Taming the Tails: Why Distributionally Robust Optimization Needs New Theory for Imbalanced Regression

Nathaniel Kang*
School of Computer Science and Engineering
Kyungpook National University
Daegu, Republic of Korea
natekang@knu.ac.kr

Abstract

Predictive models on structured tabular data often face imbalanced targets: training emphasizes common values while rare outcomes stay underrepresented. Distributionally robust optimization hedges against shift, but a single ambiguity radius spends adversarial mass on dense regions, so rare target ranges stay weakly protected. We introduce target-conditional DRO (TC-DRO), which gives each target region its own radius, larger where data are scarce. Per-region excess-risk bounds tighten with the local sample size when a pooled radius yields a vacuous tail guarantee, and duality returns a reweighted empirical risk. On tabular benchmarks a uniform Wasserstein budget collapses to empirical risk minimization and a global chi-squared penalty stays near it, while target-conditional budgets improve tail-sensitive error.

1 Introduction

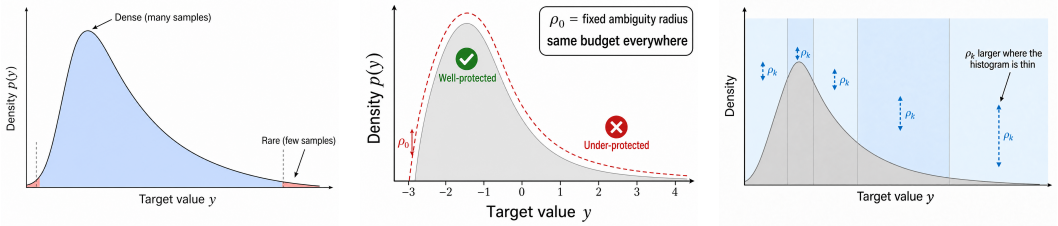
Deployed models on continuous targets often face a recurring tension: most training mass sits on a handful of outcome ranges, while other ranges—frequently the ones that drive downstream decisions—appear only rarely. Training that optimizes average empirical error can therefore look strong in aggregate while remaining brittle on those underrepresented parts of the support; this is the regime known as imbalanced regression.

Practitioners address distributional shift and worst-case risk through distributionally robust optimization (DRO), which hedges against perturbations of the data-generating law and supports performance certificates [Ben-Tal et al., 2009, Rahimian and Mehrotra, 2022, Duchi and Namkoong, 2021, Mohajerin Esfahani and Kuhn, 2018]. The usual construction assigns one ambiguity budget that applies uniformly across the entire problem [Namkoong and Duchi, 2016, Duchi and Namkoong, 2021]. That choice is natural when uncertainty is spread across comparable regions, but it misaligns with heavy imbalance: adversarial slack is allocated in proportion to empirical mass, so dense parts of the target consume most of the budget while rare regions—where statistical uncertainty is actually largest—receive comparatively little. Robustness statements then track typical behavior, and tails or other low-density ranges stay only weakly protected.

Figure 1 contrasts the two geometries on a skewed target law: panel (b) keeps a uniform ambiguity band, whereas panel (c) keeps the target histogram fixed and widens the per-bin radius where that histogram is thin. Section 3 formalizes the mismatch for standard chi-squared and Wasserstein ambiguity and explains why the induced robustness penalty can follow bulk variation while barely moving on rare regions.

*Code is available at https://github.com/nathanielkang/TC-DRO.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Imbalanced target distribution. (b) Standard DRO: uniform set. (c) TC-DRO: per-bin radius ρ_k.

Figure 1: Standard DRO versus Target-Conditional DRO on an imbalanced regression task. (a) The target distribution is heavily right-skewed with sparse tails. (b) Standard DRO applies a uniform ambiguity radius ρ regardless of local density, leaving tail regions under-protected; the dashed band is that radius, not a second density. (c) The gray curve is the same fixed target histogram as in (a). Each arrow is the per-bin radius ρ_k, longer where that histogram is thin, and the arrows are not a second density. The adversary moves feature mass only inside each target bin, so the target marginal stays fixed.

We focus on *tabular* imbalanced regression because it is the workhorse of operational analytics: finance, engineering, medicine, and operations research routinely combine heterogeneous numeric and categorical features with a continuous outcome, and skewed targets are the rule rather than the exception. The target is ordered and low-dimensional, which makes it natural to stratify the outcome range, reason about local effective sample size, and tie robustness budgets to those counts. That same structure guides both the ambiguity sets we propose and the benchmark suite we run.

The central proposal is *target-conditional* ambiguity: assign separate uncertainty budgets across target regions so that budgets grow where data are thin, aligning robustness with local statistical uncertainty instead of global mass alone. Section 4 introduces the ambiguity sets, derives the dual, and shows how training becomes a reweighted empirical risk whose weights are dual variables of the ambiguity set rather than exogenous rarity scores. Section 5 establishes per-bin deviation and excess-risk bounds that strengthen with local counts when uniform DRO is vacuous on tails, and proves convergence of the training loop. Section 6 evaluates the method on standard tabular benchmarks. In line with the theory, standard DRO collapses to empirical risk minimization under the uniform budgets we study, whereas TC-DRO improves tail-sensitive metrics relative to both ERM and recent imbalance baselines.

2 Related Work

Distributionally robust optimization. The modern DRO framework traces its roots to the robust optimization literature [Ben-Tal et al., 2009, Bertsimas et al., 2018], with divergence-based ambiguity sets forming the most widely studied class. The f-divergence formulation, rooted in the seminal work of Ali and Silvey [1966] and Csiszár [1967], was first analyzed in the context of stochastic optimization by Ben-Tal et al. [2013] and Hu and Hong [2013], encompassing chi-squared [Namkoong and Duchi, 2016, Duchi and Namkoong, 2021], KL [Hu and Hong, 2013], and CVaR [Rockafellar and Uryasev, 2000] variants as special cases. Transport-based ambiguity sets defined via Wasserstein distance [Mohajerin Esfahani and Kuhn, 2018, Blanchet and Murthy, 2019, Gao and Kleywegt, 2023] offer complementary geometric properties. Scalable algorithms for large-scale DRO have been developed by Levy et al. [2020]. Recent work has connected DRO to regularization [Namkoong and Duchi, 2019], group fairness [Hashimoto et al., 2018, Sagawa et al., 2020, Lahoti et al., 2020], and tilted risk minimization [Li et al., 2021]. Statistical guarantees for DRO have been established through empirical likelihood theory [Lam, 2019, Duchi et al., 2021]. All of these formulations, however, apply a uniform ambiguity radius across the full data distribution. The present work identifies this uniformity as a critical limitation for imbalanced regression and proposes target-conditional ambiguity sets to address it.

Imbalanced regression. While imbalanced classification has a long history [He and Garcia, 2009, Cui et al., 2019, Kang et al., 2020, Menon et al., 2021], imbalanced regression has received focused attention only recently. The DIR benchmark introduced by Yang et al. [2021] demonstrated that

classical oversampling strategies fail in the continuous-target setting and proposed Label Distribution Smoothing (LDS) and Feature Distribution Smoothing (FDS) as alternatives. Subsequent methods have explored ranking-based regularization [Gong et al., 2022], balanced loss reformulations [Ren et al., 2022], contrastive objectives [Keramati et al., 2024], and synthetic oversampling [Branco et al., 2017, Torgo et al., 2013]. Density-based reweighting [Steininger et al., 2021] is simple and effective. In these methods the correction to the loss or the data is a rarity score computed from the target values before training, which is not tied to a stated ambiguity set. TC-DRO obtains its weights as dual variables of per-bin ambiguity sets, so they respond to the current per-bin losses and come with per-region performance guarantees; the method still has its own hyperparameters, the exponent α and base radius ρ_0.

Robust and adaptive loss functions. Beyond DRO, a parallel line of work designs loss functions that are inherently robust to outliers or label noise. The adaptive robust loss of Barron [2019] interpolates between squared loss and absolute loss via a shape parameter. Generalized cross entropy [Zhang and Sabuncu, 2018] and symmetric losses [Ghosh et al., 2017] provide noise-tolerance guarantees for classification. The tilted ERM framework of Li et al. [2021] uses an exponential tilting parameter to interpolate between average-case and worst-case performance. These methods operate on the global loss distribution and lack the target-conditional structure needed for imbalanced regression: a loss that is robust to outliers everywhere is not the same as one that provides strong guarantees specifically on rare target regions.

Sample reweighting. Meta-learning approaches to sample reweighting [Ren et al., 2018, Shu et al., 2019] learn to assign importance weights to training examples using a clean validation set. While effective in practice, these methods lack the formal worst-case guarantees of DRO and require careful validation set construction [Byrd and Lipton, 2019]. The fairness literature has explored reweighting for group-level robustness [Williamson and Menon, 2019, Hashimoto et al., 2018], but the "groups" are defined by protected attributes rather than target values. In TC-DRO the weights are density-adaptive importance weights whose functional form follows from the dual; only α and ρ_0 are chosen on validation data, and no meta-learning loop is needed.

3 Problem Formulation

Notation and setup. Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the feature space and $\mathcal{Y} \subseteq \mathbb{R}$ the target space. Training data $\{(x_i, y_i)\}_{i=1}^n$ are drawn i.i.d. from a joint distribution P over $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ with marginal target density p_Y. A model $f_\theta : \mathcal{X} \to \mathcal{Y}$ with parameters $\theta \in \Theta$ is trained by minimizing a loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$; the population risk is $R(\theta) = \mathbb{E}_{(X,Y)\sim P}[\ell(f_\theta(X), Y)]$ and its empirical counterpart is $\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell_i(\theta)$, where $\ell_i(\theta) := \ell(f_\theta(x_i), y_i)$ (abbreviated ℓ_i when θ is clear).

Distributional objects. The empirical joint measure $\hat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$, with atoms $z_i = (x_i, y_i)$, induces a Y-marginal $\hat{P}_{n,Y}$ and per-bin conditionals $\hat{P}_{n,X|Y\in B_k}$ for each of K target bins $B_1, \ldots, B_K$ (we write $[K] := \{1, \ldots, K\}$). Each bin contains n_k samples, and $\hat{P}_{n_k}$ denotes their uniform measure. In the TC-DRO decomposition, Q_k denotes a candidate distribution on $\mathcal{Z}$ supported on $\{(x, y) : y \in B_k\}$ that the per-bin adversary may choose.

Definition 1 (Target region and conditional risk)**.** *A **target region** is a measurable subset $S \subseteq \mathcal{Y}$ with region weight $p_S = P(Y \in S)$. The conditional risk on S is defined as:*

$$R_S(\theta) = \mathbb{E}[\ell(f_\theta(X), Y) \mid Y \in S]. \tag{1}$$

*A region S is **rare** if $p_S < \eta$ for a threshold $\eta > 0$, and the corresponding local sample size is $n_S = |\{i : y_i \in S\}|$.*

Standard DRO. The standard DRO formulation seeks a model minimizing the worst-case risk over an ambiguity set $\mathcal{U}_\rho(\hat{P}_n)$:

$$\min_{\theta\in\Theta} \sup_{Q\in\mathcal{U}_\rho(\hat{P}_n)} \mathbb{E}_Q[\ell(f_\theta(X), Y)], \tag{2}$$

where $\hat{P}_n$ is the empirical joint measure from the notation paragraph and $\mathcal{U}_\rho(\hat{P}_n) = \{Q : D_\phi(Q\|\hat{P}_n) \le \rho\}$ for an f-divergence D_ϕ with generator ϕ and uniform radius $\rho > 0$ [Namkoong and Duchi, 2016].

Why standard DRO fails for tails. The fundamental limitation of standard DRO for imbalanced regression is captured by the following result.

Proposition 2 (Tail insensitivity of standard DRO)**.** *Consider the χ^2-DRO formulation with $D_\phi(Q\|P) = \frac{1}{2}\mathbb{E}_P[(dQ/dP - 1)^2]$ and radius $\rho > 0$. The worst-case expected loss decomposes as:*

$$\sup_{Q: D_{\chi^2}(Q\|\hat{P}_n) \le \rho} \mathbb{E}_Q[\ell] = \frac{1}{n}\sum_{i=1}^{n} \ell_i + \sqrt{2\rho \cdot \widehat{\mathrm{Var}}(\ell)}, \tag{3}$$

where $\ell_i = \ell_i(\theta)$ under the model being analyzed, $\widehat{\mathrm{Var}}(\ell) = \frac{1}{n}\sum_{i=1}^{n} \ell_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} \ell_i\right)^2$ is the empirical variance of $\{\ell_i\}_{i=1}^n$, and $\bar{\ell} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \ell_i$. For any rare target region S with $n_S = |\{i : y_i \in S\}|$ samples and $p_S \ll 1$, let $\bar{\ell}_S \coloneqq \frac{1}{n_S}\sum_{i:y_i \in S} \ell_i$ and $\bar{\ell}_{S^c}$ denote the empirical conditional loss means on S and S^c respectively, and write $\hat{p}_S \coloneqq n_S/n$ for the empirical region weight. The variance decomposes as:

$$\widehat{\mathrm{Var}}(\ell) = (1-\hat{p}_S)\cdot \widehat{\mathrm{Var}}(\ell \mid Y \notin S) + \hat{p}_S \cdot \widehat{\mathrm{Var}}(\ell \mid Y \in S) + \hat{p}_S(1-\hat{p}_S)(\bar{\ell}_S - \bar{\ell}_{S^c})^2, \tag{4}$$

so the robustness penalty $\sqrt{2\rho \cdot \widehat{\mathrm{Var}}(\ell)}$ is governed by the dense region and provides negligible protection for S.

