DISTRIBUTIONALLY ROBUST SURVIVAL MODELS UNDER SUBPOPULATION SHIFT AND OUTLIER CONTAMINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning robust survival models under distribution shift is an important but challenging problem in many applications. In heterogeneous populations, a model that performs well on average may still perform poorly on certain subpopulations, and this issue becomes even more severe when the training data are contaminated by outliers. In this paper, we propose a novel distributionally robust framework for survival analysis that jointly addresses latent subpopulation shift and outlier contamination. The proposed method combines an outer minimization that selects a refined nominal distribution by reducing the influence of contaminated samples and an inner maximization that focuses on the most challenging subpopulation. This formulation directly accommodates non-decomposable survival losses while preserving interactions across samples, including the risk-set structure of the Cox negative partial log-likelihood. We develop an alternating gradient-based algorithm with outer updates derived from the KKT conditions of the inner maximization. Experiments on simulated data and two survival benchmarks demonstrate that the proposed method remains robust when subpopulation shift and outlier contamination occur simultaneously. It stabilizes training in contaminated settings and substantially improves worst-group performance across both linear and nonlinear survival models, while maintaining competitive and sometimes superior overall performance.

1 INTRODUCTION

Survival analysis is a fundamental problem in statistics and machine learning, with broad applications in healthcare, reliability analysis, and customer retention. In many applications, the population comprises subpopulations with distinct survival patterns. Such heterogeneity may arise across demographic groups such as age, sex, and race, as well as across hospitals, clinical environments, or latent patient subtypes (Hu & Chen, 2024; Rahman & Purushotham, 2022; Fan et al., 2024). When the mixture proportions of these subpopulations differ between training and test distributions, a model with strong average performance can still perform poorly on some subpopulations (Sagawa et al., 2020; Yang et al., 2023; Hu & Chen, 2024).

To address this challenge, distributionally robust learning, or distributionally robust optimization[1] (DRO), has been widely studied as a principled framework for improving robustness under distributional shift (Mohajerin Esfahani & Kuhn, 2018; Duchi & Namkoong, 2021). Applying DRO to survival analysis, however, is not straightforward. Many commonly used survival losses are non-decomposable, involving interactions across samples. For example, the Cox negative partial log-likelihood depends on risk-set interactions, whereas ranking-based survival losses are defined through pairwise comparisons across samples (Cox, 1972; Raykar et al., 2007; Lee et al., 2018). In contrast, many existing DRO methods assume additive losses, which decompose into a sum of terms that each depend on a single sample. This structural difference prevents these methods from being directly applied to such survival losses (Jin et al., 2025; Hu & Chen, 2022; 2024).

[1]While DRO was originally studied in the context of optimization, we use the term more broadly to refer to learning methods for robustness under distribution shift.

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

Recently, Hu & Chen (2024) proposed an f-divergence based DRO (f-DRO) framework for survival models, focusing on the χ^2-divergence, and introduced a sample-splitting strategy to handle the interaction structure of survival losses. Their results show that f-DRO can improve robustness to latent subpopulation shift in survival analysis. However, this approach is highly sensitive to outliers or contaminated observations, since it improves robustness by assigning larger weights to high-loss samples. DORO (Zhai et al., 2021) addresses both subpopulation shift and outlier contamination, but is designed for additive losses and does not directly apply to non-decomposable survival losses with cross-sample dependence.

To address these limitations, we propose a coupled distributionally robust framework for survival analysis that jointly addresses latent subpopulation shift and outlier contamination. The proposed formulation is based on a mixture-based ambiguity set with different roles in a coupled minimax problem. The outer minimization refines the nominal distribution to reduce the influence of potentially contaminated samples. Based on this refined distribution, the inner maximization concentrates on the most challenging subpopulation. In this way, the proposed method iteratively refines the nominal distribution by reducing the influence of contaminated samples and reweights the most challenging subpopulation within the refined distribution, providing robustness to both outlier contamination and subpopulation shift.

For differentiable non-decomposable survival losses, we develop an efficient alternating gradient-based algorithm with outer updates derived from the KKT conditions of the inner maximization. We instantiate the framework with the Cox negative partial log-likelihood, preserving its risk-set interactions. We further show that, for additive losses, the formulation admits a two-stage structure: the outer minimization performs ϵ-trimming, and the inner maximization reweights the highest-loss α-fraction of the remaining probability mass.

Through simulation studies and experiments on two widely used survival benchmarks, METABRIC and FLC, we show that the proposed method remains robust under subpopulation shift and outlier contamination. In simulations, it exhibits stable training trajectories even in contaminated settings, whereas f-DRO exhibits unstable trajectories. On real benchmarks, it yields substantially improved worst-group performance across both linear and nonlinear survival models. It also achieves competitive overall performance, often even outperforming existing baselines in contaminated settings.

Our main contributions can be summarized as follows.

- We propose a novel coupled distributionally robust framework for survival analysis that jointly addresses two practically important challenges: latent subpopulation shift and outlier contamination.
- We develop an alternating gradient-based algorithm that directly optimizes non-decomposable survival losses while preserving interactions across samples. Its outer updates are derived from the KKT conditions of the inner maximization.
- We demonstrate through simulation studies and real data experiments that the proposed method remains stable under outlier contamination and achieves substantially improved worst-group performance under subpopulation shift.

2 Related Work

Survival Analysis under Subpopulation Shift Survival analysis has traditionally been developed based on classical statistical frameworks, including nonparametric estimators, proportional hazards models, and related semiparametric approaches (Cox, 1972; Kaplan & Meier, 1958; Buckley & James, 1979; Bennett, 1983; Kalbfleisch & Prentice, 2002). More recently, it has been extended to machine learning methods that allow more flexible nonlinear modeling of survival outcomes (Ishwaran et al., 2008), and further to deep learning based approaches designed for richer representation learning and more flexible risk prediction (Lee et al., 2018; Ranganath et al., 2016; Katzman et al., 2018; Avati et al., 2020; Tang et al., 2023). However, these models do not explicitly account for subpopulation shift, and therefore often perform poorly on specific groups.

This issue has been studied in survival analysis mainly from two directions: fairness across observed groups and federated survival analysis under institutional heterogeneity. Fairness-based approaches typically impose constraints or regularization to encourage similar predictions or predictive performance across demographic or clinically defined groups (Keya et al., 2021; Zhang & Weiss, 2022; 2023; Rahman & Purushotham, 2022; Do et al., 2023). Federated survival methods, in contrast,

address heterogeneity across hospitals or institutions with different local patient populations and data distributions (Andreux et al., 2020; Archetti et al., 2023; Seidi et al., 2024; Archetti & Matteucci, 2023; Rahman & Purushotham, 2023; 2025; Wang et al., 2022). However, these approaches generally rely on explicit group information, institution-level partitioning, or other prespecified heterogeneous structure.

Distributionally Robust Optimization Distributionally robust learning has emerged as a promising framework for handling various forms of distributional shift (Lin et al., 2022; Delage & Ye, 2010; Mohajerin Esfahani & Kuhn, 2018; Kuhn et al., 2019; Gao et al., 2024; Sagawa et al., 2020). In particular, DRO has been widely studied for latent subpopulation shift, where robustness must be achieved without explicit group annotations. A prominent line of work improves worst-case performance by emphasizing high-loss regions of the distribution, as in f-DRO and related CVaR-type objectives (Namkoong & Duchi, 2016; Duchi & Namkoong, 2021; Hashimoto et al., 2018; Shapiro et al., 2021). Recent work has extended this idea to survival analysis without relying directly on predefined sensitive attributes (Hu & Chen, 2022; 2024).

However, these methods can be highly vulnerable to outliers and may become unstable under contamination (Zhai et al., 2021). This issue is particularly important in survival analysis, where outliers are pervasive and can substantially affect model fitting and prediction. Existing robust survival methods mitigate outlier effects (Heritier et al., 2009; Sinha, 2019; Farcomeni & Ventura, 2012; Luo et al., 2022; Farcomeni & Viviani, 2011), but do not directly address latent subpopulation shift. Zhai et al. (2021) improve outlier robustness in f-DRO but focus on additive losses and do not cover non-decomposable survival losses with cross-sample dependence. Thus, jointly addressing outlier contamination and latent subpopulation shift remains an important problem in survival analysis. Independently of our work, Huang et al. (2026) propose a nested DRO framework for robust learning with additive losses. Their formulation does not directly apply to non-decomposable survival losses with cross-sample dependence.

Other Approaches to Subpopulation Shift Subpopulation shift has also been studied through latent subgroup discovery and annotation-free group robustness (Liu et al., 2021; Hong et al., 2024; Stromberg et al., 2024), group-aware priors (Rudner et al., 2024), outlier-robust group inference (Zeng et al., 2023), and DRO frameworks with hierarchical or structured ambiguity sets (Jo et al., 2026; Kim et al., 2026). Our work focuses on jointly addressing subpopulation shift and outlier contamination in survival models with non-decomposable losses.

3 Problem Setup

In this section, we introduce the notation and formalize the problem setup. Although our primary interest lies in survival analysis, we first present our framework in a general statistical learning setting, and then discuss its application to survival models in Section 4.3.

Let $Z \in \mathcal{Z}$ denote a random variable. For any distribution Q, we denote expectation under Q by $\mathbb{E}_Q[\cdot]$. For a positive integer n, let $[n] := \{1, \dots, n\}$ denote the index set, and let $\mathcal{S}_{n-1} := \{p \in \mathbb{R}^n : p_i \geq 0,\ \sum_{i=1}^n p_i = 1\}$ denote the $(n-1)$-simplex. Given a model f_θ parameterized by θ, we denote the corresponding loss function by $\ell(Z, \theta)$.

We consider a learning problem under subpopulation shift with outlier contamination. Specifically, the target distribution P_{test} is assumed to be a mixture of K latent groups, $P_{\text{test}} = \sum_{k=1}^K \pi_k P_k$, where $\sum_{k=1}^K \pi_k = 1$ and $\pi_k \in (0, 1)$. Here, P_k denotes the distribution of the k-th latent group, and π_k denotes its mixture proportion. Let P_{train} denote the clean training distribution before outlier contamination, which may have different mixing proportions from P_{test}. The observed training distribution follows Huber's ϵ-contamination model (Huber, 1992), $\widetilde{P}_{\text{train}} = (1-\epsilon)P_{\text{train}} + \epsilon\widetilde{P}$, where $\widetilde{P}$ is an arbitrary outlier distribution and $\epsilon \in [0, 1)$ denotes the contamination level. Let $\mathcal{D} = \{z_i\}_{i=1}^n$ be a dataset drawn i.i.d. from $\widetilde{P}_{\text{train}}$.

