mathbf{z}_i) \le \frac{1}{2}\|\mathbf{z}_i\|_2^2 - f^*_i,$$

and its equality holds when $P^*_{ij} > 0$. Thus, we have $\Phi(\mathbf{z}_i) = \langle \mathbf{z}_i, \mathbf{u}_j \rangle - \frac{1}{2}\|\mathbf{u}_j\|_2^2 + g^*_j$ for active $\mathbf{u}_j$, though active $\mathbf{u}_j$ for $\mathbf{z}_i$ may not be unique. Regardless of the uniqueness, we have $\mathbf{u}_j \in \partial\Phi(\mathbf{z}_i)$, which implies

$$\left\{ \mathbf{u}_j \;\middle|\; P^*_{ij} > 0 \right\} \subseteq \partial\Phi(\mathbf{z}_i).$$

Lastly, the convex combination $T_{\mathbf{P}^*}(\mathbf{z}_i) = n \sum_j P^*_{ij} \mathbf{u}_j$ must lie in $\partial\Phi(\mathbf{z}_i)$ because $\partial\Phi(\mathbf{z}_i)$ is a closed convex set. Therefore, $T_{\mathbf{P}^*}(\mathbf{z}_i) \in \partial\Phi(\mathbf{z}_i)$ for every $i \in [n]$, which is the desired argument. □

A.3 Proof of Theorem 3

First, we recapitulate the setup. Let μ be an absolutely continuous probability measure on $\mathbb{R}^d$ and $\nu = \sum_{j=1}^{k} b_j \delta_{\mathbf{u}_j}$ for $\mathbf{b} \in \triangle^{k-1}$, and $\mathbf{g}^* \in \mathbb{R}^k$ be the optimal potential to the semi-discrete OT problem:

$$\sup_{\mathbf{g} \in \mathbb{R}^k} \left\{ \int_{\mathbb{R}^d} g^c \, \mathrm{d}\mu + \sum_{j=1}^{k} b_j g_j \;\middle|\; g^c(\mathbf{z}) + g_j \le \|\mathbf{z} - \mathbf{u}_j\|_2^2 \;\text{ for any } \mathbf{z} \in \mathbb{R}^d \text{ and } j \in [k] \right\},$$

where g^c is the c-transform of $g : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ associated with the squared L_2 cost:

$$g^c(\mathbf{z}) \coloneqq \min_{1 \le j \le k} \left\{ \|\mathbf{z} - \mathbf{u}_j\|_2^2 - g_j \right\} \quad \text{for } \mathbf{z} \in \mathbb{R}^d.$$

The Laguerre map $T_{\mathbf{g}^*} : \mathbb{R}^d \to \mathbb{R}^d$ is given by

$$\begin{aligned} T_{\mathbf{g}^*}(\mathbf{z}) &\in \left\{ \mathbf{u}_j \;\middle|\; \|\mathbf{z} - \mathbf{u}_j\|_2^2 - g^*_j \le \|\mathbf{z} - \mathbf{u}_{j'}\|_2^2 - g^*_{j'} \;\text{ for } j' \ne j \right\} \\ &= \left\{ \mathbf{u}_j \;\middle|\; \langle \mathbf{z}, \mathbf{u}_j \rangle - \frac{1}{2}\|\mathbf{u}_j\|_2^2 + \frac{1}{2} g^*_j \ge \langle \mathbf{z}, \mathbf{u}_{j'} \rangle - \frac{1}{2}\|\mathbf{u}_{j'}\|_2^2 + \frac{1}{2} g^*_{j'} \;\text{ for } j' \ne j \right\}, \end{aligned}$$

[3] Note that Rockafellar (1970, Theorem 24.8) establishes the equivalence between the existence of a closed proper convex function and cyclic monotonicity of (any subset of) its subgradient. Precisely speaking, this requires that the subgradient subset is defined over the entire $\mathbb{R}^d$, but the barycentric map $T_{\mathbf{P}^*}$ for discrete OT is only defined over $\{\mathbf{z}_1, \dots, \mathbf{z}_n\}$. Nevertheless, this does not cause any issue because we only use the necessity of Rockafellar (1970, Theorem 24.8), where we do not need to take care of whether $T_{\mathbf{P}^*}$ is defined over the entire $\mathbb{R}^d$ or not.

where the tie can broken arbitrarily as long as $T_{\mathbf{g}^*}$ is measurable.

Construct a closed proper convex function

$$\Phi(\mathbf{z}) \coloneqq \max_{1\le j\le k}\left\{\langle \mathbf{z}, \mathbf{u}_j\rangle - \frac{1}{2}\|\mathbf{u}_j\|_2^2 + g_j^*\right\}.$$

By following the same strategy as the proof of Theorem 2, we can show that $T_{\mathbf{g}^*}(\mathbf{z}) \in \partial\Phi(\mathbf{z})$, which implies that $T_{\mathbf{g}^*}$ has a cyclically monotone graph. □

B MONOTONICITY NOTIONS

In this section, we briefly discuss a different notion of monotonicity, called the *intra-order preserving* (IOP) property. This property is recently used by Rahimi et al. (2020) for recalibration. Specifically, they learn a neural network with the IOP property for recalibration because it is expected to be a reasonable property as a recalibrator but does not loses the expressivity too much. Here, we discuss that cyclic monotonicity induces a slightly weaker notion of the IOP property.