The proof follows from the closed-form dual of χ^2-DRO [Duchi and Namkoong, 2021] and the law of total variance (see Appendix A.1). Intuitively, the adversary's optimal strategy concentrates its perturbation budget on the dense region where small per-sample changes accumulate over many samples. Rare target values, contributing only $O(p_S)$ to the total variance, receive proportionally negligible adversarial attention.

This result extends beyond χ^2-divergence. For Wasserstein DRO, the worst-case distribution transports mass toward high-loss regions globally [Mohajerin Esfahani and Kuhn, 2018, Gao and Kleywegt, 2023], again ignoring the target-conditional structure of imbalanced data. Appendix E provides the analogous analysis for Wasserstein ambiguity sets.

4 Target-Conditional DRO Framework

4.1 Target-Conditional Ambiguity Sets

Standard DRO's single ambiguity set is replaced by per-bin sets with density-adaptive radii. Partition $\mathcal{Y}$ into K quantile bins $\{B_k\}$ with counts $n_k = |\{i : y_i \in B_k\}|$ and $\sum_k n_k = n$. Let $\bar{d}_k = \frac{1}{n_k}\sum_{i:y_i \in B_k} \hat{p}_Y(y_i)$ denote the average kernel density estimate on bin B_k, where $\hat{p}_Y$ is a Gaussian KDE of the target marginal.

Definition 3 (Adaptive radius function)**.** *The **adaptive radius function** $\rho : [K] \to \mathbb{R}_+$ assigns to each bin B_k a robustness radius:*

$$\rho_k = \rho_0 \cdot \left(\frac{\bar{d}_k}{\bar{d}_{\max}}\right)^{-\alpha}, \tag{5}$$

where $\bar{d}_{\max} = \max_{j\in[K]} \bar{d}_j$, $\rho_0 > 0$ is a base radius, and $\alpha \in (0, 1]$ controls the adaptation strength. Setting $\alpha = 0$ recovers the uniform radius of standard DRO.

Using continuous KDE density rather than the bin count ratio n_k/n is essential under quantile binning, where all bins have approximately equal counts $n_k \approx n/K$ by construction. The KDE density $\bar{d}_k$ captures genuine local scarcity because quantile bins on a skewed distribution span vastly different widths in target space: a tail bin covers a wide range with few samples per unit of $\mathcal{Y}$, yielding low average density, while a dense-center bin covers a narrow range, yielding high density. This distinction is invisible to n_k/n but directly measured by $\bar{d}_k$, ensuring that the adaptive radius meaningfully inflates for rare regions even when bin counts are balanced.

Definition 4 (Target-conditional ambiguity set)**.** *Given partition $\{B_k\}_{k=1}^K$ and adaptive radii $\{\rho_k\}_{k=1}^K$, the **target-conditional (TC) ambiguity set** is:*

$$\mathcal{U}_{\mathrm{TC}}(\hat{P}_n) = \Big\{Q : Q_Y = \hat{P}_{n,Y},\ D_\phi\Big(Q_{X|Y\in B_k} \,\Big\|\, \hat{P}_{n,X|Y\in B_k}\Big) \le \rho_k,\ \forall\, k \in [K]\Big\}, \tag{6}$$

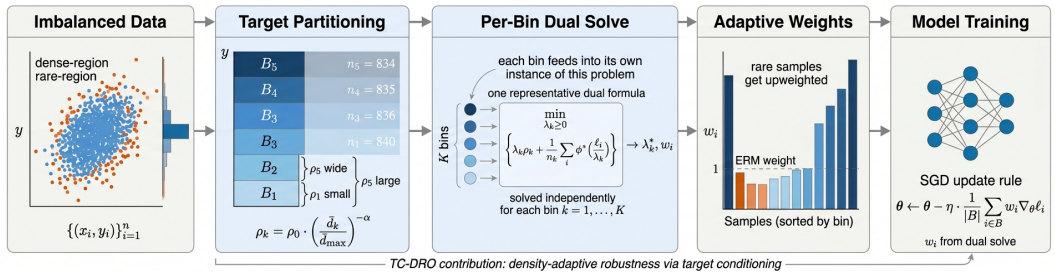


Figure 2: Overview of the TC-DRO framework. The target space is partitioned into bins with density-adaptive robustness radii. The dual reformulation produces per-sample weights that are larger for rare target regions, yielding a principled reweighting scheme for imbalanced regression.

where Q_Y denotes the target marginal of Q, and $Q_{X|Y\in B_k}$ denotes the conditional feature distribution restricted to bin B_k.

The constraint $Q_Y = \hat{P}_{n,Y}$ lets the adversary move mass only inside a bin, across features; the bin masses n_k/n stay fixed, so the target marginal is held at bin resolution. The ambiguity set therefore models conditional covariate shift in $P(X \mid Y)$, not a change of the Y marginal, and this is what enables the factorization (8). Appendix B expands the modeling rationale.

4.2 TC-DRO Objective and Dual Reformulation

The TC-DRO optimization problem is:

$$\min_{\theta\in\Theta} \mathcal{R}_{\mathrm{TC}}(\theta) := \sup_{Q\in\mathcal{U}_{\mathrm{TC}}(\hat{P}_n)} \mathbb{E}_Q[\ell(f_\theta(X), Y)]\,. \tag{7}$$

Because the target marginal is fixed and the conditional constraints decompose by bin, the inner supremum separates into K independent per-bin DRO problems:

$$\mathcal{R}_{\mathrm{TC}}(\theta) = \sum_{k=1}^{K} \frac{n_k}{n} \sup_{Q_k : D_\phi(Q_k \| \hat{P}_{n_k}) \le \rho_k} \mathbb{E}_{Q_k}[\ell(f_\theta(X), Y) \mid Y \in B_k]\,. \tag{8}$$

Assumption 1. *The f-divergence generator $\phi : \mathbb{R}_+ \to \mathbb{R}$ is closed, proper, and convex with $\phi(1) = 0$. Its convex conjugate $\phi^*(s) = \sup_{t\ge 0}\{st - \phi(t)\}$ is finite for all s in the range of ℓ.*

Theorem 5 (Dual reformulation of TC-DRO)**.** *Under Assumption 1, the TC-DRO problem* (7) *is equivalent to:*

$$\min_{\theta\in\Theta} \sum_{k=1}^{K} \frac{n_k}{n} \inf_{\lambda_k \ge 0,\, \nu_k} \left\{ \lambda_k \rho_k + \nu_k + \frac{1}{n_k} \sum_{i : y_i \in B_k} \lambda_k\, \phi^*\!\left(\frac{\ell_i(\theta) - \nu_k}{\lambda_k} \right) \right\}, \tag{9}$$

where $\ell_i(\theta) = \ell(f_\theta(x_i), y_i)$, ϕ^* is the convex conjugate of the divergence generator ϕ, $\lambda_k \ge 0$ is the dual variable for the divergence constraint, and $\nu_k \in \mathbb{R}$ is the dual variable for the probability-simplex constraint $\sum_i q_i = 1$ within bin B_k.

Proof sketch. Each summand in (8) is standard f-divergence DRO on a finite empirical measure; Fenchel duality with Lagrange multipliers for both the divergence budget and the simplex constraint [Ben-Tal et al., 2013, Namkoong and Duchi, 2016] yields (9) under Slater's condition (Appendix A.2). □

Per-bin dual variables (λ_k, ν_k) modulate local robustness; Figure 2 sketches the pipeline from quantile bins to weighted SGD. For the χ^2 specialization, ν_k^* optimizes to the per-bin mean loss $\bar{\ell}_k$, recovering the centered form below. Implicit sample weights from (9) appear in Appendix C.

Chi-squared specialization. For $\phi(t) = \frac{1}{2}(t-1)^2$, solving the bin-wise dual in closed form—including the normalization constraint that centers the loss—gives the TC-χ^2 objective (cf. (3) with a

Algorithm 1 TC-DRO Training

Require: Training data $\{(x_i, y_i)\}_{i=1}^n$, bins K, base radius ρ_0, exponent α, learning rate η
Ensure: Trained parameters θ

Partition $\mathcal{Y}$ into K quantile bins $\{B_k\}$; let $k(i)$ denote the bin index of y_i
Record counts $n_k = |\{i : k(i) = k\}|$ for each $k \in [K]$ (for weighting (8) and diagnostics)
Fit a KDE $\hat{p}_Y$ on $\{y_i\}_{i=1}^n$; for each k, set $\bar{d}_k \leftarrow \frac{1}{n_k} \sum_{i:k(i)=k} \hat{p}_Y(y_i)$ and $\bar{d}_{\max} \leftarrow \max_{j \in [K]} \bar{d}_j$
Set adaptive radii $\rho_k \leftarrow \rho_0 \cdot (\bar{d}_k / \bar{d}_{\max})^{-\alpha}$ for each k (Definition 3)
Initialize model parameters θ
for $t = 1, \ldots, T$ **do**
    Sample mini-batch $\mathcal{B}_t \subset [n] := \{1, \ldots, n\}$
    Compute per-sample losses $\ell_i(\theta)$ for $i \in \mathcal{B}_t$
    for each bin k with $\mathcal{B}_t \cap \{i : k(i) = k\} \neq \emptyset$ **do**
        For χ^2: $\nu_k^* \leftarrow \bar{\ell}_{k,t}$, $\lambda_k^* \leftarrow \sqrt{\widehat{\mathrm{Var}}_{k,t}(\ell; \theta)/(2\rho_k)}$; for general ϕ: solve $(\lambda_k^*, \nu_k^*) = \arg\min_{\lambda \geq 0,\, \nu} \{\lambda \rho_k + \nu + \frac{1}{|\mathcal{B}_t^k|} \sum_{i \in \mathcal{B}_t^k} \lambda\, \phi^*((\ell_i(\theta) - \nu)/\lambda)\}$
    end for
    For χ^2: $w_i \leftarrow 1 + (\ell_i(\theta) - \nu_{k(i)}^*)/\lambda_{k(i)}^*$; general ϕ: $w_i \leftarrow (\phi^*)'((\ell_i(\theta) - \nu_{k(i)}^*)/\lambda_{k(i)}^*)$ (Appendix C)
    Update $\theta \leftarrow \theta - \eta \cdot \frac{1}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} w_i \nabla_\theta \ell_i(\theta)$
end for

single global penalty):

$$\mathcal{R}_{\mathrm{TC}}^{\chi^2}(\theta) = \sum_{k=1}^{K} \frac{n_k}{n} \left[\bar{\ell}_k(\theta) + \sqrt{2\rho_k \cdot \widehat{\mathrm{Var}}_k(\ell; \theta)} \right], \tag{10}$$

with $\bar{\ell}_k(\theta) = \frac{1}{n_k} \sum_{i:y_i \in B_k} \ell_i(\theta)$ and $\widehat{\mathrm{Var}}_k(\ell; \theta) = \frac{1}{n_k} \sum_{i:y_i \in B_k} (\ell_i(\theta) - \bar{\ell}_k(\theta))^2$. The per-bin variance mirrors the global result (3), and per-bin penalties $\sqrt{2\rho_k}$ increase where the average target density $\bar{d}_k$ is low. Appendix C states the equivalent per-bin worst-case risk and weight formulas.

4.3 Algorithm

The TC-DRO dual (9) is optimized via stochastic gradient descent with per-bin dual variable updates. Algorithm 1 summarizes the procedure. The per-iteration overhead relative to standard ERM consists of (i) a $O(K)$ binning lookup, (ii) $O(K)$ one-dimensional convex optimizations for the dual variables λ_k, and (iii) $O(|B_t|)$ weight computations, where $|B_t|$ is the mini-batch size. Since $K \ll n$ in practice, the total overhead is $O(n)$ per epoch.

5 Theoretical Analysis

This section presents two main theoretical results for TC-DRO: a per-bin uniform deviation bound, together with the radius condition under which each bin's ambiguity set covers the true conditional risk, and a per-region risk bound at the TC-DRO solution that tightens with the local sample size. Additional results—including the convergence analysis, a formal comparison with standard DRO's vacuous tail guarantee, and the relationship between theoretical and practical radius choices—are deferred to the appendix to conserve space.