Our goal is to learn a model that is robust to both outlier contamination in the training data and subpopulation shift under the latent mixture distribution P_{test}. A classical baseline is empirical risk minimization (ERM), which minimizes the average loss over the observed training data. However,

ERM does not account for distributional shift between the clean training distribution P_{train} and the target distribution P_{test}, or for outlier contamination in the observed training distribution $\widetilde{P}_{\text{train}}$.

A promising alternative is distributionally robust learning, which minimizes the worst-case risk over a set of plausible distributions. Specifically, it solves

$$\min_{\theta} \max_{Q \in \mathcal{B}_\rho(\bar{P})} \mathbb{E}_Q\big[\ell(Z,\theta)\big],$$

where $\bar{P}$ is a nominal distribution and $\mathcal{B}_\rho(\bar{P})$ is an ambiguity set that captures distributional uncertainty around $\bar{P}$. Two prominent DRO formulations for subpopulation shift are Group DRO (Sagawa et al., 2020) and f-DRO (Namkoong & Duchi, 2016; Duchi & Namkoong, 2021). Group DRO directly optimizes the worst-case group risk but requires access to group labels, which are unavailable in many practical settings, including ours. In contrast, f-DRO defines the ambiguity set via an f-divergence and achieves robustness by reweighting high-loss samples, which can effectively address latent subpopulation shift. However, it remains inherently vulnerable to outliers (Zhai et al., 2021). Therefore, it may fail to provide robustness when both subpopulation shift and outlier contamination are present.

4 Proposed Method

We first introduce the general coupled DRO formulation and illustrate its structure with additive losses. We then extend the framework to non-decomposable survival losses, the main focus of this work.

4.1 Formulation

To motivate the proposed formulation, we first define a mixture ambiguity set. For a given nominal distribution $\bar{P}$ and a constant $\gamma \in (0,1]$, define

$$\mathcal{Q}_\gamma(\bar{P}) = \left\{ Q : \bar{P} = \gamma Q + (1-\gamma)\widetilde{Q} \text{ for some } \widetilde{Q} \in \mathcal{P}(\mathcal{Z}) \right\}, \tag{1}$$

where $\mathcal{P}(\mathcal{Z})$ denotes the set of all probability distributions on $\mathcal{Z}$. This set consists of γ-mixture components of $\bar{P}$, each of which is absolutely continuous with respect to $\bar{P}$, denoted by $Q \ll \bar{P}$.

This ambiguity set serves different roles in minimization and maximization. For a fixed θ, consider

$$\min_{Q \in \mathcal{Q}_{1-\epsilon}(\bar{P})} \mathbb{E}_Q[\ell(Z,\theta)] \quad \text{and} \quad \max_{Q \in \mathcal{Q}_\alpha(\bar{P})} \mathbb{E}_Q[\ell(Z,\theta)]. \tag{2}$$

Here, $\epsilon \in [0,1)$ and $\alpha \in (0,1]$. The minimization selects a distribution $Q \in \mathcal{Q}_{1-\epsilon}(\bar{P})$ that minimizes the expected loss by assigning smaller weights to high-loss regions, including potential outliers. The resulting distribution can be viewed as a refined version of $\bar{P}$, with ϵ reflecting the degree of potential outlier contamination. In contrast, the maximization selects a distribution $Q \in \mathcal{Q}_\alpha(\bar{P})$ that maximizes the expected loss by assigning larger weights to high-loss regions. This distribution emphasizes regions that are particularly difficult under the current model, with α controlling the size of the subpopulation considered. This maximization is closely related to CVaR-type DRO (Shapiro et al., 2021), which focuses on the high-loss tail of the distribution.

Motivated by these two roles, we propose the following coupled DRO problem:

$$\min_{\theta} \left[\min_{Q' \in \mathcal{Q}_{1-\epsilon}(\bar{P})} \max_{Q \in \mathcal{Q}_\alpha(Q')} \mathbb{E}_Q\big[\ell(Z,\theta)\big] \right]. \tag{3}$$

We refer to the minimization over Q' and the maximization over Q as the outer minimization and the inner maximization, respectively. The formulation couples nominal refinement with adversarial reweighting through the dependence of the inner ambiguity set $\mathcal{Q}_\alpha(Q')$ on the refined distribution Q'.

4.2 Structural analysis under additive losses

To facilitate understanding of the proposed formulation, we consider a class of losses for which the optimization structure becomes explicit. Specifically, given a weight vector $q \in \mathcal{S}_{n-1}$ over the

dataset $\mathcal{D}$, we view the empirical risk as a function $\mathcal{L} : \mathcal{S}_{n-1} \times \Theta \to \mathbb{R}$. We call this class of losses additive if the empirical risk can be decomposed as $\mathcal{L}(q, \theta) = \sum_{i=1}^{n} q_i \ell(z_i, \theta)$, where each loss term depends on a single sample. This class includes standard supervised learning losses such as the cross-entropy loss and the squared loss. We now take the nominal distribution $\bar{P}$ to be supported on $\mathcal{D}$ with probability mass vector $\bar{p} \in \mathcal{S}_{n-1}$. Under this choice, each feasible distribution can be represented by its probability mass vector over $\mathcal{D}$. Thus, for additive losses, the proposed formulation in (3) admits the following equivalent discrete representation:

$$\min_{\theta} \left[\min_{q' \in \mathcal{Q}_{1-\epsilon}(\bar{p})} \max_{q \in \mathcal{Q}_{\alpha}(q')} \mathcal{L}(q, \theta) \right]. \tag{4}$$

Here, for a reference weight vector $p \in \mathcal{S}_{n-1}$ and $\gamma \in (0, 1]$, the ambiguity set $\mathcal{Q}_{\gamma}(p)$ is defined as

$$\mathcal{Q}_{\gamma}(p) = \left\{ q \in \mathcal{S}_{n-1} : 0 \leq q_i \leq \frac{p_i}{\gamma},\ \forall i \in [n] \right\}. \tag{5}$$

Details of this finite-support representation of the ambiguity set are provided in Lemma 1.

The discrete coupled formulation in (4) admits an explicit two-stage characterization. Specifically, the outer minimization performs ϵ-trimming, and the inner maximization reweights the highest-loss α-fraction of the remaining probability mass. A precise characterization and the corresponding procedure are given in Appendix A.1.1 and Algorithm 2. Thus, our formulation includes CVaR-DORO (Zhai et al., 2021) as a special case for additive losses.

However, this structural result relies critically on the additivity of the loss and does not extend to settings where the loss depends on multiple samples simultaneously. In this paper, we extend the proposed framework beyond the additive setting to handle such non-decomposable losses. Survival analysis provides a primary example of this class, which we study in the following section.

4.3 Non-decomposable survival losses

In this section, we focus on right-censored survival data and specify the corresponding survival loss, although the framework can be adapted to other censoring or truncation mechanisms. Let $Z = (X, Y, \Delta)$ be a random variable, where $X \in \mathcal{X}$ denotes the covariate, T is the survival time, and C is the censoring time. The observed time is given by $Y = \min\{T, C\}$, and the event indicator is $\Delta = \mathbb{I}(T \leq C)$. Each observed sample is denoted by $z_i = (x_i, y_i, \delta_i)$.

Survival losses such as the Cox negative partial log-likelihood and ranking-based losses involve cross-sample dependence. To apply the proposed coupled DRO formulation, we express the empirical survival risk as $\mathcal{L}(q, \theta)$ to make its dependence on the sample weights explicit. As a representative example, we consider the Cox proportional hazards model (Cox, 1972). Given $q \in \mathcal{S}_{n-1}$, the empirical risk is defined as the weighted negative partial log-likelihood:

$$\mathcal{L}(q, \theta) = -\sum_{i=1}^{n} q_i \delta_i \left[f_\theta(x_i) - \log \left(\sum_{j \in [n]: y_j \geq y_i} q_j \exp(f_\theta(x_j)) \right) \right].$$

We incorporate this loss into the coupled formulation in (4). Because the sample weights also appear inside the logarithmic risk-set terms, $\mathcal{L}(q, \theta)$ is nonlinear in q. Thus, the closed-form trimming and reweighting structure derived for additive losses does not carry over directly. Moreover, standard DRO formulations based on dual reformulations are difficult to apply directly to non-decomposable survival losses, since they typically rely on additive losses (Hu & Chen, 2024; Jin et al., 2025). We therefore develop a gradient-based optimization scheme that preserves the original survival loss structure with cross-sample dependence.

5 Algorithm

In this section, we present the learning algorithm and its intuition. Throughout this section, we assume that $\mathcal{L}(q, \theta)$ is continuously differentiable in both q and θ, which enables gradient-based updates. Our approach treats the optimization as a three-block problem over (q, q', θ) and adopts an alternating update strategy. A key challenge lies in updating q', because the outer objective is defined implicitly through the inner maximization $\Phi_{\alpha}(q'; \theta) := \max_{q \in \mathcal{Q}_{\alpha}(q')} \mathcal{L}(q, \theta)$. Thus, its gradient with

Algorithm 1 Learning algorithm for the coupled DRO objective

Input: dataset $\mathcal{D}$; step sizes $\eta_q, \eta_{q'}, \eta_\theta$; update steps T, K; initial values $\theta^{(0)}, q^{(0)}, q'^{(0)}$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  Set $q^{(k,0)} \leftarrow q^{(k)}$, $q'^{(k,0)} \leftarrow q'^{(k)}$ ▷ Update (q, q')
  for $t = 0, \ldots, T-1$ **do**
    Update $q^{(k,t+1)}$ using the exponentiated gradient ascent and projection in (6)
    Update $q'^{(k,t+1)}$ using the exponentiated gradient descent and projection in (7)
  end for
  Set $q^{(k+1)} \leftarrow q^{(k,T)}$, $q'^{(k+1)} \leftarrow q'^{(k,T)}$ ▷ Update θ
  Update $\theta^{(k+1)}$ using the gradient descent step in (8)
end for

respect to q' is not directly available. The following proposition characterizes a supergradient through the KKT multipliers of the inner problem.

Proposition 1 (Outer supergradient via KKT multipliers)**.** *Fix θ and let $q^\star$ be a global maximizer of the inner problem. Under the regularity and local stability assumptions stated in Appendix A.1.2, any KKT multiplier vector $\mu^\star$ for $q_i \le q'_i/\alpha$ at $q^\star$ yields a supergradient $\mu^\star/\alpha$ of $\Phi_\alpha(\cdot;\theta)$ at q':*

$$\Phi_\alpha(q'+h;\theta) \le \Phi_\alpha(q';\theta) + \left\langle \frac{\mu^\star}{\alpha}, h \right\rangle + o(\|h\|).$$

Here, $q'+h \in \mathcal{S}_{n-1}$ and $o(\|h\|)/\|h\| \to 0$ as $h \to 0$.