Definition 3. *A function $f : \mathbb{R}^d \to \mathbb{R}^d$ is said to be* weakly intra-order preserving *if for every $\mathbf{x} \in \mathbb{R}^d$ and every pair $(i, j) \in [d]^2$, we have $x_i \le x_j \implies f_i(\mathbf{x}) \le f_j(\mathbf{x})$. Equivalently, f is weakly intra-order preserving if we have $(x_i - x_j)(f_i(\mathbf{x}) - f_j(\mathbf{x})) \ge 0$ for any $\mathbf{x} \in \mathbb{R}^d$ and $(i, j) \in [d]^2$.*

Definition 4. *A function $f : \mathbb{R}^d \to \mathbb{R}^d$ is said to be* permutation equivariant *if we have $f(\mathbf{P}\mathbf{x}) = \mathbf{P}f(\mathbf{x})$ for every permutation matrix $\mathbf{P}$ and every $\mathbf{x} \in \mathbb{R}^d$.*

Proposition 5. *Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be permutation equivariant and cyclically monotone. Then f is weakly intra-order preserving. In particular, if $x_i = x_j$ then $f_i(\mathbf{x}) = f_j(\mathbf{x})$.*

Proof. Since f is cyclically monotone, Proposition 2 yields that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\langle f(\mathbf{x}) - f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \ge 0.$$

For a fixed $\mathbf{x} \in \mathbb{R}^d$ and indices $i, j \in [n]$ such that $i \ne j$. Let $\mathbf{P} \in \mathbb{R}^{d\times d}$ be the following permutation matrix:

$$P_{i_0 j_0} = \begin{cases} 1 & \text{if } (i_0, j_0) \in \{(i, j), (j, i)\} \cup \{(i_0, i_0) \mid i_0 \ne i, j\}, \\ 0 & \text{otherwise,} \end{cases}$$

which swaps the i-th and j-th elements. If we set $\mathbf{x}' \coloneqq \mathbf{P}\mathbf{x}$, then we have

$$0 \le \langle f(\mathbf{x}) - f(\mathbf{x}'), \mathbf{x} - \mathbf{x}'\rangle = \langle f(\mathbf{x}) - \mathbf{P}f(\mathbf{x}), \mathbf{x} - \mathbf{P}\mathbf{x}\rangle = 2(f_i(\mathbf{x}) - f_j(\mathbf{x}))(x_i - x_j),$$

which is the weakly IOP property. If we have $x_i = x_j$ additionally, then $\mathbf{x}' = \mathbf{x}$ and $f(\mathbf{x}) = \mathbf{P}f(\mathbf{x})$, which implies $f_i(\mathbf{x}) = f_j(\mathbf{x})$. □

C DETAILED EXPERIMENTS

C.1 Full setup

Datasets. The full list of datasets we used in the experiments is shown in Table 3. All datasets are available on OpenML (Bischl et al., 2025). For each dataset, we created a random train-test split with the ratio 8 to 2 first. For calibration experiments (Section 4), we applied 3-fold stratified cross validation further to the train split, split into train and calibration splits. The base models were trained with the train splits, while the recalibrators were trained with calibration splits and their hyperparameters (if any) were chosen based on the train splits with the validation log loss. Then, each recalibrator trained with an individual calibration split were averaged upon recalibration, forming model averaging. This practice is common in literature (Kull et al., 2019).

Table 3: Dataset details.

Dataset	Sample size	# of features	# of classes
balance-scale	625	4	3
car	1728	6	4
cleveland	297	13	5
dermatology	358	34	6
glass	214	9	6
segment	2310	19	7
vehicle	846	18	4
yeast	1484	8	10

Recalibrators. We describe the implementation of each baseline recalibrator (used in Section 4). Except for Bin and IR, all the other baselines are multiclass recalibrators.

- Bin (OvR binning) (Zadrozny and Elkan, 2001): We used the implementation by Kull et al. (2019).[4] The number of bins of the base OvR binning estimator was fixed to 15, and each bin was created uniformly over $[0, 1]$. Then each base estimator was aggregated in the OvR fashion, turned into a multiclass recalibrator.

- IR (OvR isotonic regression) (Zadrozny and Elkan, 2002): We used `scikit-learn` implementation of binary isotonic regression (Pedregosa et al., 2011). Each binary isotonic regressor was aggregated in the OvR fashion, turned into a multiclass recalibrator.

- MS (matrix scaling) (Guo et al., 2017): Let $\mathbf{z}_i \in \mathbb{R}^d$ be a vector of multinomial logits. Then matrix scaling estimates the confidence of class i by $\widehat{p}_i = \max_{c\in[d]} \mathrm{softmax}_c(\mathbf{W}\mathbf{z}_i + \mathbf{b})$, where $\mathrm{softmax}_c$ is the c-th coordinate of the softmax output, and $\mathbf{W} \in \mathbb{R}^{d\times d}$ and $\mathbf{b} \in \mathbb{R}^d$ are optimized with respect to the NLL. Later, Kull et al. (2019) proposed to add the L_2 regularization to make $\mathbf{W}$ stay close to the identity matrix $\mathbf{I}$. We used the implementation by Kull et al. (2019), which optimizes the NLL by L-BFGS. The regularization hyperparameter was chosen from $\{10^2, 10^0, 10^{-2}, 10^{-4}\}$ by the aforementioned cross-validation on the calibration splits.