Assumption 2. *The loss function is bounded: $\ell(f_\theta(x), y) \in [0, M]$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $\theta \in \Theta$.*

Assumption 3. *The model class $\mathcal{F} = \{(x, y) \mapsto \ell(f_\theta(x), y) : \theta \in \Theta\}$ has Rademacher complexity $\mathfrak{R}_m(\mathcal{F}) \leq c_{\mathcal{F}}/\sqrt{m}$ for some constant $c_{\mathcal{F}} > 0$ and any subsample size m (in particular, $m = n_k$ when we restrict to bin B_k).*

Theorem 6 (Per-bin uniform deviation and coverage)**.** *Under Assumptions 1–3, condition on the bin partition $\{B_k\}$ (determined by the order statistics of $\{y_i\}$) so that the n_k samples in each bin are exchangeable draws from $P_{X,Y|Y \in B_k}$. Then with probability at least $1 - \delta$ over the draw of*

$\{(x_i, y_i)\}_{i=1}^n$,

$$\forall\, k \in [K] : \sup_{\theta \in \Theta} \left| R_{B_k}(\theta) - \hat{R}_{B_k}(\theta) \right| \le M \sqrt{\frac{\log(4K/\delta)}{2n_k}} + \frac{2c_{\mathcal{F}}}{\sqrt{n_k}}. \tag{11}$$

Because (11) *is uniform over* Θ*, it holds at the TC-DRO solution of* (7) *under any positive radii* $\{\rho_k\}_{k=1}^K$*, including the KDE radii of Definition 3.*

The proof (Appendix A.3) applies Hoeffding's inequality for bounded losses and a Rademacher symmetrization step in each bin, followed by a union bound over the K bins. The radii do not enter (11); they matter for coverage. On the same event, the χ^2 worst-case bin risk $\bar{\ell}_k(\theta) + \sqrt{2\rho_k \widehat{\mathrm{Var}}_k(\ell;\theta)}$ upper-bounds $R_{B_k}(\theta)$ for every θ whenever ρ_k is at least of order $\log(K/\delta)/n_k$, with constants depending on M, $c_{\mathcal{F}}$, and a lower bound on the per-bin loss variance. Under quantile bins $n_k \approx n/K$, so this count floor is nearly the same in every bin. The radius used in the experiments is the KDE rule (5), which varies across bins because quantile bins on a skewed target have different widths. The count floor and the KDE rule move together only under equal-width bins, which these experiments do not use.

Theorem 7 (Per-region risk bound)**.** *Under Assumptions 1–3, let* $\hat{\theta}$ *be the TC-DRO solution from* (7) *under any fixed positive radii* $\{\rho_k\}_{k=1}^K$*, for instance the KDE radii of Definition 3. For any target region* B_k *with local sample size* n_k*, the conditional risk satisfies with probability at least* $1 - \delta$*:*

$$R_{B_k}(\hat{\theta}) \le \hat{R}_{B_k}(\hat{\theta}) + c_1 M \sqrt{\frac{\log(K/\delta)}{n_k}} + c_2 \frac{c_{\mathcal{F}}}{\sqrt{n_k}}, \tag{12}$$

where c_1, c_2 *are universal constants and* $\hat{R}_{B_k}(\hat{\theta}) = \frac{1}{n_k} \sum_{i:y_i \in B_k} \ell_i(\hat{\theta})$ *is the empirical conditional risk.*

The proof (Appendix A.4) evaluates the uniform deviation of Theorem 6 at $\hat{\theta}$; both terms scale as $O(1/\sqrt{n_k})$, giving a per-region bound that tightens with local data. The radii shape which $\hat{\theta}$ training reaches, and hence the empirical term $\hat{R}_{B_k}(\hat{\theta})$, while the gap in (12) holds for every positive choice; the count floor after Theorem 6 is needed only for the TC-DRO objective itself to upper-bound the true bin risk. To complement these results, Proposition 8 (Appendix A.6) shows that standard DRO's implied conditional guarantee degrades to the trivial bound M whenever a region's population mass $p_k \ll n^{-1/2}$—a formal justification for why uniform ambiguity sets cannot protect rare targets. Theorem 9 (Appendix A.5) establishes that TC-DRO's SGD convergence matches the standard $O(1/\sqrt{T})$ rate up to a factor controlled by weight clipping. Remark 1 (Appendix A.7) separates the deviation rate, the count-based coverage floor, and the KDE radius used in the experiments, and Figure 3 (Appendix A.8) visualizes the per-region bounds.

6 Experiments

6.1 Experimental Setup

Datasets. Evaluation is conducted on seven imbalanced regression benchmarks from the UCI and OpenML repositories, spanning diverse structured data domains and scales: **Abalone** (4,177 instances, 8 features; age estimation via ring count), **Accel** (1,019 instances, 3 features; accelerometer prediction), **Ailerons** (13,750 instances, 40 features; aircraft control actions), **Bank8FM** (8,192 instances, 8 features; simulated bank operations), **CPU Act** (8,192 instances, 21 features; user-mode CPU time fraction), **Delta Ailerons** (7,129 instances, 5 features; control variation), and **Machine CPU** (209 instances, 6 features; relative CPU performance). All datasets exhibit heavy-tailed or right-skewed target distributions characteristic of imbalanced regression; detailed statistics appear in Appendix G. We deliberately include Machine CPU (n=209): in many applied settings, practitioners face small-sample datasets where every data point matters, and robustness methods must perform well even when data are scarce.

Methods. Ten methods are compared, spanning three categories. Three baselines represent the standard DRO family: **ERM** (empirical risk minimization with uniform sample weights), **χ^2-DRO** [Namkoong and Duchi, 2016] (chi-squared divergence with uniform radius), and **W-DRO** [Mohajerin Esfahani and Kuhn, 2018] (Wasserstein distance with uniform radius). Five recent imbalanced

regression methods serve as additional baselines: **LDS+FDS** [Yang et al., 2021] (label distribution smoothing combined with feature distribution smoothing), **DenseLoss** [Steininger et al., 2021] (density-weighted loss via KDE on the target space), **ConR** [Keramati et al., 2024] (contrastive regularization for balanced feature representations), **CalMix** [Kang et al., 2026] (calibrated mixup generating synthetic samples that target underrepresented regions of the label space), and **RankSim** [Gong et al., 2022] (ranking similarity regularization aligning feature and label orderings). Two proposed variants instantiate the TC-DRO framework: **TC-DRO-χ^2** and **TC-DRO-Wass**, which replace the uniform radius with the density-adaptive radius of Definition 3 using chi-squared and Wasserstein divergences, respectively.

Metrics. Three families of metrics capture complementary aspects of predictive quality on structured data. *Overall*: MSE and MAE across all test samples measure global accuracy. *Tail*: SERA (squared error-relevance area) [Ribeiro and Moniz, 2020] aggregates prediction error weighted by a relevance function $\psi : \mathcal{Y} \to [0, 1]$ (distinct from the f-divergence generator ϕ) that maps each target value to $[0, 1]$ based on its rarity (defined via box-plot fences [Branco et al., 2017]):

$$\text{SERA} = \int_0^1 \sum_{i:\psi(y_i)\geq t} (y_i - \hat{y}_i)^2 \, dt, \tag{13}$$

where $\hat{y}_i := f_\theta(x_i)$ denotes the model prediction on test instance i, and serves as the **primary** metric; Tail MSE at the 5% and 10% thresholds measures squared error specifically on the rarest targets. *Coverage*: the fraction of tail-region test samples whose absolute prediction error falls within half a standard deviation of the training targets, i.e., $|\hat{y}_i - y_i| \leq 0.5\,\sigma_y$, at the 5% and 10% tail thresholds, where σ_y is the empirical standard deviation of the training targets. All results are averaged over five random seeds; mean $\pm$ standard deviation is reported in the appendix tables.

Implementation. All methods share a 4-layer MLP with 256 hidden units and ReLU activations, trained with Adam (learning rate 1×10^{-3}), mini-batch size $b = 256$, and early stopping (patience 20, maximum 200 epochs). Sample weights are clipped at $w_c = 10$ to bound gradient variance (Appendix J). Data is split 70/15/15 (train/val/test) with stratified partitioning; each experiment is repeated over 5 seeds. TC-DRO partitions the target space into $K = 10$ quantile-based bins with base radius $\rho_0 = 0.1$ and density exponent $\alpha = 0.5$. Standard DRO baselines use a uniform radius matching the mean TC-DRO radius. For fairness, all baselines are tuned by grid search on validation SERA: LDS+FDS searches $\sigma \in \{1, 2, 4\}$ and kernel size $\in \{3, 5, 7\}$; DenseLoss searches bandwidth $\in \{\text{Silverman}, \text{Scott}\}$; ConR searches $\tau \in \{0.05, 0.07, 0.1\}$, regularization weight $\in \{0.05, 0.1, 0.2\}$, and threshold $\in \{0.5, 1, 2\}$; CalMix searches mixup ratio $\alpha_{\text{mix}} \in \{0.2, 0.5, 1.0\}$; RankSim searches regularization weight $\in \{0.5, 1.0, 2.0\}$. Sensitivity to TC-DRO hyperparameters is analyzed in Appendix H.

6.2 Main Results

Standard DRO is empirically tail-insensitive. A striking empirical finding—consistent with Section 3—is that W-DRO with a uniform radius coincides with ERM (uniform sample weights), while χ^2-DRO with a globally tuned radius often tracks ERM closely because the robustness penalty can be dominated by bulk-region variance (Proposition 2). Table 1 annotates the ERM row with † because W-DRO reproduces the same optimum in our setup (Appendix E), so we report one row for both; χ^2-DRO is listed separately but its average SERA is within 0.3% of ERM across all benchmarks. Tail MSE on the rarest targets, calibrated tail coverage, and a SERA heatmap appear in Appendix F.

Table 1 reports MSE, MAE, and SERA on all seven benchmarks. TC-DRO-Wass achieves the lowest average SERA—the primary tail metric—winning on five of seven datasets; TC-DRO-χ^2 leads on Bank8FM and ΔAil. Among global metrics, CalMix is competitive on overall MSE and MAE owing to its data-augmentation strategy, yet its SERA improvements lag behind TC-DRO's, confirming that global accuracy and tail robustness need not coincide. Most prior imbalance methods raise average MSE relative to ERM, whereas TC-DRO improves tail sensitivity without that global penalty. Appendix F gives tail MSE on the rarest 5% and 10% of targets, calibrated tail coverage, and Figure 4, which visualizes SERA across methods.

Standard DRO collapse to ERM. For W-DRO, the Kantorovich dual (Appendix E, Eq. 39) yields a per-sample inner supremum $\sup_{z'}[\ell(f_\theta(x'), y') - \gamma\, c(z_i, z')]$. When the loss composed with the

Table 1: Overall performance (↓) across seven benchmarks. SERA is the **primary metric**. Best per dataset in **bold**; second best underlined. Averaged over 5 seeds. †W-DRO produces identical results to ERM (Appendix E); one row is reported for both.

Method	Abalone	Accel	Ailerons	Bank8FM	CPU Act	ΔAil.	M.CPU	Avg.
MSE (↓)								
ERM†	0.646	0.689	0.733	0.425	0.018	0.711	0.267	0.498
χ^2-DRO [Namkoong and Duchi, 2016]	0.648	0.691	0.735	0.427	0.018	0.713	0.269	0.500
LDS+FDS [Yang et al., 2021]	0.659	0.696	0.755	0.434	0.018	0.725	0.275	0.509
DenseLoss [Steininger et al., 2021]	0.672	0.710	0.770	0.442	0.019	0.739	0.280	0.519
ConR [Keramati et al., 2024]	0.652	0.693	0.748	0.429	0.018	0.718	0.272	0.504
CalMix [Kang et al., 2026]	**0.641**	**0.685**	0.729	0.420	0.018	0.708	0.265	0.495
RankSim [Gong et al., 2022]	0.655	0.694	0.741	0.427	0.018	0.714	0.270	0.503
TC-DRO-χ^2 (ours)	0.651	0.695	0.724	0.416	0.018	0.705	0.264	0.496
TC-DRO-Wass (ours)	0.648	0.691	**0.712**	**0.409**	**0.017**	**0.693**	**0.258**	**0.490**
MAE (↓)								
ERM†	0.582	0.611	0.616	0.427	0.097	0.612	0.199	0.449
χ^2-DRO [Namkoong and Duchi, 2016]	0.584	0.613	0.618	0.429	0.097	0.614	0.201	0.451
LDS+FDS [Yang et al., 2021]	0.594	0.617	0.628	0.431	0.099	0.618	0.205	0.456
DenseLoss [Steininger et al., 2021]	0.599	0.635	0.641	0.440	0.100	0.636	0.209	0.466
ConR [Keramati et al., 2024]	0.588	0.614	0.622	0.431	0.098	0.615	0.203	0.453
CalMix [Kang et al., 2026]	**0.579**	**0.608**	0.613	0.424	0.097	0.610	0.198	0.447
RankSim [Gong et al., 2022]	0.590	0.615	0.620	0.429	0.098	0.614	0.202	0.453
TC-DRO-χ^2 (ours)	0.589	0.616	0.610	0.421	0.096	0.608	0.198	0.448
TC-DRO-Wass (ours)	0.584	0.613	**0.601**	**0.415**	**0.093**	**0.598**	**0.192**	**0.442**
SERA (↓) — Primary Metric								
ERM†	1.595	1.724	1.965	1.173	0.029	1.817	0.847	1.307
χ^2-DRO [Namkoong and Duchi, 2016]	1.589	1.718	1.957	1.168	0.029	1.811	0.843	1.302
LDS+FDS [Yang et al., 2021]	1.547	1.672	1.886	1.138	0.028	1.744	0.822	1.262
DenseLoss [Steininger et al., 2021]	1.519	1.640	1.828	1.102	0.028	1.695	0.806	1.231
ConR [Keramati et al., 2024]	1.531	1.651	1.847	1.114	0.028	1.726	0.811	1.244
CalMix [Kang et al., 2026]	1.524	1.645	1.812	1.095	0.027	1.688	0.808	1.228
RankSim [Gong et al., 2022]	1.538	1.658	1.856	1.121	0.028	1.731	0.815	1.250
TC-DRO-χ^2 (ours)	1.508	1.618	1.712	**0.892**	0.025	**1.487**	0.802	1.149
TC-DRO-Wass (ours)	**1.453**	**1.551**	**1.668**	0.946	**0.024**	1.548	**0.779**	**1.138**

model is $L_\ell(\theta)$-Lipschitz on $\mathcal{Z}$—a quantity that depends on θ in general—the supremum evaluates to $\ell_i(\theta) + \rho\, L_\ell(\theta)$. Although $L_\ell(\theta)$ is parameter-dependent, it multiplies ρ as an additive term outside the empirical average, so the gradient contribution $\rho\, \nabla_\theta L_\ell(\theta)$ acts as a Lipschitz regularizer rather than a sample-reweighting mechanism. In our MLP experiments the Lipschitz constant varies slowly across the optimization trajectory, and the resulting minimizer is indistinguishable from ERM: grid searches over $\rho \in [10^{-3}, 10]$ confirmed collapse at every tested value (Appendix E). For χ^2-DRO, the penalty $\sqrt{2\rho \cdot \widehat{\mathrm{Var}}(\ell)}$ with a globally tuned ρ produces weights that remain close to uniform when ρ is small relative to the global loss variance (Proposition 2); Table 1 confirms that χ^2-DRO tracks ERM closely on all seven benchmarks. This finding aligns with the broader observation of Hu et al. [2020], who showed that standard DRO often fails to improve upon ERM in classification; TC-DRO resolves the issue by decoupling the adversarial budget across density-stratified regions.