Based on Proposition 1, we can derive a supergradient with respect to q' from the KKT multipliers $\mu^\star$ of the inner problem. The detailed characterization of these multipliers and the derivation of this result are provided in Appendix A.1.2.

Combining the q' update based on this supergradient with gradient steps for q and θ yields the alternating procedure in Algorithm 1. We detail the full-batch updates below for clarity; stochastic mini-batch implementations are also possible (Kvamme et al., 2019; Zeng et al., 2026).

Update of q. Given θ and q', we update q to $\hat{q}$ using the projected exponentiated gradient ascent update (Sagawa et al., 2020)

$$\begin{aligned} \tilde{q}_i &\leftarrow \frac{q_i \exp\left(\eta_q [\nabla_q \mathcal{L}(q,\theta)]_i\right)}{\sum_{j=1}^n q_j \exp\left(\eta_q [\nabla_q \mathcal{L}(q,\theta)]_j\right)}, \quad i = 1, \ldots, n, \\ \hat{q} &\leftarrow \Pi_{\mathcal{Q}_\alpha(q')}(\tilde{q}), \end{aligned} \tag{6}$$

where $\nabla_q \mathcal{L}(q,\theta)$ denotes the gradient of the map $q \mapsto \mathcal{L}(q,\theta)$, $\eta_q > 0$ is a step size, $\Pi_{\mathcal{A}}(\cdot)$ denotes projection onto the set $\mathcal{A}$, and $[x]_j$ denotes the j-th component of a vector x.

Update of q'. Given θ and $\hat{q}$, we update q' to $\hat{q}'$ using projected exponentiated descent with a supergradient estimate s:

$$\begin{aligned} \tilde{q}'_i &\leftarrow \frac{q'_i \exp\left(-\eta_{q'} s_i\right)}{\sum_{j=1}^n q'_j \exp\left(-\eta_{q'} s_j\right)}, \quad i = 1, \ldots, n, \\ \hat{q}' &\leftarrow \Pi_{\mathcal{Q}_{1-\epsilon}(\bar{p})}(\tilde{q}'), \end{aligned} \tag{7}$$

where $\eta_{q'} > 0$ is a step size. For a nonempty interior set $\mathcal{I}$, we compute s from $\hat{q}$ as follows:

$$\mathcal{I} := \left\{ i \in [n] : 0 < \hat{q}_i < \tfrac{1}{\alpha} q'_i \right\}, \quad \nu := \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} [\nabla_q \mathcal{L}(\hat{q},\theta)]_i,$$

$$s_i = \begin{cases} \frac{1}{\alpha}\left([\nabla_q \mathcal{L}(\hat{q},\theta)]_i - \nu\right)_+, & \text{if } 0 < \hat{q}_i = \frac{1}{\alpha} q'_i, \\ 0, & \text{otherwise.} \end{cases}$$

If $\mathcal{I} = \emptyset$, we leave q' unchanged, as in Algorithm 3.

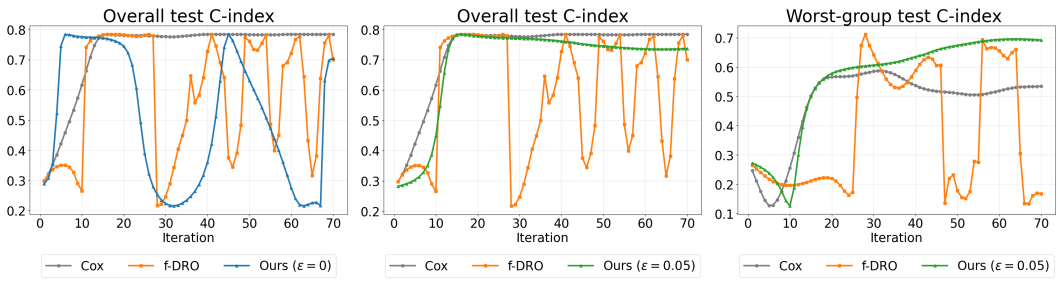

Figure 1: Test C-index trajectories under subpopulation shift with outlier contamination. (Left) Overall test C-index for the Cox model, f-DRO, and the proposed method with $(\alpha, \epsilon) = (0.1, 0)$. (Middle) Overall test C-index for the proposed method with $(\alpha, \epsilon) = (0.1, 0.05)$. (Right) Worst-group test C-index for the proposed method with $(\alpha, \epsilon) = (0.1, 0.05)$. For the proposed method, only ϵ is varied from 0 to 0.05, while all other training configurations are kept fixed.

Update of θ. After alternating the updates of q and q' for several iterations, we update θ to $\hat{\theta}$ using the resulting $\hat{q}$ via a gradient descent step.

$$\hat{\theta} \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}(\hat{q}, \theta), \tag{8}$$

where $\eta_\theta > 0$ is a step size.

To interpret the overall procedure, recall the finite-support representation of our ambiguity set:

$$\mathcal{Q}_{1-\epsilon}(\bar{p}) = \left\{ q' \in \mathcal{S}_{n-1} : 0 \le q'_i \le \frac{\bar{p}_i}{1-\epsilon},\ \forall i \in [n] \right\},$$
$$\mathcal{Q}_{\alpha}(q') = \left\{ q \in \mathcal{S}_{n-1} : 0 \le q_i \le \frac{q'_i}{\alpha},\ \forall i \in [n] \right\}.$$

The bound q'_i/α determines the maximum weight that can be assigned to sample i. Within these bounds, the inner update uses the gradient of the full survival loss to increase the loss. The outer update uses KKT multipliers from the inner problem to update q', thereby adjusting the upper bounds q'_i/α on the sample weights. Both updates account for how changes in sample weights affect the full loss, including other samples' risk-set terms in the Cox model. Thus, q and q' can be updated while preserving the cross-sample interactions in the survival loss.

Computational complexity is analyzed in Appendix A.1.3. A local convergence analysis is provided in Appendix A.1.4.

6 Experiments

6.1 Simulation

We evaluate robustness under subpopulation shift with outlier contamination; results without outliers are provided in Appendix A.3.1. For f-DRO, we use the χ^2-divergence and sample splitting following Hu & Chen (2024).

The training distribution mixes two groups with an outlier component: $\widetilde{P}_{\text{train}} = \pi_1^{\text{tr}} P_1 + \pi_2^{\text{tr}} P_2 + \tilde{\pi}^{\text{tr}} \widetilde{P}$, with $(\pi_1^{\text{tr}}, \pi_2^{\text{tr}}, \tilde{\pi}^{\text{tr}}) = (0.9, 0.08, 0.02)$ and $N_{\text{train}} = 3{,}000$. The test distribution is clean and is given by $P_{\text{test}} = \pi_1^{\text{te}} P_1 + \pi_2^{\text{te}} P_2$, with $(\pi_1^{\text{te}}, \pi_2^{\text{te}}) = (0.8, 0.2)$ and $N_{\text{test}} = 5{,}000$. For each group $g \in \{1, 2\}$, covariates are generated as $X \sim \mathcal{N}(0, 2I_2)$, and survival times follow an exponential hazard model $h_g(t \mid x) = h_0 \exp(x^\top \beta_g)$. Censoring times are drawn from an exponential distribution, yielding 10% censoring. We set $\beta_1 = (1, 1)$ and $\beta_2 = (1, -1)$ for the two groups, and use $\tilde{\beta} = (-2, -2)$ for the outlier component.

Figure 1 compares overall and worst-group test C-index trajectories, where the latter is the minimum across the two groups (Sagawa et al., 2020). Both f-DRO and the proposed method with $(\alpha, \epsilon) = (0.1, 0)$ exhibit large fluctuations, consistent with their emphasis on high-loss observations under contamination. The standard Cox model remains relatively stable but has a worst-group C-index of approximately 0.55. Setting $\epsilon = 0.05$ stabilizes the proposed method's overall and worst-group

Table 1: Results on METABRIC under subpopulation shift with outlier contamination. The 10% and 5% settings specify the validation tail loss for Robust-Cox and the DRO-based methods; Elastic-Cox and Cox-NN use validation log-likelihood. Best results are in bold, and second-best results are underlined. Values are mean (standard deviation) over 10 independent runs.

Model	Method	IBS ↓	Test log-likelihood ↑	Overall C-index ↑	Worst-group C-index ↑	Test tail loss ↓
Linear (10%)	Elastic-Cox	0.176 (0.006)	-5.061 (0.035)	0.618 (0.022)	0.501 (0.058)	6.305 (0.111)
	Robust-Cox	**0.167** (0.005)	**-5.016** (0.028)	0.665 (0.017)	0.562 (0.042)	6.229 (0.048)
	f-DRO	0.183 (0.005)	-5.121 (0.033)	0.667 (0.015)	0.574 (0.030)	**5.825** (0.015)
	Exact-Cox	0.184 (0.006)	-5.121 (0.035)	0.583 (0.028)	0.459 (0.057)	6.179 (0.164)
	Ours	0.184 (0.005)	-5.125 (0.034)	**0.684** (0.015)	**0.613** (0.028)	5.835 (0.011)
Linear (5%)	Elastic-Cox	0.176 (0.006)	-5.061 (0.035)	0.618 (0.022)	0.501 (0.058)	6.305 (0.111)
	Robust-Cox	**0.168** (0.006)	**-5.020** (0.026)	0.658 (0.022)	0.547 (0.049)	6.240 (0.056)
	f-DRO	0.186 (0.003)	-5.141 (0.018)	0.652 (0.033)	0.558 (0.046)	**5.824** (0.014)
	Exact-Cox	0.184 (0.006)	-5.121 (0.035)	0.583 (0.028)	0.459 (0.057)	6.179 (0.164)
	Ours	0.187 (0.004)	-5.145 (0.026)	**0.680** (0.015)	**0.605** (0.030)	5.831 (0.007)
Nonlinear (10%)	Cox-NN	0.181 (0.006)	-5.101 (0.050)	0.601 (0.030)	0.508 (0.091)	6.561 (0.310)
	Robust-Cox	**0.175** (0.005)	**-5.063** (0.032)	0.635 (0.028)	0.545 (0.049)	6.295 (0.161)
	f-DRO	0.182 (0.004)	-5.119 (0.027)	0.644 (0.036)	0.533 (0.047)	**5.839** (0.041)
	Exact-Cox	0.197 (0.007)	-5.210 (0.049)	0.456 (0.091)	0.395 (0.078)	6.226 (0.178)
	Ours	0.176 (0.004)	-5.074 (0.027)	**0.661** (0.020)	**0.585** (0.040)	5.902 (0.014)
Nonlinear (5%)	Cox-NN	0.181 (0.006)	-5.101 (0.050)	0.601 (0.030)	0.508 (0.091)	6.561 (0.310)
	Robust-Cox	**0.175** (0.005)	**-5.063** (0.032)	0.635 (0.028)	0.545 (0.049)	6.295 (0.161)
	f-DRO	0.186 (0.004)	-5.147 (0.029)	0.627 (0.036)	0.511 (0.051)	**5.834** (0.021)
	Exact-Cox	0.197 (0.007)	-5.210 (0.049)	0.456 (0.091)	0.395 (0.078)	6.226 (0.178)
	Ours	0.177 (0.005)	-5.079 (0.033)	**0.656** (0.017)	**0.570** (0.036)	5.901 (0.026)

trajectories and yields a worst-group C-index near 0.70, illustrating the benefit of controlling outlier influence while protecting minority-group performance.