- TS (temperature scaling) (Guo et al., 2017): It can be seen as a more restricted version of MS, such as $\widehat{p}_i = \max_{c\in[d]} \mathrm{softmax}_c(\mathbf{z}_i/\tau)$, where $\tau > 0$ is the temperature parameter optimized with respect to the NLL. We used the implementation by Kull et al. (2019).

- Dir (Dirichlet calibration) (Kull et al., 2019): It models the calibration map $\eta(\mathbf{q}) = [\Pr(Y = i \mid \widehat{\mathbf{p}}(X) = \mathbf{q})]_{i\in[d]}$ by the Dirichlet distribution. The linear parametrization of Dirichlet recalibrator is $\eta(\mathbf{q}) = \mathrm{softmax}(\mathbf{W} \ln \mathbf{q} + \mathbf{b})$, where $\mathbf{q} \in \triangle^{d-1}$ is the input of a calibration map, or can be seen as a probabilistic output of the base model. The parameters $\mathbf{W} \in \mathbb{R}^{d\times d}$ and $\mathbf{b} \in \mathbb{R}^d$ are optimized with respect to the NLL. This is very similar to MS, with the only difference replacing the logit input $\mathbf{z}$ with $\ln \mathbf{q}$. We used the implementation by Kull et al. (2019), which optimizes the NLL by L-BFGS. In addition, we used the same L_2 regularization as MS, with the regularization hyperparameter chosen from $\{10^2, 10^0, 10^{-2}, 10^{-4}\}$.

- OI (order-invariant network) (Rahimi et al., 2020): To boost the expressivity of recalibrators with minimal inductive bias, order-invariant networks were proposed. That is, a recalibrator must be equivariant to a permutation applied to the input logit vector. This can be implemented by $\mathbf{f}(\mathbf{z}) = S(\mathbf{z})^{-1}\mathbf{U}\mathbf{w}(\mathbf{S}(\mathbf{z})\mathbf{z})$, where $S(\mathbf{z}) \in \mathbb{R}^{d\times d}$ the permutation matrix yielding the descending order of $\mathbf{z} \in \mathbb{R}^d$, $\mathbf{U} \in \mathbb{R}^{d\times d}$ is an upper-triangular matrix of ones, and $\mathbf{w} : \mathbb{R}^d \to \mathbb{R}^d$ is subject to some constraints relevant to the relative order of $\mathbf{z}$-components (see the original paper for the detail). The function $\mathbf{w}$ has a small degree-of-freedoms, and the original paper proposed to model it by a two-layer multilayer perceptron. Based on the official implementation,[5] we reimplemented the whole recalibrator by optimizing the NLL with L-BFGS. Upon the recalibrator training, we applied the L_2 regularization to all perceptron parameters, and the regularization hyperparameter was chosen from $\{10^2, 10^0, 10^{-2}, 10^{-4}\}$.

[4] https://github.com/dirichletcal/experiments_neurips
[5] https://github.com/AmirooR/IntraOrderPreservingCalibration/

Table 4: Averaged running times (in seconds) over 10 trials.

Dataset	# of classes	Sample size	BrenierIR (Ours)			IRP
			$k = 15$	$k = 30$	$k = 50$	
balance-scale	3	625	3.60	13.69	41.19	0.29
car	4	1728	16.84	50.91	221.91	3.20
cleveland	5	297	4.68	16.52	35.05	4.26
dermatology	6	358	7.17	25.82	57.81	48.52
glass	6	214	4.51	18.26	55.76	7.00
segment	7	2310	39.02	119.03	279.27	124.82
vehicle	4	846	7.18	18.39	49.37	1.48
yeast	10	1484	27.72	104.01	256.01	3243.40

- IRP (iterative recursive splitting) (Berta et al., 2024): This is a direct extension of Bin to the probability simplex with higher dimensions. The basic procedure is to recursively find a point to split the probability simplex, and eventually creating bins over the simplex. After terminating binning, the recalibrator is defined as the mean prediction over each bin. Specifically, by creating grid points over the entire probability simplex, the algorithm sweeps the grid points to find out the point where the entropy criterion is maximized after the resulting simplex splitting. Based on the official implementation,[6] we reimplemented the algorithm by fixing the grid resolution (along each dimension) to $k_0 = 10$. Although this is not satisfactory, the entire number of grid points becomes $O(k_0^d)$ with d classes, which is computationally intractable with too large k_0.

Single-index models. We describe the implementation of each baseline (multinomial) single-index models (used in Section 5). Note that we intentionally focus on multiclass classifiers based on multinomial GLMs among numerous candidate baselines. All of the baselines described here model a multinomial inverse link function (more or less).

- Log (multinomial logistic regression): We used `scikit-learn` (Pedregosa et al., 2011) to implement with the default hyperparameters. The loss function is optimized by L-BFGS. The multinomial link function is fixed to the softmax function.

- CLS (calibrated least squares) (Agarwal et al., 2014): This is an alternative algorithm to fit a linear predictor to the residual and calibrate the linear predictor to the observed outcomes by a cyclically monotone link. The monotone link is modeled by an affine combination of bases. As suggested by the original paper, we used $\{z, z^2, z^3\}$ for the bases. Both steps can be executed analytically by solving least squares. We alternated these two steps for 2000 steps, with the L_2 regularization for the latter step (the regularization strength was set to 10^{-6}).