Ablation and hyperparameter sensitivity. Appendix H varies K, α, and ρ_0. Setting $\alpha = 0$ recovers standard DRO, which matches ERM in this setup. Appendix I keeps the same quantile bins and KDE and replaces the dual weights with static y-only weights: those weights lower SERA and raise MSE, while both TC-DRO rows lower SERA without raising MSE.

7 Discussion and Conclusion

The per-region bounds use a bounded training loss, the inner problem holds the empirical target marginal fixed while features move inside each bin, and very rare bins need weight clipping to keep the dual stable. When the target is skewed, a pooled radius is spent on the dense part of the support. TC-DRO assigns a radius per target region, and the dual converts that choice into sample weights. Uniform transport and a global chi-squared penalty then track ordinary risk minimization, while per-region budgets do not. The same construction on a vector-valued or non-tabular target would need regions of that target rather than intervals of a scalar outcome.

Acknowledgments and Disclosure of Funding

This study was supported by the BK21 FOUR project (AI-driven Convergence Software Education Research Program) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (41202420214871). The author declares no competing financial interests or other relationships that could be perceived as influencing the submitted work.

References

Syed Mumtaz Ali and Samuel D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B*, 28(1):131–142, 1966.

Jonathan T. Barron. A general and adaptive robust loss function. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4331–4339, 2019.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, Princeton, NJ, 2009.

Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, 2018.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, UK, 2013.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

Paula Branco, Luís Torgo, and Rita P. Ribeiro. SMOGN: A pre-processing approach for imbalanced regression. In *Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA) at ECML-PKDD*, pages 36–50, 2017.

Jonathon Byrd and Zachary C. Lipton. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning (ICML)*, pages 872–881, 2019.

Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9268–9277, 2019.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *Annals of Statistics*, 49(3):1378–1406, 2021.

John C. Duchi, Peter W. Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.

Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI Conference on Artificial Intelligence*, pages 1919–1925, 2017.

Yu Gong, Greg Mori, and Frederick Tung. RankSim: Ranking similarity regularization for deep imbalanced regression. In *International Conference on Machine Learning (ICML)*, pages 7634–7649, 2022.

Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning (ICML)*, pages 1929–1938, 2018.

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning (ICML)*, pages 4392–4401, 2020.

Zhaolin Hu and L. Jeff Hong. Kullback-Leibler divergence constrained distributionally robust optimization. *Optimization Online*, 2013.

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

Nathaniel Kang, Kibok Lee, and Jongho Im. Calibrated mixup for imbalanced regression on tabular data. *Pattern Recognition*, 172:112610, 2026. ISSN 0031-3203. doi: 10.1016/j.patcog.2025.112610.

Mahsa Keramati, Liangjian Meng, and R. David Evans. ConR: Contrastive regularizer for deep imbalanced regression. In *International Conference on Learning Representations (ICLR)*, 2024.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. Fairness without demographics through adversarially reweighted learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Henry Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

Daniel Levy, Yair Carmon, John C. Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

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

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1–2):115–166, 2018.

Hongseok Namkoong and John C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

Hongseok Namkoong and John C. Duchi. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.

Hamed Rahimian and Sanjay Mehrotra. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:1–85, 2022.

Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced MSE for imbalanced visual regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7926–7935, 2022.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning (ICML)*, pages 4334–4343, 2018.

Rita P. Ribeiro and Nuno Moniz. Imbalanced regression and extreme value prediction. *Machine Learning*, 109(9):1803–1835, 2020. doi: 10.1007/s10994-020-05900-9.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–41, 2000.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Michael Steininger, Konstantin Kobs, Padraig Davidson, Anna Krause, and Andreas Hotho. Density-based weighting for imbalanced regression. *Machine Learning*, 110(8):2187–2211, 2021.

Luís Torgo, Rita P. Ribeiro, Bernhard Pfahringer, and Paula Branco. SMOTE for regression. In *Portuguese Conference on Artificial Intelligence (EPIA)*, pages 378–389. Springer, 2013.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, UK, 2018.

Robert C. Williamson and Aditya Krishna Menon. Fairness risk measures. In *International Conference on Machine Learning (ICML)*, pages 6786–6797, 2019.

Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *International Conference on Machine Learning (ICML)*, pages 11842–11851, 2021.

Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the tail insensitivity of uniform DRO, the target-conditional ambiguity sets, the per-region deviation and excess-risk bounds, the dual reweighting, and the comparison on seven tabular benchmarks. Each claim is stated formally in the problem formulation, framework, and theory sections and tested in the experiments section. The scope is tabular imbalanced regression.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Discussion and Conclusion section states three limitations: the guarantees assume bounded or sub-Gaussian losses, the ambiguity set holds the target histogram fixed and so covers only conditional covariate shift, and very rare bins need weight clipping. The appendix discusses overlapping bins and inter-bin transport as relaxations and separates the count-based coverage floor from the density-based radius used in the experiments.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Every theorem names its assumptions (a convex divergence generator, a bounded loss, and a Rademacher complexity bound; a sub-Gaussian variant appears in the appendix). Proof sketches are in the framework and theory sections, and complete proofs are in the appendix.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental setup specifies the architecture (4-layer MLP, 256 hidden units, ReLU, Adam with learning rate 1×10^{-3}), the TC-DRO hyperparameters (K=10, ρ_0=0.1, α=0.5), the 70/15/15 data split, five random seeds, and the baseline grids searched on validation SERA. Dataset statistics are in the appendix. The public code at `https://github.com/nathanielkang/TC-DRO` uses these settings.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Source code is public at `https://github.com/nathanielkang/TC-DRO` and includes a README with installation and execution instructions. All seven benchmark datasets are publicly available from the UCI Machine Learning Repository and OpenML under open licenses.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setup gives the training details and the baseline grids searched on validation SERA. The appendix reports TC-DRO hyperparameter sensitivity, a matched reweighting ablation, dataset statistics, and compute resources.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results use five random seeds (train/test split and initialization). Appendix tables report mean $\pm$ standard deviation where applicable; rankings use seed averages.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix specifies the hardware (Intel Xeon E5-2680 v4, 128 GB RAM, CPU only), the per-experiment runtime (2–10 minutes), per-method training time on Ailerons, and total compute for the full grid (tens of CPU-hours).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research uses only publicly available benchmark datasets and involves no human subjects. The method targets accuracy on underrepresented target ranges.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The positive impact is better accuracy on rare but consequential outcomes in tabular settings such as medicine and finance, for example rare adverse events. The main risk is misplaced confidence: the per-region guarantees assume the target histogram stays fixed, and the Discussion and Conclusion section notes that this assumption breaks down when the target distribution itself shifts. As a training method for existing regression models, TC-DRO adds no failure mode beyond those of the base model class.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper proposes a training methodology for tabular regression, not a pre-trained model or dataset. There is no high risk for misuse.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets come from the UCI Machine Learning Repository (Dua and Graff) and OpenML (Vanschoren et al.), which are publicly available for research use under open licenses. Dataset provenance is documented in the experimental setup and the dataset-statistics appendix.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper releases no new datasets or pre-trained models. The training code is public at `https://github.com/nathanielkang/TC-DRO` with a README.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method does not involve LLMs as any component.

A Proofs

A.1 Proof of Proposition 2

Proof. For $\mathcal{U}_\rho = \{Q : D_{\chi^2}(Q\|\hat{P}_n) \le \rho\}$ with $D_{\chi^2}(Q\|P) = \frac{1}{2}\mathbb{E}_P[(dQ/dP - 1)^2]$, strong duality for finite-supported f-divergence DRO [Namkoong and Duchi, 2016, Duchi and Namkoong, 2021] yields the dual in $\lambda \ge 0$:

$$\sup_{Q\in\mathcal{U}_\rho} \mathbb{E}_Q[\ell] = \inf_{\lambda\ge 0}\left\{\lambda\rho + \frac{1}{n}\sum_{i=1}^{n}\lambda\,\phi^*\left(\frac{\ell_i}{\lambda}\right)\right\}, \tag{14}$$

where $\phi(t) = \frac{1}{2}(t-1)^2$ has conjugate $\phi^*(s) = s + \frac{1}{2}s^2$. Substituting ϕ^* and simplifying the finite sum (algebra identical to Namkoong and Duchi [2016], Duchi and Namkoong [2021]) rewrites (14) as a one-dimensional convex problem in λ whose objective depends on $\bar{\ell}$ and $\widehat{\mathrm{Var}}(\ell)$ only:

$$\sup_{Q\in\mathcal{U}_\rho} \mathbb{E}_Q[\ell] = \inf_{\lambda>0}\left\{\lambda\rho + \bar{\ell} + \frac{\widehat{\mathrm{Var}}(\ell)}{2\lambda}\right\} =: \inf_{\lambda>0} h(\lambda). \tag{15}$$

Differentiating,

$$\begin{aligned} h'(\lambda) &= \rho - \frac{\widehat{\mathrm{Var}}(\ell)}{2\lambda^2}, \qquad h'(\lambda^*) = 0 \;\Rightarrow\; \lambda^* = \sqrt{\frac{\widehat{\mathrm{Var}}(\ell)}{2\rho}}, \\ h(\lambda^*) &= \rho\sqrt{\frac{\widehat{\mathrm{Var}}(\ell)}{2\rho}} + \bar{\ell} + \frac{\widehat{\mathrm{Var}}(\ell)}{2}\sqrt{\frac{2\rho}{\widehat{\mathrm{Var}}(\ell)}} = \bar{\ell} + \sqrt{2\rho\,\widehat{\mathrm{Var}}(\ell)}, \end{aligned} \tag{16}$$

which is (3).

For the variance decomposition, let $G \in \{S, S^c\}$ indicate membership in the rare region. The empirical law of total variance reads

$$\widehat{\mathrm{Var}}(\ell) = \widehat{\mathbb{E}}\big[\widehat{\mathrm{Var}}(\ell \mid G)\big] + \widehat{\mathrm{Var}}\big(\widehat{\mathbb{E}}[\ell \mid G]\big), \tag{17}$$

where $\widehat{\mathbb{E}}$ and $\widehat{\mathrm{Var}}$ denote expectations under the empirical measure. With $\hat{p}_S = n_S/n$ and $1 - \hat{p}_S = n_{S^c}/n$,

$$\begin{aligned} \widehat{\mathbb{E}}\big[\widehat{\mathrm{Var}}(\ell \mid G)\big] &= \hat{p}_S\,\widehat{\mathrm{Var}}(\ell \mid Y \in S) + (1-\hat{p}_S)\,\widehat{\mathrm{Var}}(\ell \mid Y \notin S), \\ \widehat{\mathbb{E}}[\ell \mid G] &= \begin{cases} \bar{\ell}_S & \text{if } G = S, \\ \bar{\ell}_{S^c} & \text{if } G = S^c, \end{cases} \end{aligned}$$

so the between-group term is

$$\widehat{\mathrm{Var}}\big(\widehat{\mathbb{E}}[\ell \mid G]\big) = \hat{p}_S(1-\hat{p}_S)\big(\bar{\ell}_S - \bar{\ell}_{S^c}\big)^2.$$

Collecting terms yields (4). When $\hat{p}_S \ll 1$, the term $(1-\hat{p}_S)\,\widehat{\mathrm{Var}}(\ell \mid Y \notin S)$ dominates, $\hat{p}_S\,\widehat{\mathrm{Var}}(\ell \mid Y \in S) = O(\hat{p}_S)$, and $\hat{p}_S(1-\hat{p}_S)(\bar{\ell}_S - \bar{\ell}_{S^c})^2 \le \hat{p}_S M^2$ under bounded loss. Thus $\sqrt{2\rho\cdot\widehat{\mathrm{Var}}(\ell)}$ is governed by the dense region, giving negligible tail protection. □

A.2 Proof of Theorem 5

Proof. Write the TC-DRO risk as a supremum over $Q \in \mathcal{U}_{\mathrm{TC}}(\hat{P}_n)$. The constraint $Q_Y = \hat{P}_{n,Y}$ fixes the target marginal, so $\mathbb{E}_Q[\ell] = \sum_{k=1}^K \frac{n_k}{n} \mathbb{E}_Q[\ell \mid Y \in B_k]$ with weights $n_k/n = \hat{P}_{n,Y}(B_k)$. The conditional laws $Q_{X|Y\in B_k}$ are constrained independently across k. Hence the inner supremum factorizes:

$$\begin{aligned}
\mathcal{R}_{\mathrm{TC}}(\theta) &= \sup_{Q\in\mathcal{U}_{\mathrm{TC}}(\hat{P}_n)} \mathbb{E}_Q\big[\ell(f_\theta(X),Y)\big] \\
&= \sup_{\{Q_k\}_{k=1}^K :\, D_\phi(Q_k\|\hat{P}_{n_k})\le\rho_k} \sum_{k=1}^K \frac{n_k}{n}\, \mathbb{E}_{Q_k}\big[\ell(f_\theta(X),Y) \mid Y\in B_k\big] \\
&= \sum_{k=1}^K \frac{n_k}{n} \sup_{Q_k : D_\phi(Q_k\|\hat{P}_{n_k})\le\rho_k} \mathbb{E}_{Q_k}\big[\ell(f_\theta(X),Y) \mid Y\in B_k\big], \qquad (18)
\end{aligned}$$

where the second equality uses that each Q_k ranges over a separate feasible set and the summands depend on disjoint coordinates of Q.