Further simulation analyses, including nonlinear risk functions, higher-dimensional covariates, smaller training samples, heavier censoring, and IBS trajectories, are provided in Appendices A.3.2–A.3.4. Sensitivity to the robustness parameters α and ϵ is examined in Appendix A.2.1.

6.2 Real data analysis

Baselines. We compare the proposed method with standard Cox models, f-DRO and its exact Cox variant (Exact-Cox) for latent subpopulation shift, and Robust-Cox (Minder & Bednarski, 1996) for outlier contamination. Both f-DRO and Exact-Cox use the χ^2-divergence following Hu & Chen (2024); f-DRO uses sample splitting, whereas Exact-Cox uses the full likelihood without sample splitting. For nonlinear Robust-Cox, we extend the original estimator to a neural predictor.

Datasets. We evaluate METABRIC (Curtis et al., 2012) and FLC (Dispenzieri et al., 2012), defining four subgroups in each dataset based on clinical covariates following Hu & Chen (2024). Group labels are unavailable during training and validation. Further details, including sample sizes, censoring rates, and subgroup definitions, are provided in Appendix A.3.5.

Experimental Setup. Although the framework applies to general survival models, we consider linear Cox models and neural Cox models (Cox-NN) with a single hidden layer of $h = 2$ units. Motivated by Duchi & Namkoong (2021), we select hyperparameters for Robust-Cox and the DRO-based methods by minimizing validation tail loss without using group labels. We evaluate overall performance using the integrated Brier score (IBS), test log-likelihood, and C-index, and robustness using worst-group C-index and 5% test tail loss. Details of the datasets, data splits, baselines, training, and hyperparameter selection are provided in Appendices A.3.5–A.3.7.

Outlier Construction. We evaluate subpopulation shift with outlier contamination; results without outliers appear in Appendix A.3.8. We generate new survival times by inverting the fitted Cox risks of the highest-risk training samples: the top 2% for linear models and 5% for nonlinear models. These settings are motivated by survival outliers whose event times are unusually long or short relative to predictions based on their covariates (Luo et al., 2022). We use risk inversion to introduce such departures in a small fraction of the training samples, with effects that may vary across datasets and model classes. Construction details and additional contamination experiments are provided in Appendices A.3.5, A.3.9, and A.3.10.

Results under outlier contamination. On METABRIC (Table 1), the proposed method outperforms all baselines in both overall and worst-group C-index in every setting, while achieving the

Table 2: Results on FLC under subpopulation shift with outlier contamination. Settings and notation follow Table 1.

Model	Method	IBS ↓	Test log-likelihood ↑	Overall C-index ↑	Worst-group C-index ↑	Test tail loss ↓
Linear (10%)	Elastic-Cox	0.100 (0.003)	-6.573 (0.066)	0.793 (0.003)	0.618 (0.038)	7.650 (0.114)
	Robust-Cox	**0.096** (0.001)	**-6.491** (0.016)	0.793 (0.005)	0.611 (0.040)	7.854 (0.031)
	f-DRO	0.107 (0.003)	-6.708 (0.031)	0.797 (0.004)	0.626 (0.037)	7.409 (0.035)
	Exact-Cox	0.103 (0.005)	-6.634 (0.063)	0.789 (0.006)	0.609 (0.037)	7.581 (0.127)
	Ours	0.119 (0.002)	-6.941 (0.009)	**0.798** (0.004)	**0.631** (0.036)	**7.283** (0.006)
Linear (5%)	Elastic-Cox	0.100 (0.003)	-6.573 (0.066)	0.793 (0.003)	0.618 (0.038)	7.650 (0.114)
	Robust-Cox	**0.096** (0.001)	**-6.491** (0.016)	0.793 (0.005)	0.611 (0.040)	7.854 (0.031)
	f-DRO	0.125 (0.002)	-7.052 (0.008)	0.795 (0.005)	0.626 (0.030)	7.294 (0.004)
	Exact-Cox	0.109 (0.010)	-6.747 (0.192)	0.722 (0.140)	0.576 (0.085)	7.535 (0.072)
	Ours	0.119 (0.002)	-6.941 (0.009)	**0.798** (0.004)	**0.631** (0.036)	**7.283** (0.006)
Nonlinear (10%)	Cox-NN	**0.104** (0.006)	**-6.617** (0.065)	0.782 (0.010)	0.623 (0.029)	7.801 (0.208)
	Robust-Cox	0.106 (0.007)	-6.690 (0.126)	0.760 (0.031)	0.602 (0.030)	8.134 (0.324)
	f-DRO	0.111 (0.004)	-6.725 (0.033)	**0.785** (0.009)	0.624 (0.028)	7.447 (0.043)
	Exact-Cox	0.112 (0.008)	-6.783 (0.155)	0.723 (0.125)	0.561 (0.081)	7.622 (0.188)
	Ours	0.116 (0.003)	-6.837 (0.057)	0.777 (0.019)	**0.630** (0.033)	**7.371** (0.067)
Nonlinear (5%)	Cox-NN	**0.104** (0.006)	**-6.617** (0.065)	**0.782** (0.010)	0.623 (0.029)	7.801 (0.208)
	Robust-Cox	0.106 (0.007)	-6.687 (0.129)	0.761 (0.032)	0.603 (0.031)	8.134 (0.324)
	f-DRO	0.126 (0.002)	-7.055 (0.009)	0.779 (0.013)	0.620 (0.023)	7.298 (0.004)
	Exact-Cox	0.119 (0.012)	-6.948 (0.214)	0.625 (0.181)	0.508 (0.117)	7.579 (0.207)
	Ours	0.123 (0.004)	-6.976 (0.043)	0.781 (0.017)	**0.628** (0.041)	**7.284** (0.008)

Table 3: Performance across hidden-layer sizes h on METABRIC under outlier contamination. Settings and notation follow the 10% validation setting in Table 1, whose nonlinear block corresponds to $h = 2$.

h	Method	IBS ↓	Test log-likelihood ↑	Overall C-index ↑	Worst-group C-index ↑	Test tail loss ↓
4	Cox-NN	0.183 (0.007)	-5.118 (0.047)	0.597 (0.012)	0.474 (0.073)	6.658 (0.188)
	Robust-Cox	0.183 (0.009)	-5.139 (0.062)	0.612 (0.019)	0.506 (0.065)	6.914 (0.322)
	f-DRO	0.185 (0.004)	-5.135 (0.021)	0.627 (0.028)	0.519 (0.036)	**5.832** (0.021)
	Exact-Cox	0.199 (0.006)	-5.225 (0.036)	0.452 (0.060)	0.391 (0.054)	6.245 (0.079)
	Ours	**0.177** (0.004)	**-5.077** (0.023)	**0.646** (0.027)	**0.546** (0.042)	5.946 (0.045)
8	Cox-NN	0.189 (0.005)	-5.184 (0.068)	0.581 (0.029)	0.482 (0.043)	6.600 (0.662)
	Robust-Cox	0.186 (0.007)	-5.179 (0.066)	0.611 (0.017)	0.504 (0.061)	7.106 (0.295)
	f-DRO	0.185 (0.005)	-5.138 (0.037)	0.611 (0.012)	0.492 (0.055)	**5.870** (0.053)
	Exact-Cox	0.201 (0.011)	-5.243 (0.070)	0.439 (0.072)	0.387 (0.073)	6.344 (0.178)
	Ours	**0.175** (0.005)	**-5.068** (0.037)	**0.654** (0.027)	**0.582** (0.045)	5.971 (0.084)
16	Cox-NN	0.191 (0.006)	-5.191 (0.070)	0.571 (0.032)	0.467 (0.074)	6.313 (0.731)
	Robust-Cox	0.190 (0.009)	-5.231 (0.076)	0.608 (0.021)	0.477 (0.051)	7.320 (0.374)
	f-DRO	0.186 (0.004)	-5.140 (0.028)	0.600 (0.020)	0.494 (0.056)	**5.888** (0.049)
	Exact-Cox	0.201 (0.010)	-5.250 (0.062)	0.465 (0.064)	0.394 (0.045)	6.367 (0.207)
	Ours	**0.175** (0.003)	**-5.065** (0.028)	**0.653** (0.017)	**0.568** (0.057)	5.980 (0.089)

second-lowest test tail loss. On FLC (Table 2), the results show a trade-off between average predictive performance and robustness: the proposed method achieves the highest worst-group C-index and the lowest test tail loss in every setting, but has higher IBS and lower test log-likelihood than the standard Cox models and Robust-Cox. Additionally, we conduct experiments with appended outliers on FLC using linear models and with 2% and 5% contamination across all dataset–model combinations (Appendices A.3.9 and A.3.10). Across these experiments, the proposed method consistently achieves the lowest or second-lowest test tail loss. Runtime comparisons are provided in Appendix A.3.11.

Ablation across network configurations. We evaluate larger networks on METABRIC using hidden-layer sizes $h = 4, 8, 16$ under the same contamination and model-selection settings. As shown in Table 3, the proposed method achieves the highest overall and worst-group C-index, the lowest IBS, and the highest test log-likelihood in all three configurations. Its test tail loss is consistently second lowest, with f-DRO attaining the lowest values. These results show that the gains in overall and worst-group discrimination persist across the larger networks evaluated here.

7 Conclusion

We proposed a coupled DRO framework for survival analysis that jointly addresses subpopulation shift and outlier contamination. The formulation couples nominal refinement with adversarial reweighting while preserving cross-sample interactions in non-decomposable survival losses. We derived outer updates from the inner problem's KKT multipliers, yielding an alternating algorithm that directly optimizes the coupled objective. Experiments demonstrate improved worst-group C-index and consistently low test tail loss under contamination, with competitive overall C-index. Future work could develop ambiguity sets that also capture within-group covariate shifts while retaining outlier robustness.

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and editing and implementation details. The authors reviewed all AI-assisted content and take responsibility for the final manuscript.