- LT (LegendreTron) (Lam et al., 2023): This models the multinomial inverse link function via input-convex neural networks (ICNNs) (Amos et al., 2017), which is applied on top of a linear predictor. We can immediately obtain a monotone link by taking the gradient of ICNNs with respect to inputs. The original paper suggests ICNN with 2 blocks, 4 hidden units, and 4 layers. We used the official implementation,[7] without changing the default hyperparameter setups, and trained for 240 epochs.

C.2 Computational time

We additionally measure the computational time to compare BIR and IRP, the two most promising approaches in probability calibration. We report the computational time of fitting recalibrators and test prediction, averaged over 10 trials. As seen in Table 4, the running time of BIR becomes more advantageous with a mildly large number of classes ($\sim$10).

[6] https://github.com/eugeneberta/Calibration-ROC-IR/
[7] https://github.com/khflam/legendretron/

C.3 More illustration of calibration maps

We show more illustrations of the calibration maps, complementing Fig. 3 in Section 4. In Fig. 5, we show the calibration maps of nonparametric recalibrators: BIR (Brenier isotonic regression), Bin (OvR binning), IRP (iterative recursive splitting), and IR (OvR isotonic regression). The OvR approaches Bin and IR do not learn inter-class information, and hence the resulting calibration maps exhibit contours parallel to the simplex boundary. Moreover, these two methods do not significantly perform binning so that the counter maps have considerably many bins. This is because the number of bins k grows exponentially in the number of classes d, such as $k = O(k_0^d)$, suppose k_0 is the number of bins of a single base recalibration model of the whole OvR model. This can cause overfitting, leading to inferior calibration performance, as seen in Table 1. To the contrary, BIR and IRP produce coarser and adaptive binning results. One notable difference is that IRP is restricted to the simplex splitting perpendicular to the simplex boundary. Indeed, each boundary of the calibration map levels is perpendicular to the simplex boundary.[8] This is attributed to their simplex splitting criterion R_k in Berta et al. (2024, Eq. (3)), and can also be seen in their Figure 3. Our BIR does not have such a restriction because the simplex splitting is based on the Laguerre cells, which is fully adaptive to the Kantorovich potential. As we increase the number of bins k for BIR, the simplex binning results become finer, but not excessively detailed unlike Bin and IR. Moreover, BIR has outputs clearly concentrated on a few number of points over the simplex, as can bee seen in the calibration map visualization, which nicely trades off the bias and variance in practice.

In Fig. 6, we show the calibration maps of parametric recalibrators: Dir (Dirichlet calibration), MS (matrix scaling), OI (order-invariant network), and TS (temperature scaling). By comparing the calibration maps, the vector fields produced by these parametric recalibrators entail the non-degenerating behaviors, unlike the nonparametric recalibrators. In the right three contour plots, the level-set boundaries are more smooth than those of the nonparametric recalibrators, which results in worse expressivity.

Further, we tested the same setup but with a different based model, naive Bayes, instead of MLP. The obtained calibration maps for all recalibrators are shown in Fig. 7 (for nonparametric recalibrators) and Fig. 8 (for parametric recalibrators). The overall trends resemble the MLP case.

C.4 More calibration results

In addition to the calibration results shown in Section 4, which only measures the L_1 calibration error, we expand in this section the benchmark results for the other metrics. Specifically, Table 5 shows accuracy, Table 6 shows the classwise calibration error, and Table 7 shows the confidence calibration error.

We observe the following points.

1. In terms of the classwise/confidence/L_1 calibration error, IRP is quite a powerful baseline, and our BrenierIR performs on par with it.
2. In terms of accuracy, IRP is no longer strong enough but BrenierIR achieves the reasonable standard.

[8] Although the theoretical binning boundaries given by IRP are completely straight and perpendicular to the simplex boundaries, the computed boundaries in Fig. 5 look a bit "jaggy" due to the insufficient resolution when we run IRP algorithm. For this visualization, we used 200 grids to run IRP to get as close boundaries to theory as possible.

Table 5: Recalibration results for MLP **(upper table)** and linear SVM **(lower table)**. Each number indicates **accuracy** (higher is better) with averaging 10 trials, and bold-faced if the recalibrator achieves the best or second best performance or statistically indistinguishable from them by the Mann–Whitney U test (significance level: 5%).

Dataset \ Recalibrator	—	Bin	Dir	IRP	IR	MS	OI	TS	BrenierIR (Ours)		
									$k = 15$	$k = 30$	$k = 50$
balance-scale	0.928	**0.968**	**0.968**	0.960	**0.968**	0.928	0.928	0.928	**0.965**	0.960	0.959
car	**0.991**	**0.988**	**0.988**	0.986	0.986	**0.991**	**0.991**	**0.991**	0.965	**0.988**	0.988
cleveland	0.525	0.492	**0.607**	0.541	0.557	**0.590**	0.525	0.525	0.549	0.528	0.536
dermatology	0.919	0.919	**0.932**	0.919	**0.932**	0.919	0.919	0.919	0.916	**0.934**	0.923
glass	0.698	0.674	0.698	0.698	**0.791**	0.698	0.698	0.698	**0.744**	0.730	0.702
vehicle	0.829	0.824	0.829	0.812	**0.835**	0.824	0.829	0.829	**0.844**	0.809	0.829