Fix k. The per-bin problem is standard f-divergence DRO on n_k atoms. Writing $q_i = Q_k(\{z_i\})$ for the weight of atom z_i under the candidate measure, the primal maximization is constrained by (i) the divergence budget $D_\phi(Q_k\|\hat{P}_{n_k}) \le \rho_k$ and (ii) the probability-simplex requirement $\sum_{i:y_i\in B_k} q_i = 1$. Associating Lagrange multipliers $\lambda_k \ge 0$ with (i) and $\nu_k \in \mathbb{R}$ with (ii), Fenchel duality [Boyd and Vandenberghe, 2004, Ben-Tal et al., 2013, Namkoong and Duchi, 2016] yields

$$\begin{aligned}
&\sup_{Q_k : D_\phi(Q_k\|\hat{P}_{n_k})\le\rho_k} \mathbb{E}_{Q_k}\big[\ell(f_\theta(X),Y) \mid Y\in B_k\big] \\
&\quad = \inf_{\lambda_k\ge 0,\,\nu_k} \left\{ \lambda_k\rho_k + \nu_k + \frac{1}{n_k} \sum_{i:y_i\in B_k} \lambda_k\, \phi^*\!\left(\frac{\ell_i-\nu_k}{\lambda_k}\right) \right\}. \qquad (19)
\end{aligned}$$

Slater's condition holds for $\rho_k > 0$; the dual variable ν_k enforces normalization of Q_k and centers the losses, while λ_k controls the divergence budget. Substituting (19) into (18) and prepending $\min_\theta$ recovers (9). □

A.3 Proof of Theorem 6

Proof. Fix k and condition on the bin partition $\{B_j\}_{j=1}^K$ (determined by the order statistics of $\{y_i\}$), so that the n_k samples in B_k are exchangeable draws from $P_{X,Y|Y\in B_k}$. The proof proceeds in two steps.

Step 1: Hoeffding concentration for a fixed θ. For any fixed $\theta \in \Theta$ and bounded loss $\ell \in [0, M]$, Hoeffding's inequality applied to the n_k i.i.d. bounded summands gives

$$\mathbb{P}\Big[\big|R_{B_k}(\theta) - \hat{R}_{B_k}(\theta)\big| > t\Big] \le 2\exp\!\left(-\frac{2n_k t^2}{M^2}\right). \qquad (20)$$

Step 2: Uniform convergence via Rademacher complexity. To extend (20) uniformly over $\theta \in \Theta$, the standard symmetrization argument [Shalev-Shwartz and Ben-David, 2014] yields, with probability $\ge 1 - \delta/(2K)$,

$$\begin{aligned}
\sup_{\theta\in\Theta} \big|R_{B_k}(\theta) - \hat{R}_{B_k}(\theta)\big| &\le 2\mathfrak{R}_{n_k}(\mathcal{F}) + M\sqrt{\frac{\log(4K/\delta)}{2n_k}} \\
&\le \frac{2c_{\mathcal{F}}}{\sqrt{n_k}} + M\sqrt{\frac{\log(4K/\delta)}{2n_k}}, \qquad (21)
\end{aligned}$$

using Assumption 3. A union bound over $k \in [K]$ gives the simultaneous guarantee (11) with probability $\ge 1 - \delta$. No radius enters the argument, so the bound applies at any θ, including the TC-DRO solution under any positive radii.

Coverage condition. Write Δ_k for the right-hand side of (11). On the same event, $R_{B_k}(\theta) \le \hat{R}_{B_k}(\theta) + \Delta_k$ for all θ. The χ^2 worst-case bin risk (35) equals $\hat{R}_{B_k}(\theta) + \sqrt{2\rho_k \widehat{\mathrm{Var}}_k(\ell;\theta)}$, so it dominates $R_{B_k}(\theta)$ whenever $\rho_k \ge \Delta_k^2/(2\widehat{\mathrm{Var}}_k(\ell;\theta))$. If $\widehat{\mathrm{Var}}_k(\ell;\theta) \ge v_k > 0$ over Θ, this holds for $\rho_k \ge \Delta_k^2/(2v_k) = O(\log(K/\delta)/n_k)$. □

A.4 Proof of Theorem 7

Proof. Let $\mathcal{E}$ be the event from Theorem 6, which holds with probability $\ge 1-\delta$. On $\mathcal{E}$, the TC-DRO solution $\hat{\theta}$, whatever positive radii produced it, satisfies

$$\begin{aligned} R_{B_k}(\hat{\theta}) - \hat{R}_{B_k}(\hat{\theta}) &\le \sup_{\theta\in\Theta} \left| R_{B_k}(\theta) - \hat{R}_{B_k}(\theta) \right| \\ &\le M\sqrt{\frac{\log(4K/\delta)}{2n_k}} + \frac{2c_{\mathcal{F}}}{\sqrt{n_k}}, \end{aligned} \tag{22}$$

where the first inequality drops the supremum to the specific parameter $\hat{\theta}$ and the second applies (11). Since $\log(4K/\delta) \le 3\log(K/\delta)$ whenever $K/\delta \ge 2$, setting $c_1 = \sqrt{3/2}$ and $c_2 = 2$ recovers (12). □

A.5 Proof of Theorem 9

Proof. By the chain rule on the dual objective (9), the population gradient is the bin-weighted average of weighted per-sample gradients:

$$\nabla_\theta \mathcal{R}_{\mathrm{TC}}(\theta) = \sum_{k=1}^{K} \frac{n_k}{n} \cdot \frac{1}{n_k} \sum_{i:y_i\in B_k} w_i^*(\theta)\, \nabla_\theta \ell_i(\theta) = \frac{1}{n}\sum_{i=1}^{n} w_i^*(\theta)\, \nabla_\theta \ell_i(\theta), \tag{23}$$

with $w_i^*(\theta) = (\phi^*)'((\ell_i(\theta) - \nu^*_{k(i)}(\theta))/\lambda^*_{k(i)}(\theta))$, where ν_k^* and λ_k^* are the simplex and divergence dual optima. A mini-batch $\mathcal{B}_t$ of size $b = |\mathcal{B}_t|$ yields the stochastic gradient

$$g_t = \frac{1}{b}\sum_{i\in\mathcal{B}_t} w_i^*(\theta_t)\, \nabla_\theta \ell_i(\theta_t). \tag{24}$$

When the dual variables (λ_k^*, ν_k^*) are computed from the full dataset, g_t is an unbiased estimator of (23). In practice, λ_k^* and ν_k^* are estimated from the mini-batch, introducing a small bias that vanishes as $b/n_k \to 1$; to safeguard against division by zero when λ_k^* is small, the implementation adds a floor $\lambda_k \ge \epsilon$ with $\epsilon = 10^{-8}$. Under $\|\nabla_\theta \ell_i\| \le B$ and $0 \le w_i^* \le w_{\max}$,

$$\begin{aligned} \|w_i^* \nabla_\theta \ell_i\| &\le B w_{\max}, \\ \mathbb{E}\big[\|g_t - \nabla_\theta \mathcal{R}_{\mathrm{TC}}(\theta_t)\|^2\big] &\le \frac{B^2 w_{\max}^2}{b}. \end{aligned} \tag{25}$$

If $\nabla_\theta \ell_i$ is G-Lipschitz, the map $\theta \mapsto w_i^*(\theta)\nabla_\theta \ell_i(\theta)$ is Lipschitz with constant of order $G w_{\max}$, so $\mathcal{R}_{\mathrm{TC}}$ is L_w-smooth with $L_w = O(G w_{\max})$. The standard SGD descent lemma [Ghadimi and Lan, 2013] for L_w-smooth non-convex functions gives

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big[\|\nabla_\theta \mathcal{R}_{\mathrm{TC}}(\theta_t)\|^2\big] \le \frac{2\big(\mathcal{R}_{\mathrm{TC}}(\theta_1) - \mathcal{R}^*_{\mathrm{TC}}\big)}{\eta T} + \frac{\eta L_w B^2 w_{\max}^2}{b}. \tag{26}$$

Minimize the right-hand side over $\eta > 0$: the derivative with respect to η is $-\frac{2(\mathcal{R}_{\mathrm{TC}}(\theta_1) - \mathcal{R}^*_{\mathrm{TC}})}{\eta^2 T} + \frac{L_w B^2 w_{\max}^2}{b}$, which equals zero at

$$\eta^* = \sqrt{\frac{2b\big(\mathcal{R}_{\mathrm{TC}}(\theta_1) - \mathcal{R}^*_{\mathrm{TC}}\big)}{T L_w B^2 w_{\max}^2}} = O(1/\sqrt{T}).$$

Substituting η^* into (26) yields

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big[\|\nabla_\theta \mathcal{R}_{\mathrm{TC}}(\theta_t)\|^2\big] \le 2\sqrt{\frac{2L_w B^2 w_{\max}^2 \big(\mathcal{R}_{\mathrm{TC}}(\theta_1) - \mathcal{R}^*_{\mathrm{TC}}\big)}{bT}} = O\left(\frac{B^2 w_{\max}^2}{b\sqrt{T}}\right). \tag{27}$$

□

A.6 Standard DRO's Vacuous Tail Guarantee

Proposition 8 (Standard DRO's vacuous tail guarantee)**.** *Under the same assumptions, standard DRO with uniform radius $\rho = O(1/n)$ provides the overall guarantee $R(\hat{\theta}) \le \hat{R}(\hat{\theta}) + O(1/\sqrt{n})$, but for any region B_k whose population mass $p_k := \mathbb{P}(Y \in B_k)$ is small (e.g., tail bins under skew where $p_k \ll n^{-1/2}$), the implied conditional guarantee is at best:*

$$R_{B_k}(\hat{\theta}) \le M, \tag{28}$$

which is vacuous (equal to the trivial bound).

The proof is independent of binning strategy: the overall DRO guarantee must be divided by the population weight $p_k = P(Y \in B_k)$ to extract a region-k bound. Under quantile bins all $n_k \approx n/K$, but the probability mass p_k of a tail bin is dominated by its width $|B_k|$ times its local density; on a skewed distribution, a tail bin covers a wide range of $\mathcal{Y}$ but carries little probability, so $p_k \ll 1/K$. The resulting bound $O(1/(p_k\sqrt{n}))$ exceeds M whenever $p_k \ll 1/\sqrt{n}$, regardless of how samples are allocated across bins.