REFERENCES

Mathieu Andreux, Andre Manoel, Romuald Menuet, Charlie Saillard, and Chloé Simpson. Federated survival analysis with discrete-time Cox models. *arXiv:2006.08997*, 2020.

Alberto Archetti and Matteo Matteucci. Federated survival forests. In *International Joint Conference on Neural Networks*, pp. 1–9, 2023.

Alberto Archetti, Eugenio Lomurno, Francesco Lattari, André Martin, and Matteo Matteucci. Heterogeneous datasets for federated survival analysis simulation. In *International Conference on Performance Engineering Companion*, pp. 173–180, 2023.

Anand Avati, Tony Duan, Sharon Zhou, Kenneth Jung, Nigam H. Shah, and Andrew Y. Ng. Countdown regression: Sharp and calibrated survival predictions. In *Conference on Uncertainty in Artificial Intelligence*, pp. 145–155, 2020.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Steve Bennett. Analysis of survival data by the proportional odds model. *Statistics in Medicine*, 2(2): 273–277, 1983.

Dimitri P. Bertsekas. On the Goldstein–Levitin–Polyak gradient projection method. *IEEE Transactions on Automatic Control*, 21(2):174–184, 1976.

Jonathan Buckley and Ian James. Linear regression with censored data. *Biometrika*, 66(3):429–436, 1979.

David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403): 346–352, 2012.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

Angela Dispenzieri, Jerry A. Katzmann, Robert A. Kyle, Dirk R. Larson, Terry M. Therneau, Colin L. Colby, Raynell J. Clark, Graham P. Mead, Shaji Kumar, L. Joseph Melton, III, et al. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. *Mayo Clinic Proceedings*, 87(6):517–523, 2012.

Hyungrok Do, Yuxin Chang, Yoon Sang Cho, Padhraic Smyth, and Judy Zhong. Fair survival time prediction via mutual information minimization. In *Machine Learning for Healthcare Conference*, pp. 128–149, 2023.

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

Shaohua Fan, Renzhe Xu, Qian Dong, Yue He, Cheng Chang, and Peng Cui. Stable Cox regression for survival analysis under distribution shifts. *Nature Machine Intelligence*, 6(12):1525–1541, 2024.

Alessio Farcomeni and Laura Ventura. An overview of robust methods in medical research. *Statistical Methods in Medical Research*, 21(2):111–133, 2012.

Alessio Farcomeni and Sara Viviani. Robust estimation for the Cox regression model based on trimming. *Biometrical Journal*, 53(6):956–973, 2011.

Rui Gao, Xi Chen, and Anton J. Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 72(3):1177–1191, 2024.

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

Stephane Heritier, Eva Cantoni, Samuel Copt, and Maria-Pia Victoria-Feser. *Robust methods in biostatistics*. John Wiley & Sons, 2009.

Feng Hong, Jiangchao Yao, Yueming Lyu, Zhihan Zhou, Ivor W. Tsang, Ya Zhang, and Yanfeng Wang. On harmonizing implicit subpopulations. In *International Conference on Learning Representations*, 2024.

Shu Hu and George H. Chen. Distributionally robust survival analysis: A novel fairness loss without demographics. In *Machine Learning for Health Symposium*, pp. 62–87, 2022.

Shu Hu and George H. Chen. Fairness in survival analysis with distributionally robust optimization. *Journal of Machine Learning Research*, 25(246):1–85, 2024.

Jinyi Huang, Jinlong Lei, and Guodong Shi. Robust learning via nested distributionally robust optimization. In *International Conference on Machine Learning*, 2026.

Peter J. Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

Yeping Jin, Lauren Wise, and Ioannis Ch. Paschalidis. Distributionally robust learning in survival analysis. In *Conference on Health, Inference, and Learning*, pp. 369–380, 2025.

Sung Ho Jo, Seonghwi Kim, and Minwoo Chae. Mitigating spurious correlation via distributionally robust learning with hierarchical ambiguity sets. In *International Conference on Learning Representations*, 2026.

John D. Kalbfleisch and Ross L. Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2nd edition, 2002.

Edward L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18:24, 2018.

Kamrun Naher Keya, Rashidul Islam, Shimei Pan, Ian Stockwell, and James Foulds. Equitable allocation of healthcare resources with fair survival models. In *SIAM International Conference on Data Mining*, pp. 190–198, 2021.

Seonghwi Kim, Sung Ho Jo, Wooseok Ha, and Minwoo Chae. Distributionally robust classification for multi-source unsupervised domain adaptation. In *International Conference on Learning Representations*, 2026.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pp. 130–166. INFORMS, 2019.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and Cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.

Changhee Lee, William R. Zame, Jinsung Yoon, and Mihaela van der Schaar. DeepHit: A deep learning approach to survival analysis with competing risks. In *AAAI Conference on Artificial Intelligence*, pp. 2314–2321, 2018.

Fengming Lin, Xiaolei Fang, and Zheming Gao. Distributionally robust optimization: A review on theory and applications. *Numerical Algebra, Control and Optimization*, 12(1):159–212, 2022.

Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792, 2021.

Zhaoqi Liu, Xiang-Sun Zhang, and Shihua Zhang. Breast tumor subgroups reveal diverse clinical prognostic power. *Scientific Reports*, 4:4002, 2014.

Bin Luo, Xiaoli Gao, and Susan Halabi. Penalized weighted proportional hazards model for robust variable selection and outlier detection. *Statistics in Medicine*, 41(17):3398–3420, 2022.

C. E. Minder and T. Bednarski. A robust method for proportional hazards regression. *Statistics in Medicine*, 15(10):1033–1047, 1996.

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*, 2016.

Md Mahmudur Rahman and Sanjay Purushotham. Fair and interpretable models for survival analysis. In *SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1452–1462, 2022.

Md Mahmudur Rahman and Sanjay Purushotham. FedPseudo: Privacy-preserving pseudo value-based deep learning models for federated survival analysis. In *SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1999–2009, 2023.

Md Mahmudur Rahman and Sanjay Purushotham. Fair federated survival analysis. In *AAAI Conference on Artificial Intelligence*, pp. 20104–20112, 2025.

Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. In *Machine Learning for Healthcare Conference*, pp. 101–114, 2016.

Vikas C. Raykar, Harald Steck, Balaji Krishnapuram, Cary Dehing-Oberije, and Philippe Lambin. On ranking in survival analysis: Bounds on the concordance index. In *Advances in Neural Information Processing Systems*, pp. 1209–1216, 2007.

R. Tyrrell Rockafellar and Roger J. B. Wets. *Variational analysis*. Springer, 1998.

Tim G. J. Rudner, Ya Shi Zhang, Andrew Gordon Wilson, and Julia Kempe. Mind the GAP: Improving robustness to subpopulation shifts with group-aware priors. In *International Conference on Artificial Intelligence and Statistics*, pp. 127–135, 2024.

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*, 2020.

Navid Seidi, Satyaki Roy, Sajal K. Das, and Ardhendu Tripathy. Addressing data heterogeneity in federated learning of Cox proportional hazards models. In *IEEE International Conference on E-health Networking, Application & Services*, 2024.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: Modeling and theory*. SIAM, 3rd edition, 2021.

Sanjoy K. Sinha. Robust estimation in accelerated failure time models. *Lifetime Data Analysis*, 25 (1):52–78, 2019.

Nathan Stromberg, Rohan Ayyagari, Monica Welfert, Sanmi Koyejo, Richard Nock, and Lalitha Sankar. Robustness to subpopulation shift with domain label noise via regularized annotation of domains. *Transactions on Machine Learning Research*, 2024.

Weijing Tang, Kevin He, Gongjun Xu, and Ji Zhu. Survival analysis via ordinary differential equations. *Journal of the American Statistical Association*, 118(544):2406–2421, 2023.

Xuan Wang, Harrison G. Zhang, Xin Xiong, Chuan Hong, Griffin M. Weber, Gabriel A. Brat, Clara-Lea Bonzel, Yuan Luo, Rui Duan, Nathan P. Palmer, et al. SurvMaximin: Robust federated approach to transporting survival risk prediction models. *Journal of Biomedical Informatics*, 134: 104176, 2022.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. In *International Conference on Machine Learning*, pp. 39584–39622, 2023.

Lang Zeng, Weijing Tang, Zhao Ren, and Ying Ding. Mini-batch estimation for deep Cox models: Statistical foundations and practical guidance. *Journal of the American Statistical Association*, 121 (554):988–999, 2026.

Yuchen Zeng, Kristjan Greenewald, Luann Jung, Kangwook Lee, Justin Solomon, and Mikhail Yurochkin. Outlier-robust group inference via gradient space clustering. In *NeurIPS Workshop on Distribution Shifts*, 2023.

Runtian Zhai, Chen Dan, Zico Kolter, and Pradeep Ravikumar. DORO: Distributional and outlier robust optimization. In *International Conference on Machine Learning*, pp. 12345–12355, 2021.

Wenbin Zhang and Jeremy C. Weiss. Longitudinal fairness with censorship. In *AAAI Conference on Artificial Intelligence*, pp. 12235–12243, 2022.

Wenbin Zhang and Jeremy C. Weiss. Fairness with censorship and group constraints. *Knowledge and Information Systems*, 65(6):2571–2594, 2023.

A APPENDIX

Contents of the appendix

A.1 Formulation and algorithm details . . . 15
A.1.1 Additive losses . . . 15
A.1.2 Non-decomposable losses . . . 21
A.1.3 Computational complexity . . . 25
A.1.4 Local convergence analysis . . . 25
A.1.5 Alternating gradient algorithm . . . 27
A.2 Sensitivity analysis . . . 28
A.2.1 Robustness parameters . . . 28
A.3 Additional experiments and experimental details . . . 28
A.3.1 Simulation: Subpopulation shift . . . 28
A.3.2 Simulation: Nonlinear risk functions . . . 29
A.3.3 Simulation: Higher dimension, smaller samples, and heavier censoring . . . 30
A.3.4 Simulation: Integrated Brier score trajectories . . . 31
A.3.5 Benchmark datasets and contamination setup . . . 32
A.3.6 Baselines . . . 33
A.3.7 Training and hyperparameter selection . . . 33
A.3.8 Real data analysis: Subpopulation shift . . . 34
A.3.9 Additional outlier setting . . . 35
A.3.10 Contamination sensitivity . . . 35
A.3.11 Computational runtime . . . 36

A.1 FORMULATION AND ALGORITHM DETAILS

For additive losses, $\mathcal{L}(q,\theta) = \sum_{i=1}^n q_i \ell(z_i,\theta)$ is linear in q, so the optimal weights follow the ordering of the losses. We first derive this trimming and reweighting structure, then use KKT multipliers to obtain gradient-based updates for non-decomposable losses such as the weighted Cox loss. We also analyze computational complexity and local convergence.