Dataset \ Recalibrator	—	Bin	Dir	IRP	IR	MS	OI	TS	BrenierIR (Ours)		
									$k = 15$	$k = 30$	$k = 50$
balance-scale	0.872	**0.880**	0.736	0.456	0.872	0.736	0.736	0.736	**0.928**	**0.928**	**0.928**
car	0.882	**0.893**	0.737	0.882	**0.908**	0.754	0.260	0.260	0.860	0.881	0.879
cleveland	**0.607**	0.541	0.574	0.590	0.557	0.590	0.557	0.557	0.595	**0.597**	0.582
dermatology	**0.959**	**0.959**	0.568	0.838	**0.959**	0.568	0.311	0.311	0.915	0.934	**0.950**
glass	0.605	0.581	0.419	0.349	**0.651**	0.395	0.326	0.326	**0.577**	**0.658**	**0.644**
vehicle	**0.765**	**0.759**	0.576	0.259	**0.765**	0.588	0.412	0.412	0.749	0.733	0.747

Table 6: Recalibration results for MLP **(upper table)** and linear SVM **(lower table)**. Each number indicates the **classwise calibration error** (lower is better) with averaging 10 trials, and bold-faced if the recalibrator achieves the best or second best performance or statistically indistinguishable from them by the Mann–Whitney U test (significance level: 5%).

Dataset \ Recalibrator	—	Bin	Dir	IRP	IR	MS	OI	TS	BrenierIR (Ours)		
									$k = 15$	$k = 30$	$k = 50$
balance-scale	0.059	0.050	0.029	**0.020**	0.038	0.045	0.046	0.039	**0.020**	**0.020**	0.024
car	0.015	0.011	**0.007**	**0.007**	0.008	0.008	0.028	0.008	0.011	0.010	0.009
cleveland	0.114	0.089	0.081	**0.035**	**0.063**	0.094	0.184	0.097	0.072	0.083	0.087
dermatology	0.028	0.030	0.024	0.029	**0.023**	0.027	0.069	0.027	**0.019**	0.026	0.027
glass	0.097	0.087	0.089	**0.072**	0.082	0.087	0.132	0.098	0.081	0.085	**0.071**
vehicle	0.058	0.052	0.040	**0.021**	0.040	0.057	0.116	0.055	0.044	**0.027**	0.044

Dataset \ Recalibrator	—	Bin	Dir	IRP	IR	MS	OI	TS	BrenierIR (Ours)		
									$k = 15$	$k = 30$	$k = 50$
balance-scale	0.032	0.061	0.094	**0.004**	0.089	0.053	0.179	0.053	**0.029**	0.029	0.034
car	**0.024**	0.065	0.054	0.139	0.052	0.052	0.177	0.133	**0.028**	0.041	0.035
cleveland	**0.074**	0.076	0.083	**0.037**	0.089	0.106	0.139	0.154	0.082	0.097	0.093
dermatology	0.040	0.041	0.029	0.051	**0.013**	0.024	0.063	0.098	0.021	0.020	**0.020**
glass	0.089	0.073	0.105	**0.007**	**0.067**	0.086	0.096	0.083	0.074	0.089	0.100
vehicle	0.074	0.082	0.104	**0.002**	0.066	0.084	0.098	0.087	**0.061**	0.072	0.070

Table 7: Recalibration results for MLP **(upper table)** and linear SVM **(lower table)**. Each number indicates the **confidence calibration error** (lower is better) with averaging 10 trials, and bold-faced if the recalibrator achieves the best or second best performance or statistically indistinguishable from them by the Mann–Whitney U test (significance level: 5%).

Dataset \ Recalibrator	—	Bin	Dir	IRP	IR	MS	OI	TS	BrenierIR (Ours)		
									$k = 15$	$k = 30$	$k = 50$
balance-scale	0.062	0.062	0.023	**0.010**	0.032	0.035	0.049	0.044	**0.012**	0.020	0.023
car	0.024	0.016	0.011	0.013	**0.011**	**0.011**	0.063	0.015	0.019	0.017	0.016
cleveland	0.172	0.201	0.106	**0.061**	0.158	**0.099**	0.435	0.126	0.119	0.168	0.161
dermatology	0.058	**0.050**	0.066	0.083	0.058	0.067	0.418	0.058	**0.024**	0.058	0.057
glass	0.155	0.155	0.218	0.197	0.228	0.169	0.517	**0.136**	0.152	**0.159**	**0.144**
vehicle	0.064	0.064	0.063	**0.040**	0.057	**0.050**	0.214	0.051	0.078	0.055	0.062

Dataset \ Recalibrator	—	Bin	Dir	IRP	IR	MS	OI	TS	BrenierIR (Ours)		
									$k = 15$	$k = 30$	$k = 50$
balance-scale	**0.030**	0.075	0.141	**0.006**	0.098	0.051	0.221	0.048	0.042	0.038	0.038
car	**0.026**	0.101	0.088	0.277	0.086	0.059	0.092	0.218	**0.046**	0.062	0.049
cleveland	0.148	0.183	0.209	**0.030**	0.158	0.192	0.300	0.276	0.153	**0.147**	0.160
dermatology	0.087	0.083	0.224	0.100	**0.033**	0.064	0.102	0.061	0.054	0.043	**0.041**
glass	0.182	0.156	0.102	**0.008**	0.150	0.150	0.120	**0.076**	0.217	0.225	0.220
vehicle	0.106	0.099	0.138	**0.001**	0.088	0.084	**0.058**	0.101	0.095	0.100	0.115