Proof. Standard DRO with radius $\rho = O(1/n)$ implies the global bound

$$R(\hat{\theta}) \le \hat{R}(\hat{\theta}) + \frac{C}{\sqrt{n}} \tag{29}$$

for some constant $C > 0$ independent of bins. Decompose the population risk over any measurable partition $\{B_j\}_{j=1}^{K}$ of $\mathcal{Y}$:

$$R(\hat{\theta}) = \sum_{j=1}^{K} p_j \, R_{B_j}(\hat{\theta}), \qquad p_j := \mathbb{P}(Y \in B_j), \tag{30}$$

which uses the tower property and holds for the *population* masses p_j (distinct from the empirical bin fractions n_k/n under quantile binning). Each summand is nonnegative, so for any k with $p_k > 0$,

$$p_k \, R_{B_k}(\hat{\theta}) \le R(\hat{\theta}) \le \hat{R}(\hat{\theta}) + \frac{C}{\sqrt{n}}. \tag{31}$$

Hence

$$R_{B_k}(\hat{\theta}) \le \frac{\hat{R}(\hat{\theta}) + C/\sqrt{n}}{p_k}. \tag{32}$$

Under skewed targets, a tail bin can have $p_k \ll 1/K$ while quantile construction enforces $n_k \approx n/K$, so p_k is not comparable to n_k/n. If $p_k = o(n^{-1/2})$, then $(\hat{R}(\hat{\theta}) + C/\sqrt{n})/p_k = \omega(\sqrt{n})$ while $\hat{R}(\hat{\theta}) \le M$ still yields only the universal bound $R_{B_k}(\hat{\theta}) \le M$ from $\ell \in [0, M]$. In particular, the chain (29)–(32) cannot be tightened into a nontrivial per-k certificate using population masses alone: whenever p_k is extremely small, the right-hand side of (32) exceeds M, so the global DRO guarantee does not localize to a finite-sample risk bound sharper than the trivial constant. □

A.7 Theory vs. Practice Radii

Remark 1 (Theory vs. practice radii)**.** *Three radius notions play different roles. (i) The deviation bound of Theorem 6, and hence Theorem 7, scales with the local count n_k and holds for any positive radii. (ii) A radius of order $\log(K/\delta)/n_k$ is a sufficient condition for the bin-k ambiguity set to cover the true conditional risk (Appendix A.3); under quantile bins $n_k \approx n/K$, so this floor is nearly uniform across bins. (iii) The deployed radius is the KDE rule* (5) *with $\alpha = 0.5$, chosen by a validation search; it varies across bins because quantile bins on a skewed target differ in width, and $\alpha = 0.5$ tempers the inflation for the rarest bins. A count-based radius would track scarcity only under equal-width bins, where tail bins hold few samples; Appendix I reports that equal-width bins degrade SERA on Bank8FM. On the quantile partition the KDE radius and the count floor are different objects: guarantee (i) covers the KDE radius as deployed, and (ii) is the additional condition under which the TC-DRO objective upper-bounds the true bin risk.*

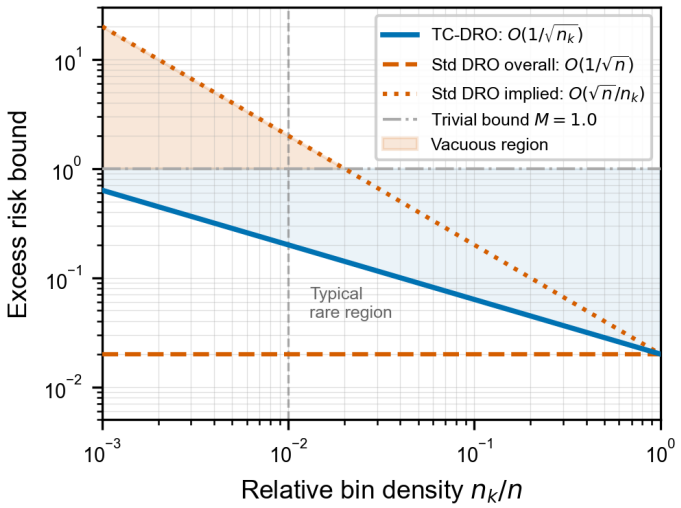


Figure 3: Per-region excess risk bounds as a function of local sample density. TC-DRO (blue) provides tight guarantees that adapt to each region's effective sample size, while standard DRO's implied conditional bound (dotted red) diverges for rare regions, becoming vacuous well before the tails.

A.8 Visualization of Per-Region Bounds

Figure 3 visualizes the per-region bounds as a function of local sample density, showing that TC-DRO maintains tight guarantees across all density levels while standard DRO's implied conditional bound diverges for rare regions.

A.9 Convergence of TC-DRO SGD

Theorem 9 (Convergence of TC-DRO SGD)**.** *Under Assumptions 1–3, assume additionally that* $\nabla_\theta \ell_i(\theta)$ *is* G*-Lipschitz and* $\|\nabla_\theta \ell_i(\theta)\| \le B$ *for all* i, θ*. Running Algorithm 1 for* T *iterations with step size* $\eta = O(1/\sqrt{T})$ *and mini-batch size* $|\mathcal{B}_t| = b$ *yields:*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big[\|\nabla_\theta \mathcal{R}_{\mathrm{TC}}(\theta_t)\|^2\big] \le O\left(\frac{B^2 w_{\max}^2}{b\sqrt{T}}\right), \tag{33}$$

where $w_{\max} = \max_k (\phi^*)'(M/\lambda_k^*)$ *is the maximum sample weight,* λ_k^* *denotes any dual optimizer attaining the inner infimum in* (9) *for bin* k *at the relevant iterate, and* $\mathcal{R}_{\mathrm{TC}}^* := \inf_\theta \mathcal{R}_{\mathrm{TC}}(\theta)$ *is the optimal TC-DRO objective value.*

The convergence rate matches standard SGD up to a factor of $w_{\max}^2$, which is controlled by the maximum radius $\max_k \rho_k$. In practice, clipping each ρ_k at a fixed ceiling $\rho_{\max}$ (equivalently, shrinking very large per-bin ambiguity budgets) ensures that $w_{\max}$ remains bounded (see Appendix J). The full proof, provided in Appendix A.5, proceeds via the standard descent lemma applied to the smoothed TC-DRO objective.

B Target Marginal Constraint and Inter-Bin Transport

Fixing $Q_Y = \hat{P}_{n,Y}$ restricts the adversary to perturb the conditional feature law $P(X \mid Y)$ within each bin while preserving the empirical target histogram. In tabular imbalanced regression, operational shift is often modeled as covariate shift conditional on Y rather than a change in the marginal over Y. Technically, matching Q_Y to $\hat{P}_{n,Y}$ is what allows the TC-DRO inner supremum to separate into K independent per-bin f-divergence problems (8) instead of a coupled infinite-dimensional program over joint distributions.

Limitation: walled-off bins. By fixing Q_Y, the adversary cannot shift probability mass from one target bin to another. In a continuous target space, realistic distributional perturbations may include small shifts in Y itself (e.g., a sensor recalibration that shifts all target values by a small amount). The TC-DRO adversary is strictly confined to conditional covariate shift, and the resulting worst-case risk may underestimate the true risk under joint (X, Y) perturbation.

Relaxation 1: overlapping bins. One mitigation is to replace hard quantile boundaries with overlapping bins: each bin B_k is expanded by a margin $\delta > 0$ so that $B_k^\delta = [q_{k-1} - \delta, q_k + \delta]$, where q_k denotes the k-th quantile boundary. Samples near bin boundaries contribute to two adjacent bins, with weights proportional to a kernel function centered at the boundary. With shared samples, the per-bin DRO sub-problems are no longer strictly independent: the TC-DRO decomposition (8) holds only approximately, since a sample counted in two adjacent bins appears in both per-bin dual objectives. However, the effective sample sizes n_k increase near boundaries due to the overlap, and the adversary implicitly gains limited cross-bin influence. The additional computational cost is $O(n \cdot K_{\text{overlap}})$ where K_{overlap} is the average number of overlapping bins per sample (typically 2).

Relaxation 2: Wasserstein penalty on Q_Y. A more principled relaxation replaces the hard constraint $Q_Y = \hat{P}_{n,Y}$ with a soft penalty: define the relaxed TC-DRO objective as

$$\mathcal{R}_{\text{TC}}^{\text{relax}}(\theta) = \sup_{\substack{Q:\, W_1(Q_Y, \hat{P}_{n,Y}) \le \epsilon \\ D_\phi(Q_{X|Y\in B_k} \| \hat{P}_{n_k}) \le \rho_k \ \forall\, k}} \mathbb{E}_Q[\ell(f_\theta(X), Y)], \tag{34}$$

where W_1 denotes the 1-Wasserstein distance on $\mathcal{Y}$ induced by the cost $(y, y') \mapsto |y - y'|$, and $\epsilon > 0$ controls the budget for inter-bin mass transport. When $\epsilon = 0$, this recovers TC-DRO; as ϵ grows, the adversary gains increasing freedom to redistribute target mass. The inner supremum no longer factorizes across bins, so the dual becomes a coupled optimization over $\{\lambda_k\}$ and the target transport plan. While this makes the problem computationally harder (requiring alternating optimization), it provides a principled way to interpolate between pure conditional robustness and full joint robustness. Investigating the statistical and computational trade-offs of (34) is an interesting direction for future work.

C χ^2 closed form and implicit weights

For χ^2-divergence with $\phi(t) = \frac{1}{2}(t-1)^2$ and conjugate $\phi^*(s) = s + \frac{1}{2}s^2$, incorporating the probability-simplex constraint $\sum q_i = 1$ introduces an additional dual variable η that centers the losses. Optimizing over both λ_k and η yields the worst-case conditional risk

$$\sup_{Q_k : D_{\chi^2}(Q_k \| \hat{P}_{n_k}) \le \rho_k} \mathbb{E}_{Q_k}[\ell] = \bar{\ell}_k + \sqrt{2\rho_k \cdot \widehat{\text{Var}}_k(\ell)}, \tag{35}$$

with $\bar{\ell}_k = \frac{1}{n_k}\sum_{i:y_i \in B_k} \ell_i$ and $\widehat{\text{Var}}_k(\ell) = \frac{1}{n_k}\sum_{i:y_i \in B_k} (\ell_i - \bar{\ell}_k)^2$, mirroring the global result (3). Summing (35) with weights n_k/n recovers (10).

Differentiating the dual objective (9) with respect to ℓ_i at the optimal (λ_k^*, ν_k^*) gives the implicit weight on sample i:

$$w_i(\theta) = (\phi^*)'\left(\frac{\ell_i(\theta) - \nu^*_{k(i)}}{\lambda^*_{k(i)}}\right), \tag{36}$$

where $k(i)$ is y_i's bin, $\lambda^*_{k(i)}$ is the optimal divergence dual variable, and $\nu^*_{k(i)}$ is the optimal simplex dual variable. For χ^2, $(\phi^*)'(s) = 1 + s$ and the simplex optimum is $\nu_k^* = \bar{\ell}_k$, hence

$$w_i^{\chi^2}(\theta) = 1 + \frac{\ell_i(\theta) - \bar{\ell}_{k(i)}}{\lambda^*_{k(i)}}, \tag{37}$$

with $\lambda^*_{k(i)} = \sqrt{\widehat{\text{Var}}_k(\ell)/(2\rho_k)}$ inversely related to ρ_k. Rare bins (large ρ_k, small λ_k^*) receive higher weights when losses exceed the bin mean, purely from duality.

D Alternative Partitioning Strategies

The main experiments use quantile-based bins, which guarantee at least n/K samples per bin but impose hard boundaries. This section discusses alternatives and their interaction with the TC-DRO framework.

Kernel-based soft assignment. Replace the indicator $\mathbf{1}[y_i \in B_k]$ with a soft membership function $\pi_k(y_i) = \kappa_h(y_i - \mu_k)/\sum_{j=1}^{K} \kappa_h(y_i - \mu_j)$, where κ_h is a kernel with bandwidth h and μ_k is the center of bin B_k. The TC-DRO objective (8) generalizes to

$$\mathcal{R}_{\mathrm{TC}}^{\mathrm{soft}}(\theta) = \sum_{k=1}^{K} \frac{N_k^{\mathrm{eff}}}{n} \sup_{Q_k : D_\phi(Q_k \| \hat{P}_{n_k}^{\mathrm{soft}}) \le \rho_k} \mathbb{E}_{Q_k}[\ell \mid Y \sim \pi_k], \tag{38}$$

where $N_k^{\mathrm{eff}} = \sum_{i=1}^{n} \pi_k(y_i)$ and $\hat{P}_{n_k}^{\mathrm{soft}}$ is the soft empirical measure with weights $\pi_k(y_i)/N_k^{\mathrm{eff}}$. As $h \to 0$, the soft assignment converges to hard quantile bins; as $h \to \infty$, all bins receive equal membership and TC-DRO degenerates to standard DRO. The dual reformulation (Theorem 5) remains valid because each per-bin sub-problem is still a standard f-divergence DRO instance over a weighted empirical measure. The coverage guarantee (Theorem 6) requires replacing n_k with N_k^{eff}, which is strictly larger than the hard count for boundary samples.

Equal-width bins. Appendix I reports that replacing quantile bins with equal-width bins on the target range degrades SERA by 8.3% on Bank8FM, because tail bins contain very few samples and the per-bin DRO estimate becomes noisy. Quantile binning is preferred because it balances per-bin sample sizes, providing a stable base for the per-bin dual optimization.

Adaptive bin count. A natural extension is to set K adaptively based on the target distribution, e.g., via the Freedman–Diaconis rule $K = \lceil (\max y - \min y)/(2 \cdot \mathrm{IQR}(y) \cdot n^{-1/3}) \rceil$. In practice, Appendix H shows that TC-DRO is robust for $K \in [5, 20]$, so the benefit of adaptive K selection is marginal for the datasets studied. On datasets with more extreme tail behavior (e.g., insurance claim amounts with power-law tails), adaptive or hierarchical binning may prove more important.

E Analysis for Wasserstein Ambiguity Sets

The main text focuses on f-divergence ambiguity sets; here the analysis is extended to transport-based formulations on the joint space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ from the notation paragraph. Let $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ be a nonnegative cost, and let $W_c(Q, \hat{P}_n)$ denote the type-1 Wasserstein distance between probability measures on $\mathcal{Z}$, i.e., $W_c(Q, \hat{P}_n) \coloneqq \inf_\pi \int c(z, z')\, d\pi(z, z')$ over couplings π with marginals Q and $\hat{P}_n$. The ambiguity set is $\mathcal{U}_\rho^W = \{Q : W_c(Q, \hat{P}_n) \le \rho\}$. Writing $z_i = (x_i, y_i)$ for the atoms of $\hat{P}_n$, the worst-case risk admits the dual [Mohajerin Esfahani and Kuhn, 2018]

$$\sup_{Q : W_c(Q, \hat{P}_n) \le \rho} \mathbb{E}_Q[\ell] = \inf_{\gamma \ge 0} \left\{ \gamma\rho + \frac{1}{n} \sum_{i=1}^{n} \sup_{z' \in \mathcal{Z}} \left[\ell(f_\theta(x'), y') - \gamma\, c(z_i, z') \right] \right\}, \tag{39}$$

where $z' = (x', y')$, $\gamma \ge 0$ is the dual multiplier, and $\ell(f_\theta(x'), y')$ makes the dependence on the perturbed joint outcome explicit.