For fixed θ, recall the inner value function

$$\Phi_\alpha(q';\theta) := \max_{q \in \mathcal{Q}_\alpha(q')} \mathcal{L}(q,\theta). \tag{9}$$

We omit θ when it is fixed.

A.1.1 ADDITIVE LOSSES

Write $\ell_i := \ell(z_i,\theta)$ and $\Phi_\alpha(q') := \Phi_\alpha(q';\theta)$. We first express the ambiguity set as bounds on sample weights, then derive ϵ-trimming for $\alpha = 1$ and the two-stage solution for general α. The uniform-weight case gives the explicit allocation in Corollary 1.

Lemma 1 (Equivalent discrete representation of the ambiguity set)**.** *Let $\bar{P}$ be the nominal distribution supported on $\mathcal{D}$, with corresponding weight vector $\bar{p} \in \mathcal{S}_{n-1}$. For $\gamma \in (0,1]$, the set $\mathcal{Q}_\gamma(\bar{P})$ admits the equivalent finite-support representation* (5)*, i.e.,*

$$\mathcal{Q}_\gamma(\bar{p}) = \left\{ q \in \mathcal{S}_{n-1} : 0 \le q_i \le \frac{\bar{p}_i}{\gamma},\ \forall i \in [n] \right\}.$$

Moreover, every such q can be reparameterized as

$$q_i = \frac{\bar{p}_i}{\gamma} w_i, \qquad w \in [0,1]^n, \qquad \sum_{i=1}^n \bar{p}_i w_i = \gamma. \tag{10}$$

Proof. We first consider the case $\gamma \in (0,1)$.

($\Rightarrow$) Suppose that $Q \in \mathcal{Q}_\gamma(\bar{P})$. Since $\bar{P}$ is supported on the dataset $\mathcal{D} = \{z_i\}_{i=1}^n$, there exists a weight vector $\bar{p} \in \mathcal{S}_{n-1}$ such that

$$\bar{P} = \sum_{i=1}^n \bar{p}_i \delta_{z_i},$$

where δ_{z_i} denotes the Dirac measure at z_i. By the definition of $\mathcal{Q}_\gamma(\bar{P})$, there exists a probability distribution $\widetilde{Q} \in \mathcal{P}(\mathcal{Z})$ such that

$$\bar{P} = \gamma Q + (1-\gamma)\widetilde{Q}.$$

Since $\bar{P}$ is supported on $\mathcal{D}$, the above identity implies that Q and $\widetilde{Q}$ are also supported on $\mathcal{D}$. Hence there exist weight vectors $q, \tilde{q} \in \mathcal{S}_{n-1}$ such that

$$Q = \sum_{i=1}^n q_i \delta_{z_i}, \qquad \widetilde{Q} = \sum_{i=1}^n \tilde{q}_i \delta_{z_i}.$$

By comparing the coefficients of the Dirac measures at each z_i, we obtain

$$\bar{p} = \gamma q + (1-\gamma)\tilde{q}.$$

Taking the i-th coordinate and using $\tilde{q}_i \ge 0$, we have

$$\bar{p}_i = \gamma q_i + (1-\gamma)\tilde{q}_i \ge \gamma q_i,$$

which implies

$$q_i \le \frac{\bar{p}_i}{\gamma}.$$

Since $q \in \mathcal{S}_{n-1}$, we also have $q_i \ge 0$ for all i and $\sum_{i=1}^n q_i = 1$. Hence $q \in \mathcal{Q}_\gamma(\bar{p})$ in (5).

($\Leftarrow$) Conversely, suppose that $q \in \mathcal{S}_{n-1}$ satisfies

$$0 \le q_i \le \frac{\bar{p}_i}{\gamma}, \qquad \forall i \in [n].$$

Define a vector $\tilde{q} \in \mathbb{R}^n$ by

$$\tilde{q}_i := \frac{\bar{p}_i - \gamma q_i}{1-\gamma}, \qquad i \in [n].$$

Since $q_i \le \bar{p}_i/\gamma$, we have $\tilde{q}_i \ge 0$ for all $i \in [n]$. Moreover,

$$\sum_{i=1}^n \tilde{q}_i = \frac{1}{1-\gamma}\sum_{i=1}^n \big(\bar{p}_i - \gamma q_i\big) = \frac{1}{1-\gamma}\big(1-\gamma\cdot 1\big) = 1.$$

Hence $\tilde{q} \in \mathcal{S}_{n-1}$.

Now define probability distributions

$$Q = \sum_{i=1}^n q_i \delta_{z_i}, \qquad \widetilde{Q} = \sum_{i=1}^n \tilde{q}_i \delta_{z_i}.$$

Then

$$\gamma Q + (1-\gamma)\widetilde{Q} = \sum_{i=1}^n \big(\gamma q_i + (1-\gamma)\tilde{q}_i\big)\delta_{z_i} = \sum_{i=1}^n \bar{p}_i \delta_{z_i} = \bar{P}.$$

Therefore, $Q \in \mathcal{Q}_\gamma(\bar{P})$.

We now show the reparameterization in (10). Suppose that q satisfies (5). For each $i \in [n]$ with $\bar{p}_i > 0$, define

$$w_i := \frac{\gamma q_i}{\bar{p}_i}.$$

If $\bar{p}_i = 0$, then (5) implies $q_i = 0$, and we define $w_i := 0$.

Then $w_i \ge 0$ for all i. Moreover, from $q_i \le \bar{p}_i/\gamma$, it follows that $w_i \le 1$ whenever $\bar{p}_i > 0$. Hence $w \in [0,1]^n$. By construction,

$$q_i = \frac{\bar{p}_i}{\gamma} w_i, \qquad \forall i \in [n].$$

Finally,

$$\sum_{i=1}^n \bar{p}_i w_i = \sum_{i:\bar{p}_i>0} \bar{p}_i \frac{\gamma q_i}{\bar{p}_i} = \gamma \sum_{i=1}^n q_i = \gamma,$$

since $q \in \mathcal{S}_{n-1}$.

If $\gamma = 1$, then the condition

$$\bar{P} = Q$$

implies $q = \bar{p}$. The discrete constraint becomes $0 \le q_i \le \bar{p}_i$ with $q \in \mathcal{S}_{n-1}$, which also implies $q = \bar{p}$. The reparameterization holds with $w_i = 1$ for $\bar{p}_i > 0$ and $w_i = 0$ for $\bar{p}_i = 0$. This completes the proof. □

Using the representation in Lemma 1, we first consider the special case $\alpha = 1$. In this case, the inner ambiguity set $\mathcal{Q}_\alpha(q')$ reduces to the singleton $\{q'\}$, and the coupled DRO problem reduces to

$$\min_{q' \in \mathcal{Q}_{1-\epsilon}(\bar{p})} \sum_{i=1}^n q'_i \ell_i. \tag{11}$$

The minimizer assigns mass to smaller-loss samples first, yielding ϵ-trimming.

Proposition 2 (Exact ϵ-trimming under $\alpha = 1$)**.** *Assume $\alpha = 1$. Consider the problem*

$$q'^\star \in \operatorname*{argmin}_{q' \in \mathcal{Q}_{1-\epsilon}(\bar{p})} \sum_{i=1}^n q'_i \ell_i.$$

Then there exists an optimal solution of the form

$$q'^\star_i = \frac{\bar{p}_i w^\star_i}{1-\epsilon},$$

where $w^\star \in [0,1]^n$ *satisfies*

$$\sum_{i=1}^{n} \bar{p}_i w_i^\star = 1 - \epsilon,$$

and, after reindexing so that

$$\ell_1 \le \ell_2 \le \cdots \le \ell_n,$$

there exists $k \in [n]$ *such that*

$$w_i^\star = \begin{cases} 1, & i < k, \\ \gamma, & i = k, \\ 0, & i > k, \end{cases}$$

for some $\gamma \in [0,1]$ *chosen to satisfy the probability mass constraint* $\sum_{i=1}^{n} q_i'^* = 1$.

Proof. The argument transfers mass from larger to smaller losses without increasing the objective. By Lemma 1, any $q' \in \mathcal{Q}_{1-\epsilon}(\bar{p})$ admits the representation

$$q_i' = \frac{\bar{p}_i w_i}{1-\epsilon}, \qquad w \in [0,1]^n, \qquad \sum_{i=1}^{n} \bar{p}_i w_i = 1 - \epsilon.$$

Then, (11) reduces to

$$\min_{w \in [0,1]^n} \quad \sum_{i=1}^{n} \bar{p}_i w_i \ell_i \quad \text{s.t.} \quad \sum_{i=1}^{n} \bar{p}_i w_i = 1 - \epsilon.$$

Let $m_i := \bar{p}_i w_i$. The problem is equivalently

$$\min_{0 \le m_i \le \bar{p}_i} \sum_{i=1}^{n} \ell_i m_i \quad \text{s.t.} \quad \sum_{i=1}^{n} m_i = 1 - \epsilon.$$

The objective is linear in the allocated mass m_i with coefficient ℓ_i, while $\bar{p}_i$ acts as an upper bound. Reindex so that $\ell_1 \le \ell_2 \le \cdots \le \ell_n$. Suppose there exist $i < j$ such that $m_i < \bar{p}_i$ and $m_j > 0$. Then for any $0 < \delta \le \min\{\bar{p}_i - m_i,\ m_j\}$, the perturbation

$$m_i \leftarrow m_i + \delta, \qquad m_j \leftarrow m_j - \delta$$

preserves feasibility and changes the objective by $\delta(\ell_i - \ell_j) \le 0$, with strict inequality if $\ell_i < \ell_j$. Hence an optimal solution assigns mass in increasing loss order up to the bounds, including when losses are tied.

Thus there exists an index $k \in [n]$ such that

$$m_i^\star = \begin{cases} \bar{p}_i, & i < k, \\ \widetilde{m}_k, & i = k, \\ 0, & i > k, \end{cases}$$

for some $\widetilde{m}_k \in [0, \bar{p}_k]$ chosen to satisfy the probability mass constraint. Equivalently, there exist $k \in [n]$ and $\gamma \in [0,1]$ such that

$$w_i^\star = \begin{cases} 1, & i < k, \\ \gamma, & i = k, \\ 0, & i > k, \end{cases}$$

with $\sum_{i=1}^{n} \bar{p}_i w_i^\star = 1 - \epsilon$. This completes the proof. □

The next result extends this structure to the general case $\alpha \in (0,1]$ under $\epsilon \in [0,1)$. We first show that the inner maximization problem $\Phi_\alpha(q')$ admits a weighted CVaR representation.