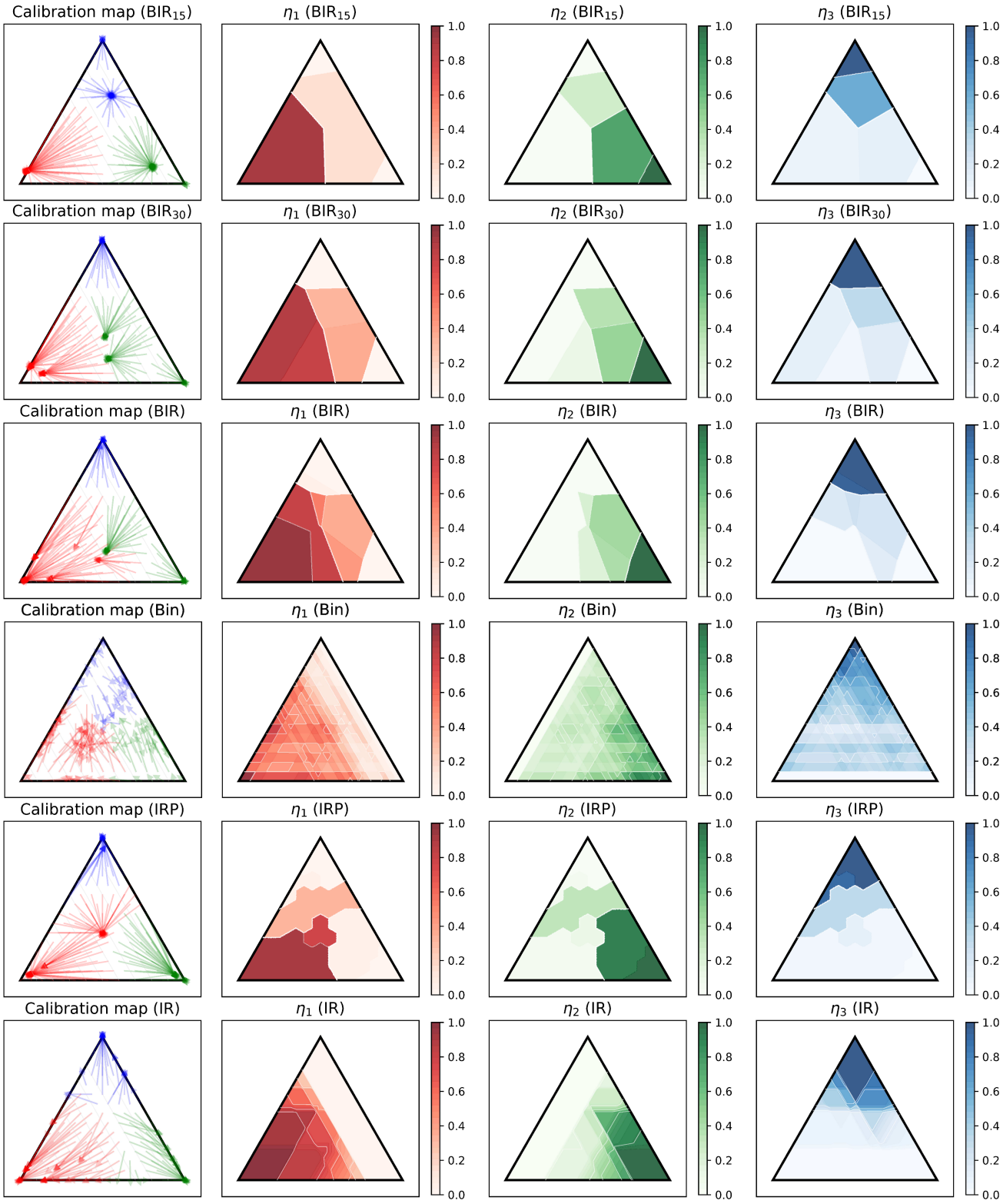

Figure 5: Calibration maps of nonparametric recalibrators. The base model is MLP and balance-scale dataset is used. In each row (corresponding to one recalibrator), we show the calibration map (vector field), the first/second/third coordinates of the calibration map, respectively, from left to right.