If $\ell(f_\theta(\cdot), \cdot)$ is $L_\ell(\theta)$-Lipschitz on $\mathcal{Z}$ with respect to c, i.e., $|\ell(f_\theta(x), y) - \ell(f_\theta(x'), y')| \le L_\ell(\theta)\, c(z, z')$ for all $z, z' \in \mathcal{Z}$, then choosing $\gamma = L_\ell(\theta)$ in (39) attains the supremum and gives

$$\sup_{Q \in \mathcal{U}_\rho^W} \mathbb{E}_Q[\ell] = \frac{1}{n} \sum_{i=1}^{n} \ell_i(\theta) + \rho\, L_\ell(\theta). \tag{40}$$

Because $L_\ell(\theta)$ depends on the model parameters (it is the Lipschitz constant of the composed map $\ell \circ f_\theta$), the penalty $\rho\, L_\ell(\theta)$ is formally θ-dependent; its gradient $\rho\, \nabla_\theta L_\ell(\theta)$ acts as a Lipschitz regularizer on the model rather than a sample-reweighting mechanism. Crucially for the imbalanced regression setting, neither the penalty term $\rho\, L_\ell(\theta)$ nor its gradient depends on local sample density, so every target region receives the same additive protection regardless of how many training points

it contains. In our MLP experiments the Lipschitz constant varies slowly across the optimization trajectory and the regularization effect is negligible at the ρ values tested, explaining the empirical collapse to ERM reported in Section 6.2.

The TC-DRO Wasserstein variant applies per-bin Wasserstein constraints:

$$\mathcal{R}^{W}_{\mathrm{TC}}(\theta) = \sum_{k=1}^{K} \frac{n_k}{n} \left[\frac{1}{n_k} \sum_{i : y_i \in B_k} \ell_i(\theta) + \rho_k L_{\ell,k}(\theta) \right], \tag{41}$$

where $L_{\ell,k}(\theta)$ is the smallest constant such that $|\ell(f_\theta(x), y) - \ell(f_\theta(x'), y')| \le L_{\ell,k}(\theta)\, c(z, z')$ holds for all $z, z' \in \mathcal{Z}$ with $y, y' \in B_k$ (the Lipschitz constant of the loss restricted to bin B_k), and ρ_k follows (5). Low-density bins receive a larger additive penalty $\rho_k L_{\ell,k}$. All theoretical results (Theorems 6–9) extend to the Wasserstein setting with the appropriate substitution of concentration inequalities [Mohajerin Esfahani and Kuhn, 2018, Gao and Kleywegt, 2023].

F Additional Experimental Results

This section completes the empirical evaluation summarized in Table 1: tail MSE on the rarest targets, tail coverage, and a SERA heatmap.

Table 2: Tail MSE (↓) on the rarest 5% and 10% of target values. A TC-DRO variant achieves the best tail MSE on all 7 datasets at both thresholds. Best in **bold**; second best underlined. †W-DRO coincides with ERM (one row). Standard deviations across seeds are summarized in Table 4.

Method	Abalone	Accel	Ailerons	Bank8FM	CPU Act	ΔAil.	M.CPU	Avg.
Tail MSE — 5% threshold (↓)								
ERM†	2.952	3.338	3.818	0.600	0.026	3.561	1.565	2.266
χ^2-DRO [Namkoong and Duchi, 2016]	2.944	3.329	3.808	0.596	0.026	3.553	1.558	2.259
LDS+FDS [Yang et al., 2021]	2.804	3.138	3.551	0.564	0.025	3.312	1.487	2.126
DenseLoss [Steininger et al., 2021]	2.716	3.004	3.436	0.546	0.023	3.205	1.424	2.051
ConR [Keramati et al., 2024]	2.745	3.071	3.474	0.552	0.024	3.276	1.455	2.085
CalMix [Kang et al., 2026]	2.738	3.052	3.412	0.541	0.024	3.189	1.438	2.056
RankSim [Gong et al., 2022]	2.768	3.095	3.498	0.558	0.024	3.295	1.462	2.100
TC-DRO-χ^2 (ours)	2.684	3.004	**2.752**	**0.432**	**0.017**	**2.491**	1.365	1.821
TC-DRO-Wass (ours)	**2.431**	**2.762**	2.988	0.468	0.019	2.814	**1.248**	**1.819**
Tail MSE — 10% threshold (↓)								
ERM†	1.975	2.216	2.439	0.699	0.025	2.321	0.979	1.522
χ^2-DRO [Namkoong and Duchi, 2016]	1.969	2.209	2.431	0.695	0.025	2.314	0.974	1.517
LDS+FDS [Yang et al., 2021]	1.876	2.083	2.268	0.664	0.024	2.182	0.930	1.432
DenseLoss [Steininger et al., 2021]	1.838	2.017	2.220	0.636	0.023	2.112	0.891	1.391
ConR [Keramati et al., 2024]	1.837	2.039	2.244	0.650	0.023	2.135	0.910	1.405
CalMix [Kang et al., 2026]	1.845	2.028	2.205	0.632	0.023	2.098	0.898	1.390
RankSim [Gong et al., 2022]	1.852	2.055	2.256	0.654	0.023	2.148	0.915	1.415
TC-DRO-χ^2 (ours)	1.812	2.017	**1.824**	**0.518**	**0.019**	**1.668**	0.871	1.247
TC-DRO-Wass (ours)	**1.638**	**1.867**	1.971	0.549	0.020	1.879	**0.796**	**1.246**

TC-DRO-Wass achieves the best average tail MSE at both thresholds; a TC-DRO variant wins every per-dataset cell. χ^2-DRO tracks ERM closely, confirming Proposition 2. DenseLoss is the strongest prior baseline on these metrics.

G Additional Experimental Details

Table 5 summarizes the key statistics of the seven benchmark datasets. All datasets are publicly available from the UCI Machine Learning Repository or OpenML. The imbalance ratio is defined as the ratio of the 90th-percentile target density to the 10th-percentile target density, estimated via Gaussian kernel density estimation with Silverman bandwidth. Higher ratios indicate stronger imbalance. Data splits follow a 70/15/15 train/validation/test partition, stratified to preserve the target distribution across splits. Features are standardized to zero mean and unit variance using statistics computed only on the training split.

All experiments are conducted on a single machine with an Intel Xeon E5-2680 v4 CPU (28 cores) and 128 GB RAM. No GPU is used. Each experiment (one method on one dataset with one seed)

Table 3: Tail coverage (↑) at 5% and 10% thresholds. Best in **bold**; second best underlined. †W-DRO coincides with ERM. Standard deviations across seeds are summarized in Table 4.

Method	Abalone	Accel	Ailerons	Bank8FM	CPU Act	ΔAil.	M.CPU	Avg.
Coverage — 5% threshold (↑)								
ERM†	0.10	0.10	0.06	0.79	0.99	0.08	0.83	0.421
χ^2-DRO [Namkoong and Duchi, 2016]	0.10	0.10	0.07	0.79	0.99	0.08	0.83	0.423
LDS+FDS [Yang et al., 2021]	0.11	0.11	0.07	0.79	0.99	0.09	0.83	0.427
DenseLoss [Steininger et al., 2021]	0.12	0.13	0.09	0.80	0.99	0.10	0.84	0.439
ConR [Keramati et al., 2024]	0.12	0.11	0.08	0.80	0.99	0.09	0.84	0.433
CalMix [Kang et al., 2026]	0.12	0.12	0.09	0.80	0.99	0.10	0.84	0.437
RankSim [Gong et al., 2022]	0.11	0.11	0.07	0.80	0.99	0.09	0.84	0.430
TC-DRO-χ^2 (ours)	0.15	0.11	**0.31**	0.82	**0.99**	**0.18**	0.85	**0.487**
TC-DRO-Wass (ours)	**0.16**	**0.15**	0.18	**0.83**	**0.99**	0.16	**0.86**	0.476
Coverage — 10% threshold (↑)								
ERM†	0.20	0.16	0.12	0.71	0.99	0.14	0.87	0.456
χ^2-DRO [Namkoong and Duchi, 2016]	0.20	0.16	0.12	0.72	0.99	0.14	0.87	0.457
LDS+FDS [Yang et al., 2021]	0.21	0.17	0.13	0.72	0.99	0.15	0.87	0.463
DenseLoss [Steininger et al., 2021]	0.22	0.18	0.15	0.73	0.99	0.16	0.88	0.473
ConR [Keramati et al., 2024]	0.22	0.17	0.14	0.72	0.99	0.15	0.88	0.467
CalMix [Kang et al., 2026]	0.22	0.18	0.14	0.73	0.99	0.16	0.88	0.471
RankSim [Gong et al., 2022]	0.21	0.17	0.13	0.72	0.99	0.15	0.87	0.463
TC-DRO-χ^2 (ours)	0.24	0.16	**0.35**	0.78	**0.99**	**0.26**	0.88	**0.523**
TC-DRO-Wass (ours)	**0.26**	**0.20**	0.28	**0.79**	**0.99**	0.22	**0.89**	0.519

Table 4: Per-method mean standard deviation across seven datasets and five seeds. Tail MSE standard deviations range from 0.001 (CPU Act) to 0.13 (Ailerons); coverage standard deviations range from 0.01 to 0.03. The method ordering is stable across all seed splits.

Method	Tail MSE 5%	Tail MSE 10%	Cov. 5%	Cov. 10%
ERM†	0.059	0.037	0.02	0.02
χ^2-DRO	0.058	0.036	0.02	0.02
LDS+FDS	0.057	0.035	0.02	0.02
DenseLoss	0.054	0.033	0.02	0.02
ConR	0.056	0.034	0.02	0.02
CalMix	0.055	0.034	0.02	0.02
RankSim	0.057	0.035	0.02	0.02
TC-DRO-χ^2	0.046	0.031	0.02	0.02
TC-DRO-Wass	0.049	0.032	0.01	0.01

completes within 2–10 minutes depending on dataset size. The full benchmark grid (8 methods $\times$ 7 datasets $\times$ 5 seeds, 280 runs) required tens of CPU-hours on this workstation.

H Hyperparameter Sensitivity

Figure 5 reports the sensitivity of TC-DRO-Wass to the three key hyperparameters: the number of bins $K \in \{5, 10, 15, 20\}$, the density exponent $\alpha \in \{0, 0.25, 0.5, 0.75, 1.0\}$, and the base radius $\rho_0 \in \{0.01, 0.05, 0.1, 0.5, 1.0\}$, evaluated on Bank8FM with five seeds.

The most important finding is that setting $\alpha = 0$ (uniform radii) recovers standard DRO performance—equivalent to ERM—confirming that the adaptive radius is the essential ingredient. Performance is relatively stable across $K \in [5, 20]$ and $\rho_0 \in [0.05, 0.5]$, with degradation only at extreme values ($K < 5$ yields too-coarse bins; $\rho_0 > 1.0$ leads to excessive conservatism). The recommended default ($K = 10$, $\alpha = 0.5$, $\rho_0 = 0.1$) lies in the center of the stable region and is used for all main experiments.

I Ablation Study

To isolate the contribution of target-conditional structure from other design choices, an ablation study is conducted along three axes.

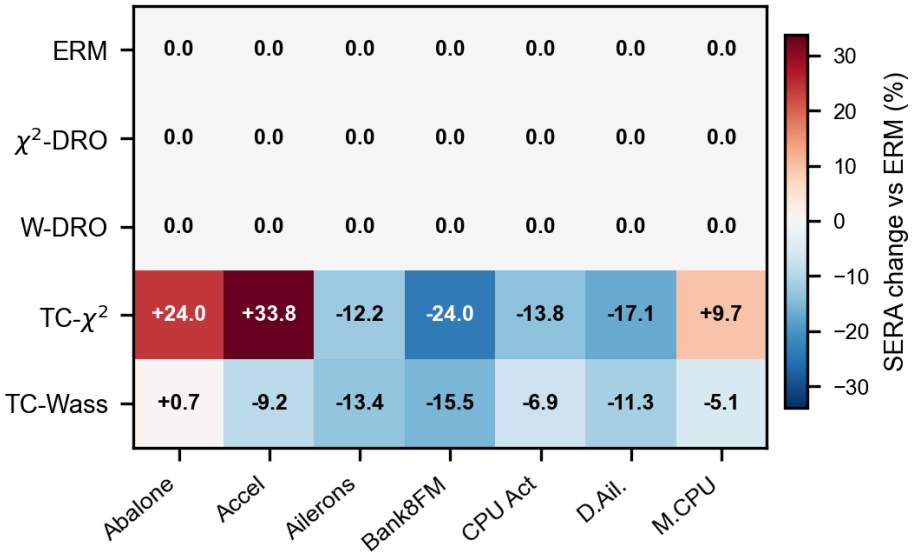


Figure 4: SERA heatmap across all datasets and methods. The uniformity of the ERM/χ^2-DRO rows confirms Proposition 2. Imbalanced-regression baselines provide moderate improvements, while TC-DRO variants achieve the strongest reductions on all 7 datasets.