Lemma 2 (Weighted CVaR dual representation of Φ_α)**.** *For any $q' \in \mathcal{S}_{n-1}$, the inner maximization problem $\Phi_\alpha(q')$ admits the dual representation*

$$\Phi_\alpha(q') = \min_{\tau \in \mathbb{R}} \left\{ \tau + \frac{1}{\alpha} \sum_{i=1}^{n} q'_i (\ell_i - \tau)_+ \right\},$$

where $(x)_+ := \max\{x, 0\}$.

Proof. Recall that the inner maximization problem is

$$\Phi_\alpha(q') = \max \left\{ \sum_{i=1}^{n} q_i \ell_i : q \in \mathcal{S}_{n-1},\ 0 \le q_i \le \frac{1}{\alpha} q'_i,\ \forall i \in [n] \right\}.$$

This is precisely the standard primal formulation of weighted CVaR under the reference weight vector q'. By the dual representation of CVaR (Shapiro et al., 2021), we obtain

$$\Phi_\alpha(q') = \min_{\tau \in \mathbb{R}} \left\{ \tau + \frac{1}{\alpha} \sum_{i=1}^{n} q'_i (\ell_i - \tau)_+ \right\}.$$

This completes the proof. □

The weighted CVaR representation yields the following two-stage solution.

Theorem 1 (Two-stage structure of the coupled DRO problem)**.** *There exists an optimal solution with the following structure. After reordering the indices so that $\ell_1 \le \ell_2 \le \cdots \le \ell_n$, i.e., in increasing order of losses, the solution exhibits a two-stage structure consisting of an outer and an inner component.*

(Outer structure) *The outer minimization assigns mass to smaller-loss samples first. Specifically,*

$$q'^{\star}_i = \frac{\bar{p}_i w^\star_i}{1-\epsilon}, \qquad w^\star \in [0,1]^n, \qquad \sum_{i=1}^{n} \bar{p}_i w^\star_i = 1 - \epsilon,$$

where

$$w^\star_i = \begin{cases} 1, & i < k, \\ \gamma, & i = k, \\ 0, & i > k, \end{cases}$$

for some $k \in [n]$ and $\gamma \in [0,1]$.

(Inner structure) *Given $q'^{\star}$, the inner maximization concentrates mass on larger-loss samples. In particular,*

$$q^\star_i = \begin{cases} 0, & i < t, \\ \eta \frac{1}{\alpha} q'^{\star}_t, & i = t, \\ \frac{1}{\alpha} q'^{\star}_i, & i > t, \end{cases} \qquad \sum_{i=1}^{n} q^\star_i = 1,$$

for some $t \in [n]$ and $\eta \in [0,1]$.

Proof. The CVaR representation turns the outer problem into a joint minimization over q' and a threshold. Mass transfers then characterize both the outer and inner optimizers.

By Lemma 2, for any $q' \in \mathcal{S}_{n-1}$,

$$\Phi_\alpha(q') = \min_{\tau \in \mathbb{R}} \left\{ \tau + \frac{1}{\alpha} \sum_{i=1}^{n} q'_i (\ell_i - \tau)_+ \right\}.$$

Therefore, the DRO objective can be written as the joint minimization

$$\min_{q' \in \mathcal{Q}_{1-\epsilon}(\bar{p})} \Phi_\alpha(q') = \min_{q' \in \mathcal{Q}_{1-\epsilon}(\bar{p}),\ \tau \in \mathbb{R}} \left\{ \tau + \frac{1}{\alpha} \sum_{i=1}^{n} q'_i (\ell_i - \tau)_+ \right\}. \tag{12}$$

Let $q'^\star$ be an outer minimizer of $\Phi_\alpha(q')$, and choose

$$\tau^\star \in \operatorname*{argmin}_{\tau \in \mathbb{R}} \left\{ \tau + \frac{1}{\alpha} \sum_{i} q'^\star_i (\ell_i - \tau)_+ \right\}.$$

For any $q' \in \mathcal{Q}_{1-\epsilon}(\bar{p})$ and $\tau \in \mathbb{R}$,

$$\tau + \frac{1}{\alpha} \sum_{i=1}^{n} q'_i (\ell_i - \tau)_+ \geq \Phi_\alpha(q') \geq \Phi_\alpha(q'^\star) = \tau^\star + \frac{1}{\alpha} \sum_{i=1}^{n} q'^\star_i (\ell_i - \tau^\star)_+.$$

Hence $(q'^\star, \tau^\star)$ attains the minimum in (12).

(Outer structure). Define $a_i^\star := (\ell_i - \tau^\star)_+$. For fixed $\tau^\star$, the objective is linear in q', so $q'^\star$ solves

$$\min_{q' \in \mathcal{Q}_{1-\epsilon}(\bar{p})} \sum_{i=1}^{n} q'_i a_i^\star.$$

By Lemma 1, write $q'_i = \bar{p}_i w_i / (1-\epsilon)$ with $w \in [0,1]^n$ and $\sum_i \bar{p}_i w_i = 1 - \epsilon$. Equivalently, letting $m_i := \bar{p}_i w_i$, we obtain

$$\min_{0 \leq m_i \leq \bar{p}_i} \sum_{i=1}^{n} a_i^\star m_i \quad \text{s.t.} \quad \sum_{i=1}^{n} m_i = 1 - \epsilon.$$

Reindex the samples so that $\ell_1 \leq \cdots \leq \ell_n$. Since $a_i^\star = (\ell_i - \tau^\star)_+$ is non-decreasing in ℓ_i, this indexing also satisfies $a_1^\star \leq \cdots \leq a_n^\star$. If there exist $i < j$ with $m_i < \bar{p}_i$ and $m_j > 0$, then for any $0 < \delta \leq \min\{\bar{p}_i - m_i,\ m_j\}$, the perturbation $m_i \leftarrow m_i + \delta$, $m_j \leftarrow m_j - \delta$ preserves feasibility and changes the objective by $\delta(a_i^\star - a_j^\star) \leq 0$ with strict decrease if $a_i^\star < a_j^\star$. Therefore, an optimal solution can be chosen that fills indices in this order, including when the $a_i^\star$ are tied. Consequently, there exist $k \in [n]$ and $\gamma \in [0,1]$ such that

$$w_i^\star = \begin{cases} 1, & i < k, \\ \gamma, & i = k, \\ 0, & i > k, \end{cases}$$

which shows that $w^\star$ has a threshold structure with at most one fractional component.

(Inner structure). Given $q'^\star$, the inner problem is the linear program

$$\max_{q \in \mathcal{Q}_\alpha(q'^\star)} \sum_{i=1}^{n} q_i \ell_i.$$

Using the same loss ordering $\ell_1 \leq \cdots \leq \ell_n$, suppose that there exist $i < j$ such that $q_i > 0$ and $q_j < \frac{1}{\alpha} q'^\star_j$. Then for any $0 < \delta \leq \min\{q_i,\ \frac{1}{\alpha} q'^\star_j - q_j\}$, the perturbation $q_i \leftarrow q_i - \delta$, $q_j \leftarrow q_j + \delta$ preserves feasibility and increases the objective by $\delta(\ell_j - \ell_i) \geq 0$, with strict increase if $\ell_j > \ell_i$. Therefore, an optimal solution can be chosen that fills indices in decreasing loss order, including when losses are tied, up to the caps $q_i \leq \frac{1}{\alpha} q'^\star_i$. Consequently, there exist $t \in [n]$ and $\eta \in [0,1]$ such that

$$q_i^\star = \begin{cases} 0, & i < t, \\ \eta \frac{1}{\alpha} q'^\star_t, & i = t, \\ \frac{1}{\alpha} q'^\star_i, & i > t, \end{cases}$$

which shows that $q^\star$ has a threshold structure with at most one fractional boundary index. □

For uniform nominal weights, the two-stage solution takes the following explicit form.

Corollary 1 (Exact trimming followed by exact worst-case allocation)**.** *Assume $\bar{p}_i = 1/n$ for all i, and reorder the samples so that*

$$\ell_1 \le \ell_2 \le \cdots \le \ell_n.$$

Assume that $(1-\epsilon)n \in \mathbb{Z}$ and $\alpha(1-\epsilon)n \in \mathbb{Z}$. Let

$$k_\epsilon := (1-\epsilon)n, \qquad m_\alpha := \alpha(1-\epsilon)n.$$

Then there exists an optimal solution $(q'^\star, q^\star)$ such that

$$q_i'^\star = \begin{cases} \dfrac{1}{(1-\epsilon)n}, & i = 1, \dots, k_\epsilon, \\ 0, & i > k_\epsilon, \end{cases}$$

and

$$q_i^\star = \begin{cases} 0, & i = 1, \dots, k_\epsilon - m_\alpha, \\ \dfrac{1}{\alpha(1-\epsilon)n}, & i = k_\epsilon - m_\alpha + 1, \dots, k_\epsilon, \\ 0, & i > k_\epsilon. \end{cases}$$

In particular, $q'^\star$ is the exact ϵ-trimmed empirical distribution, and $q^\star$ places uniform mass on the largest $\alpha(1-\epsilon)n$ losses among the retained samples.

Proof. By Theorem 1, there exists an optimal solution $(q'^\star, q^\star)$ such that the outer optimizer is obtained by filling the smallest-loss indices first. Then, we have

$$q_i'^\star = \frac{\bar{p}_i w_i^\star}{1-\epsilon},$$

where for some k and $\gamma \in [0,1]$,

$$w_i^\star = \begin{cases} 1, & i < k, \\ \gamma, & i = k, \\ 0, & i > k, \end{cases} \qquad \sum_{i=1}^n \bar{p}_i w_i^\star = 1 - \epsilon.$$

Since $\bar{p}_i = 1/n$, the normalization condition becomes

$$\sum_{i=1}^n \bar{p}_i w_i^\star = \frac{1}{n}\big((k-1) + \gamma\big) = 1 - \epsilon.$$

Because $(1-\epsilon)n = k_\epsilon \in \mathbb{Z}$, it follows that $(k-1) + \gamma = k_\epsilon$. Since $k_\epsilon \in \mathbb{Z}$ and $k_\epsilon = (k-1) + \gamma$ with $\gamma \in [0,1]$, we may take $\gamma = 1$ and $k = k_\epsilon$, which also covers $\epsilon = 0$.