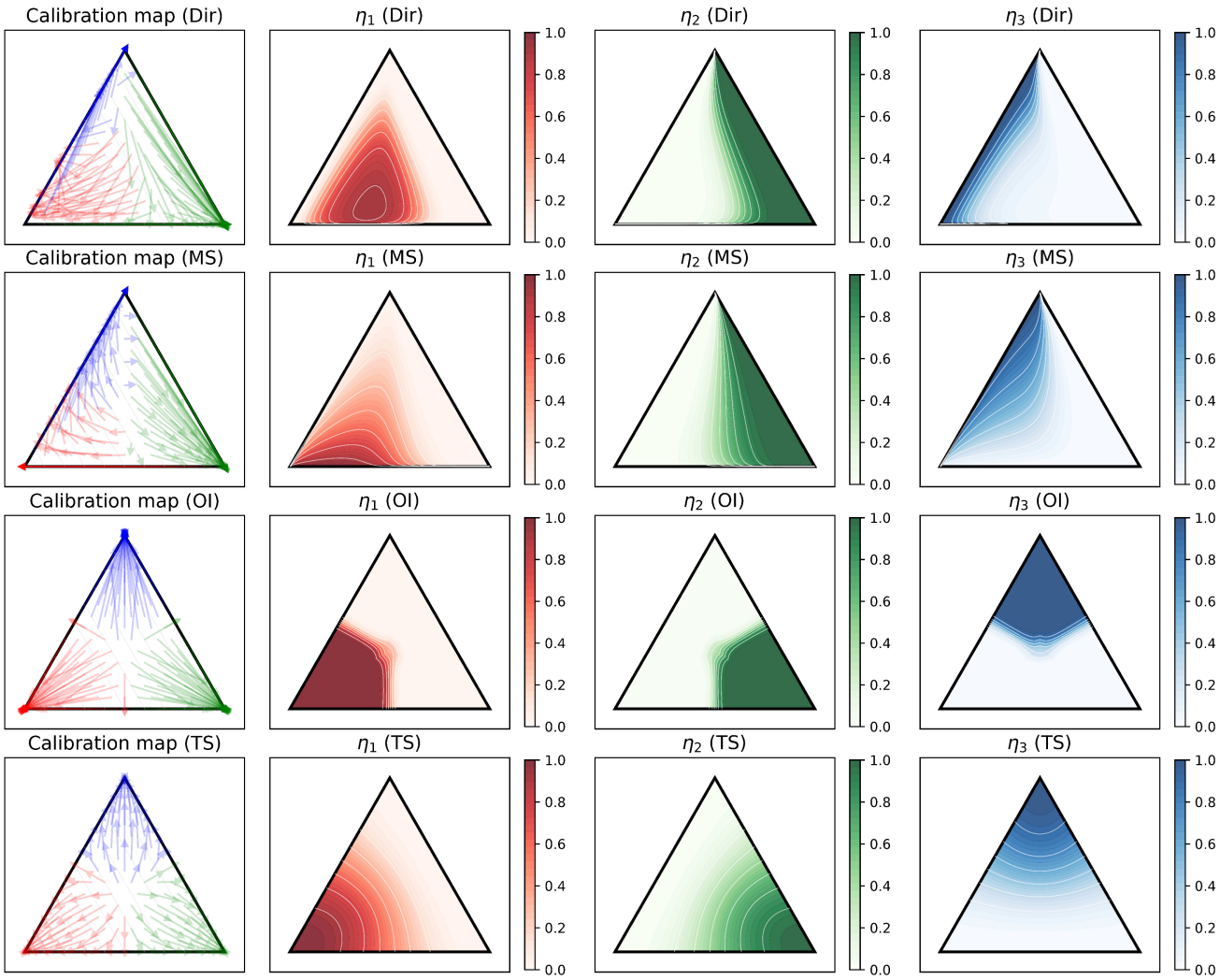

Figure 6: Calibration maps of parametric recalibrators. The base model is MLP and balance-scale dataset is used. In each row (corresponding to one recalibrator), we show the calibration map (vector field), the first/second/third coordinates of the calibration map, respectively, from left to right.