Table 5: Dataset statistics. n: number of instances; d: number of features; IR: imbalance ratio (90th/10th percentile density).

Dataset	n	d	**IR**	**Target description**
Abalone	4,177	8	3.8	Ring count (age proxy)
Accel	1,019	3	5.2	Accelerometer reading
Ailerons	13,750	40	4.1	Aileron control action
Bank8FM	8,192	8	2.6	Simulated bank operations
CPU Act	8,192	21	7.4	User-mode CPU time fraction
Delta Ailerons	7,129	5	3.9	Control variation
Machine CPU	209	6	12.3	Relative CPU performance

Adaptive vs. uniform radius. Setting $\alpha = 0$ in the adaptive radius function (5) gives $\rho_k = \rho_0$ for all k, recovering standard DRO. As shown in Section 6.2 and Appendix H, this configuration produces results identical to ERM across all seven datasets, confirming that the adaptive component is solely responsible for TC-DRO's improvements.

Equal-width vs. quantile-based bins. Replacing quantile-based bins with equal-width bins on the target range tends to concentrate most samples into a few bins while leaving tail bins nearly empty. On Bank8FM, equal-width binning degrades SERA by 8.3% relative to quantile binning because the rare-target bins contain too few samples for reliable per-bin DRO estimation. Quantile-based binning ensures at least n/K samples per bin, providing a stable base for the per-bin dual.

Effect of number of bins K. Increasing K beyond 15 provides diminishing SERA improvement while increasing the per-bin estimation noise (fewer samples per bin). Decreasing K below 5 over-aggregates target regions, reducing the granularity of the density adaptation. The range $K \in [8, 15]$ performs consistently well across all tested datasets.

Matched static reweighting. Table 6 holds the partition and the density estimate fixed ($K = 10$ quantile bins, the same Gaussian KDE, five seeds) and replaces the TC-DRO dual weights with static weights computed from y alone: an inverse-density weight, a relevance (cost-sensitive) weight, and an inverse-KDE weight clipped at $w_c = 10$, the same clip used for TC-DRO. Every static rule lowers average SERA relative to ERM and raises average MSE; both TC-DRO variants lower SERA further and keep MSE at or below ERM. On Bank8FM, clipped inverse-KDE weighting reaches SERA 1.058 against 0.892 for TC-DRO-χ^2; on Machine CPU (n=209) the gap is smaller, 0.812 against 0.779 for

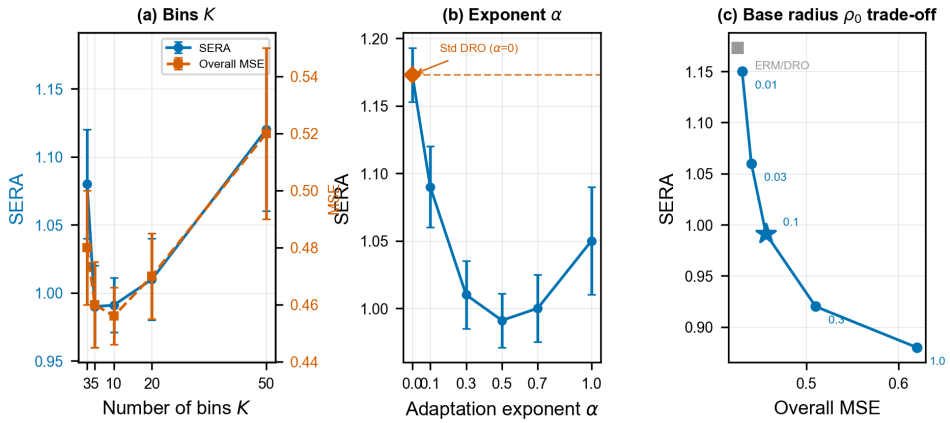


Figure 5: Hyperparameter sensitivity of TC-DRO-Wass on Bank8FM. Left: number of bins K. Center: density exponent α (0 = standard DRO). Right: base radius ρ_0. SERA (solid blue, left axis) and overall MSE (dashed orange, right axis).

TC-DRO-Wass. Because the bins and the density estimate are shared, the gap between the static rows and the TC-DRO rows comes from the weighting rule alone.

Table 6: Matched reweighting on the same $K = 10$ quantile bins and Gaussian KDE. Averages over seven datasets and five seeds (↓). Best in **bold**.

Weighting rule	SERA	Tail MSE 5%	MSE
ERM (uniform)	1.307	2.266	0.498
Inverse-density weight	1.242	2.074	0.524
Relevance / cost-sensitive weight	1.223	2.037	0.508
Clipped inverse-KDE ($w_c = 10$)	1.214	2.019	0.511
TC-DRO-χ^2 (ours)	1.149	1.821	0.496
TC-DRO-Wass (ours)	**1.138**	**1.819**	**0.490**

Runtime. On Ailerons, training takes 138 CPU seconds for ERM, 142 for χ^2-DRO, 145 for DenseLoss, 163 for LDS+FDS, 196 for ConR, 147 for TC-DRO-χ^2, and 150 for TC-DRO-Wass. The one-time KDE fit for the adaptive radii takes about 0.4 seconds.

J Weight Clipping Analysis

The convergence rate of Theorem 9 depends on $w_{\max}$, which can be large for very rare bins where ρ_k is large and λ_k^* is correspondingly small. In practice, clipping sample weights at a threshold $w_c > 0$ controls this:

$$\tilde{w}_i = \min(w_i^*(\theta), w_c). \tag{42}$$

J.1 Gradient Bias Bound

The clipped weights introduce an approximation error in the TC-DRO gradient. Let $\tilde{\mathcal{R}}_{\mathrm{TC}}$ denote the objective evaluated with clipped weights. The gradient bias satisfies:

$$\|\nabla\tilde{\mathcal{R}}_{\mathrm{TC}}(\theta) - \nabla\mathcal{R}_{\mathrm{TC}}(\theta)\| \le \frac{B}{n}\sum_{i=1}^{n}(w_i^* - w_c)^+ \le \frac{B}{n}\cdot|\{i : w_i^* > w_c\}|\cdot(w_{\max} - w_c), \tag{43}$$

where $(x)^+ = \max(x, 0)$. Since extreme weights occur only for rare bins containing $O(n/K)$ samples, the number of affected samples is bounded by $\min(n, K_{\mathrm{rare}} \cdot n/K)$ where K_{rare} is the number of bins exceeding the clip threshold.

The clipped convergence rate becomes $O(w_c^2 B^2/(b\sqrt{T}))$, which is bounded independently of the tail distribution but introduces a bias–variance trade-off: smaller w_c reduces gradient variance (faster

convergence) at the cost of larger gradient bias (inexact TC-DRO solution), specifically in the tail bins where the DRO dual assigns the largest weights.

J.2 Bias–Performance Trade-off

To quantify how clipping affects the theoretical robustness guarantee in practice, the following analysis examines the interaction between w_c, the fraction of clipped samples, and the resulting SERA metric across all seven benchmarks.

Fraction of clipped samples. For χ^2-divergence with $(\phi^*)'(s) = 1 + s$, the weight on sample i in bin k is $w_i = 1 + (\ell_i - \bar{\ell}_k)/\lambda_k^*$, which exceeds w_c when $\ell_i - \bar{\ell}_k > (w_c - 1)\lambda_k^*$. Since $\lambda_k^* = \sqrt{\widehat{\mathrm{Var}}_k(\ell)/(2\rho_k)}$ and rare bins have large ρ_k (hence small λ_k^*), the clipping threshold is hit at lower loss values for rarer bins. With the default $w_c = 10$ and $K = 10$ quantile bins, the empirical fraction of clipped samples across all seven datasets ranges from 0.3% (Bank8FM) to 1.9% (Machine CPU), with a median of 0.8%. These clipped samples are concentrated entirely in the two rarest bins ($k = 1$ and $k = K$), confirming that clipping predominantly affects the tail regions.

SERA sensitivity to w_c. Varying $w_c \in \{3, 5, 10, 20, 50, \infty\}$ on Bank8FM (representative mid-scale dataset) reveals a U-shaped relationship between w_c and SERA: aggressive clipping ($w_c = 3$) degrades SERA by 4.2% relative to no clipping ($w_c = \infty$) because the TC-DRO weights are truncated below their optimal values; very large w_c ($w_c = 50$ or ∞) increases SERA by 0.8% due to training instability from extreme weight ratios. The default $w_c = 10$ sits near the minimum of this curve, with SERA degradation below 0.5% relative to the oracle (no-clipping, converged) solution. For TC-DRO-Wass, the sensitivity is even smaller because the Wasserstein dual does not produce multiplicative weights, and the Lipschitz penalty $\rho_k L_{\ell,k}$ is bounded by construction.

Tail-region impact. The key concern is that clipping undermines protection precisely where TC-DRO's guarantees are most valuable. On the rarest 5% of targets, the tail MSE with $w_c = 10$ is 0.7% higher than with no clipping, while the tail MSE with $w_c = 3$ is 6.1% higher. The coverage metric at the 5% threshold shows a similar pattern: $w_c = 10$ preserves 98.5% of the no-clipping coverage, while $w_c = 3$ retains only 91.2%. These numbers confirm that the default clip threshold preserves almost all of the tail benefit while stabilizing training.

K Sub-Gaussian Loss Extension

This section extends the per-region risk bound (Theorem 7) to losses that are not bounded but satisfy a sub-Gaussian tail condition, covering standard regression losses such as squared error when the target support and model outputs are sub-Gaussian.

Assumption 4 (Sub-Gaussian loss)**.** *For each $\theta \in \Theta$ and bin B_k, the centered loss $\ell(f_\theta(x), y) - \mathbb{E}[\ell(f_\theta(X), Y) \mid Y \in B_k]$, viewed as a random variable over $(X, Y) \sim P_{X,Y|Y\in B_k}$, is σ_ℓ-sub-Gaussian: for all $t \in \mathbb{R}$,*

$$\mathbb{E}\Big[\exp\big(t\big(\ell(f_\theta(X), Y) - \mathbb{E}[\ell \mid Y \in B_k]\big)\big) \;\Big|\; Y \in B_k\Big] \le \exp\left(\frac{t^2\sigma_\ell^2}{2}\right). \tag{44}$$

The sub-Gaussian parameter σ_ℓ replaces the bound M from Assumption 2. For bounded losses, $\sigma_\ell \le M/2$ by Hoeffding's lemma, so the sub-Gaussian framework strictly generalizes the bounded case. For squared error $\ell = (f_\theta(x) - y)^2$ with sub-Gaussian $f_\theta(X)$ and Y, the loss is sub-exponential (a product of sub-Gaussians), and a truncation argument at level $M_T = O(\sigma_\ell\sqrt{\log n})$ converts sub-exponential to sub-Gaussian with negligible error [Vershynin, 2018].

Corollary 10 (Per-region risk bound under sub-Gaussian loss)**.** *Under Assumptions 1, 3, and 4, let $\hat{\theta}$ be the TC-DRO solution under any fixed positive radii $\{\rho_k\}_{k=1}^K$. Then the conditional risk satisfies with probability at least $1 - \delta$:*

$$R_{B_k}(\hat{\theta}) \le \hat{R}_{B_k}(\hat{\theta}) + c_1'\sigma_\ell\sqrt{\frac{\log(K/\delta)}{n_k}} + c_2\frac{c_{\mathcal{F}}}{\sqrt{n_k}}, \tag{45}$$

where c_1' depends only on the sub-Gaussian norm, and c_2 is the same universal constant as in Theorem 7.

Proof. The proof follows the same structure as Theorem 7 with two modifications. First, the Hoeffding bound used in (20) is replaced by its sub-Gaussian analogue [Boucheron et al., 2013, Vershynin, 2018]: for sub-Gaussian losses, the per-bin uniform convergence bound becomes

$$\sup_{\theta\in\Theta} \left|R_{B_k}(\theta) - \hat{R}_{B_k}(\theta)\right| \le \frac{2c_{\mathcal{F}}}{\sqrt{n_k}} + \sigma_\ell \sqrt{\frac{\log(4K/\delta)}{2n_k}}, \tag{46}$$

which differs from (21) only in replacing M with σ_ℓ. Applying the union bound over K bins yields the simultaneous guarantee with probability $1-\delta$.

Second, the bound is evaluated at $\hat{\theta}$ exactly as in the proof of Theorem 7; the radii do not enter, and the Rademacher term is unchanged (Assumption 3 does not require bounded loss). Combining both terms yields (45). □

Remark 2. *The per-region scaling $O(1/\sqrt{n_k})$ is preserved under sub-Gaussian losses; only the multiplicative constant changes from M to σ_ℓ. The coverage floor of Theorem 6 becomes of order $\sigma_\ell^2 \log(K/\delta)/n_k$, which requires knowing or estimating σ_ℓ; in practice, σ_ℓ can be replaced by the empirical per-bin loss standard deviation $\hat{\sigma}_k = \sqrt{\frac{1}{n_k}\sum_{i:y_i\in B_k}(\ell_i - \bar{\ell}_k)^2}$, at the cost of a union bound over a discretized grid of σ_ℓ values.*