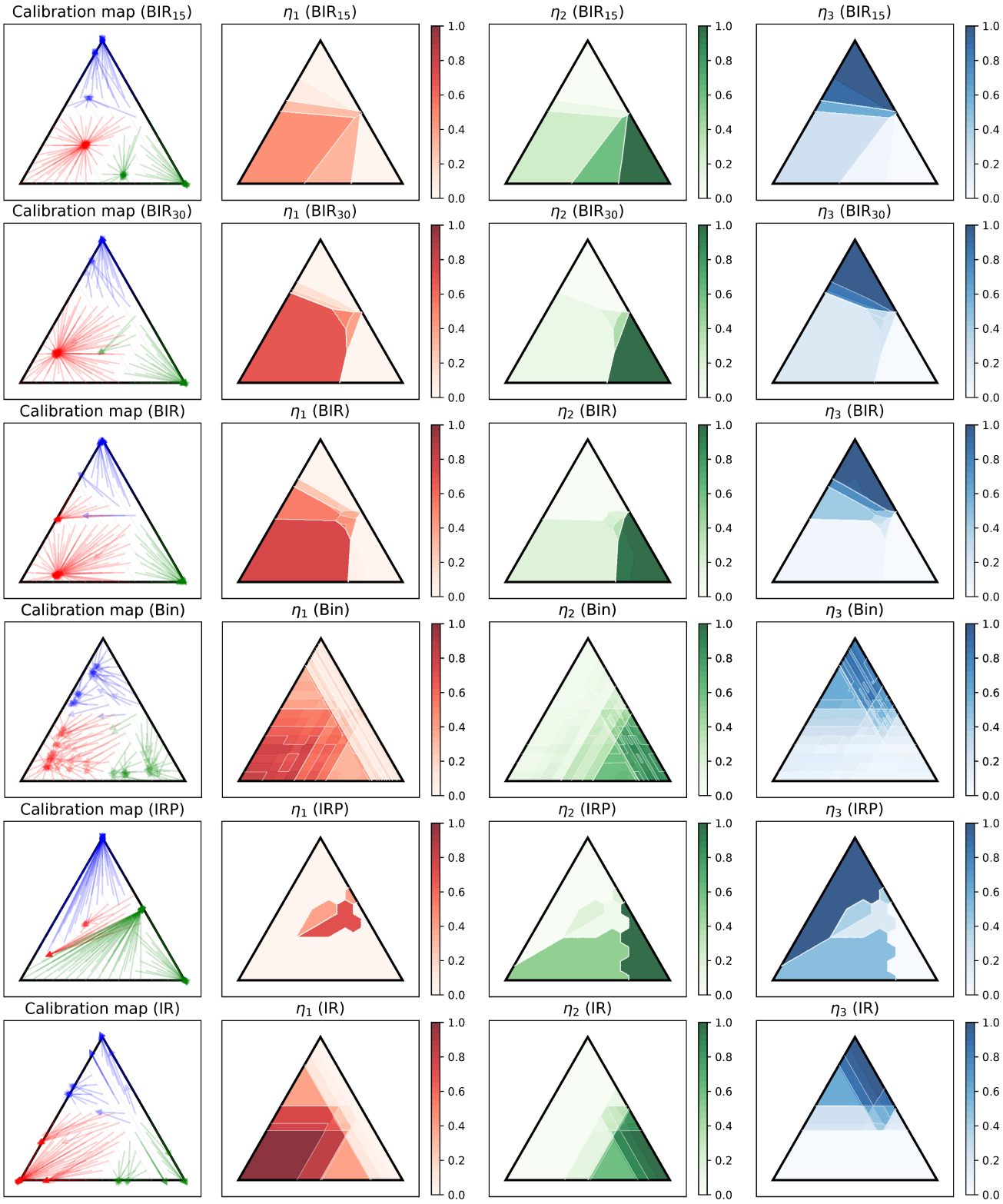

Figure 7: Calibration maps of nonparametric recalibrators. The base model is naive Bayes and balance-scale dataset is used. In each row (corresponding to one recalibrator), we show the calibration map (vector field), the first/second/third coordinates of the calibration map, respectively, from left to right.

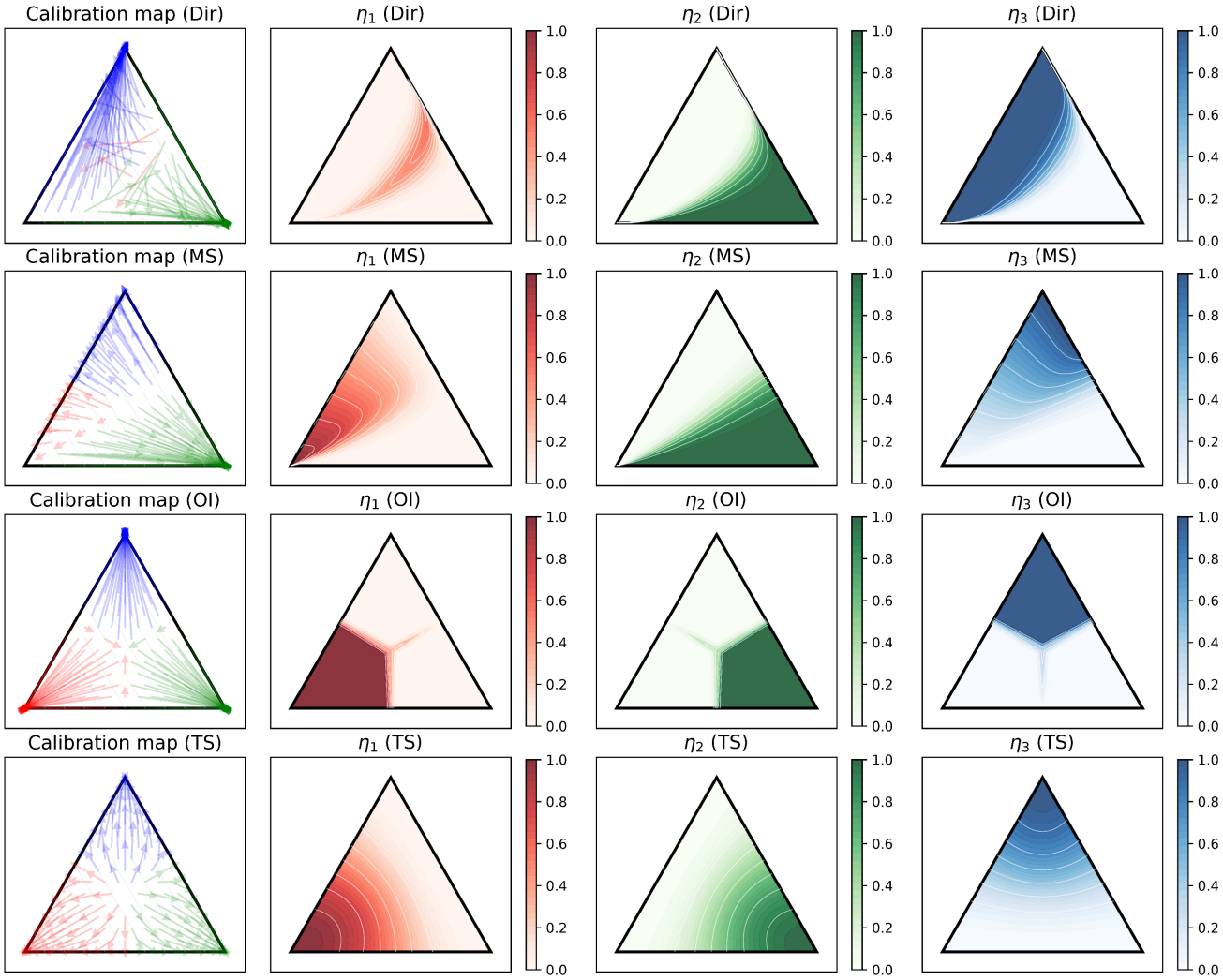

Figure 8: Calibration maps of parametric recalibrators. The base model is naive Bayes and balance-scale dataset is used. In each row (corresponding to one recalibrator), we show the calibration map (vector field), the first/second/third coordinates of the calibration map, respectively, from left to right.