Then, we have

$$q_i'^\star = \begin{cases} \dfrac{1}{(1-\epsilon)n}, & i = 1, \dots, k_\epsilon, \\ 0, & i > k_\epsilon. \end{cases}$$

Now consider the inner maximization

$$\max_{q \in \mathcal{Q}_\alpha(q'^\star)} \sum_{i=1}^n q_i \ell_i, \qquad 0 \le q_i \le \frac{1}{\alpha} q_i'^\star, \quad \sum_{i=1}^n q_i = 1.$$

Since $\ell_1 \le \cdots \le \ell_n$, the objective is maximized by assigning as much mass as possible to the largest-loss indices. Because the constraint $q_i \le \frac{1}{\alpha} q_i'^\star$ holds for every i, and $q_i'^\star = 0$ for all $i > k_\epsilon$, we have

$$q_i^\star = 0 \qquad \text{for all } i > k_\epsilon.$$

For each retained index $i \le k_\epsilon$, we have

$$\frac{1}{\alpha} q_i'^\star = \frac{1}{\alpha(1-\epsilon)n} = \frac{1}{m_\alpha},$$

since $m_\alpha = \alpha(1-\epsilon)n$. Thus every non-zero upper bound is the same, namely $1/m_\alpha$. Therefore the maximizer is obtained by saturating exactly the largest m_α indices among $\{1, \dots, k_\epsilon\}$, namely $i = k_\epsilon - m_\alpha + 1, \dots, k_\epsilon$. This yields

$$q_i^\star = \begin{cases} 0, & i = 1, \dots, k_\epsilon - m_\alpha, \\ \dfrac{1}{\alpha(1-\epsilon)n}, & i = k_\epsilon - m_\alpha + 1, \dots, k_\epsilon, \\ 0, & i > k_\epsilon. \end{cases}$$

Since there are exactly m_α such indices, the total mass is

$$m_\alpha \cdot \frac{1}{\alpha(1-\epsilon)n} = \frac{m_\alpha}{m_\alpha} = 1,$$

so the simplex constraint is satisfied. This proves the claim. □

Remark. When $(1-\epsilon)n \notin \mathbb{Z}$, the loss-ordered outer solution has a fractional boundary atom at index $\lfloor (1-\epsilon)n \rfloor + 1$. The inner optimizer fills the retained atoms in decreasing loss order, with partial allocation at the cutoff when needed.

For uniform nominal weights, Algorithm 2 computes the weights by sorting the losses, trimming the largest losses, and emphasizing the largest retained losses. It assumes that $(1-\epsilon)n$ and $\alpha(1-\epsilon)n$ are integers; otherwise, fractional allocation is required at the boundaries as described above.

Algorithm 2 Procedure for minimizing the coupled DRO objective under additive losses

Require: data $\{z_i\}_{i=1}^n$, parameters ϵ, α, step size η

1: Initialize θ

2: Set $k_\epsilon \leftarrow \lfloor (1-\epsilon)n \rfloor$ and $m_\alpha \leftarrow \lfloor \alpha(1-\epsilon)n \rfloor$

3: **while** not converged **do**

4: Compute losses $\ell_i = \ell(z_i, \theta)$ for $i = 1, \dots, n$

5: Reorder the indices so that

$$\ell_1 \le \ell_2 \le \cdots \le \ell_n$$

6: Set

$$q_i' \leftarrow \begin{cases} \frac{1}{(1-\epsilon)n}, & i = 1, \dots, k_\epsilon, \\ 0, & i > k_\epsilon \end{cases}$$

7: Set

$$q_i \leftarrow \begin{cases} 0, & i = 1, \dots, k_\epsilon - m_\alpha, \\ \frac{1}{\alpha(1-\epsilon)n}, & i = k_\epsilon - m_\alpha + 1, \dots, k_\epsilon, \\ 0, & i > k_\epsilon \end{cases}$$

8: Update

$$\theta \leftarrow \theta - \eta \nabla_\theta \left(\sum_{i=1}^n q_i \, \ell(z_i, \theta) \right)$$

9: **end while**

10: **return** θ

A.1.2 NON-DECOMPOSABLE LOSSES

For non-decomposable losses, interactions across samples generally prevent the explicit weight allocation derived above. We instead alternate between ascent in q, an update of q' based on an estimated supergradient, and descent in θ. The main step is to characterize the sensitivity of the inner value function

$$\Phi_\alpha(q'; \theta) = \max_{q \in \mathcal{Q}_\alpha(q')} \mathcal{L}(q, \theta). \tag{13}$$

Proposition 1 relates this sensitivity to the upper-bound KKT multipliers under local stability (Rockafellar & Wets, 1998). Theorem 2 gives their explicit form.

Proposition 1 (restated). *Fix θ, $0 < \alpha < 1$, and $q' \in \mathcal{S}_{n-1}$ with $q'_i > 0$ for all i. Set $b = q'/\alpha$, $\mathcal{F}(b) = \{q \in \mathcal{S}_{n-1} : q \le b\}$, and*

$$V(b) := \max_{q \in \mathcal{F}(b)} \mathcal{L}(q, \theta), \qquad \Phi_\alpha(q'; \theta) = V(b).$$

Let $q^\star(b)$ be a global maximizer, and assume that $\mathcal{L}(\cdot, \theta)$ is continuous on the simplex and continuously differentiable near $q^\star(b)$. Assume local stability: for every sufficiently small Δb, a global maximizer $q^\star(b + \Delta b)$ can be selected such that

$$\|q^\star(b + \Delta b) - q^\star(b)\| \le C\|\Delta b\| \tag{14}$$

for some constant $C > 0$ independent of Δb. Then any KKT multiplier $\mu^\star$ for the upper-bound constraints $q_i \le b_i$ at $q^\star(b)$ satisfies

$$\frac{\mu^\star}{\alpha} \in \partial^+ \Phi_\alpha(q'; \theta),$$

where ∂^+ denotes the superdifferential: $s \in \partial^+ F(u)$ if $F(u + h) \le F(u) + \langle s, h\rangle + o(\|h\|)$ as $h \to 0$ within the domain of F.

Proof. We prove the result by combining the KKT conditions with a first-order Taylor expansion. The local stability assumption controls the remainder under perturbations of b.

Write $\Delta q = q^\star(b + \Delta b) - q^\star(b)$ and $g^\star = \nabla_q \mathcal{L}(q^\star(b), \theta)$. The KKT conditions give $g^\star = \nu^\star \mathbf{1} - \lambda^\star + \mu^\star$, where $\lambda^\star$ and $\mu^\star$ are the nonnegative lower- and upper-bound multipliers. Since both solutions sum to one, feasibility and complementary slackness imply

$$\begin{aligned} \langle g^\star, \Delta q\rangle &= -\langle \lambda^\star, \Delta q\rangle + \langle \mu^\star, \Delta q\rangle \\ &\le \langle \mu^\star, \Delta b\rangle. \end{aligned}$$

By (14), $\|\Delta q\| = O(\|\Delta b\|)$. A first-order Taylor expansion therefore yields

$$\begin{aligned} V(b + \Delta b) - V(b) &= \langle g^\star, \Delta q\rangle + o(\|\Delta b\|) \\ &\le \langle \mu^\star, \Delta b\rangle + o(\|\Delta b\|). \end{aligned}$$

Thus $\mu^\star \in \partial^+ V(b)$. Substituting $b = q'/\alpha$ gives $\mu^\star/\alpha \in \partial^+ \Phi_\alpha(q'; \theta)$. □

Applying the supergradient result to weighted Cox. The following lemma gives conditions under which the weighted Cox loss satisfies Proposition 1. The key step is to bound changes in the inner optimizer by changes in its upper bounds.

For fixed θ with $f_\theta(x_j) \in \mathbb{R}$ for all j, write $a_j = \exp(f_\theta(x_j)) > 0$ and $S_i(q) = \sum_{j: y_j \ge y_i} a_j q_j$. The weighted Cox objective is $\mathcal{L}(q, \theta) = \sum_i \delta_i q_i \{\log S_i(q) - \log a_i\}$, with zero-mass terms interpreted by continuity.

Lemma 3 (Sufficient conditions for weighted Cox). *Let $0 < \alpha < 1$, $q' \in \mathcal{S}_{n-1}$ with $q'_i > 0$ for all i, and $b = q'/\alpha$. Suppose the weighted Cox objective has a unique global maximizer $q^\star$ over $\mathcal{F}(b)$, $S_i(q^\star) > 0$ for every observed event i with $\delta_i = 1$, and the coordinates of $g^\star = \nabla_q \mathcal{L}(q^\star, \theta)$ are pairwise distinct. Then the objective is continuous on the simplex and continuously differentiable near $q^\star$, and the global inner optimizer satisfies the local stability condition* (14)*. Consequently, the supergradient characterization in Proposition 1 applies.*

Proof. We first establish continuity of $\mathcal{L}$, continuous differentiability near $q^\star$, and convergence of global maximizers as the upper bounds approach b. We then use the strict ordering of the gradient coordinates to derive an explicit representation of these maximizers and establish local Lipschitz stability.

Since $S_i(q) \ge a_i q_i$, the term $q_i \log S_i(q)$ tends to zero whenever $S_i(q) \to 0$ on the simplex. This gives a continuous extension of the objective to the simplex. Positivity of all event risk-set sums at $q^\star$ ensures continuous differentiability in a neighborhood of $q^\star$.

The point q' is strictly feasible relative to the simplex because $0 < q'_i < b_i$. Let $d = \min_i(b_i - q'_i) > 0$. For sufficiently small Δb, the point $(1 - t)q^\star + tq'$, where $t = \|\Delta b\|_\infty / d$, belongs to $\mathcal{F}(b + \Delta b)$ and

converges to $q^\star$. By compactness and continuity, any limit of global maximizers for nearby bounds is therefore a global maximizer at b. Uniqueness implies that all such maximizers converge to $q^\star$.

Choose a permutation π such that $g^\star_{\pi(1)} > \cdots > g^\star_{\pi(n)}$. Continuity of the gradient ensures that this ordering persists at every nearby global maximizer $\widehat{q}$. The first-order necessary condition on the convex set $\mathcal{F}(b + \Delta b)$ gives

$$\langle \nabla_q \mathcal{L}(\widehat{q}, \theta), q - \widehat{q} \rangle \le 0, \qquad q \in \mathcal{F}(b + \Delta b).$$

Thus $\widehat{q}$ maximizes its linearized objective over this set. With strictly ordered coefficients, this linear problem has the unique solution

$$\widehat{q}_{\pi(k)} = \min \left\{ b_{\pi(k)} + \Delta b_{\pi(k)}, \ \left[1 - \sum_{j<k} (b_{\pi(j)} + \Delta b_{\pi(j)}) \right]_+ \right\}, \qquad k \in [n].$$

This expression is a continuous piecewise affine function of $b + \Delta b$, and hence is locally Lipschitz. It equals $q^\star$ at $\Delta b = 0$, proving (14). □

We next express the multipliers in terms of the loss gradient.

Theorem 2 (Outer supergradient characterization)**.** *Fix $\theta \in \Theta$ and $q' \in \mathcal{S}_{n-1}$. Let $\Phi_\alpha(q'; \theta)$ be the inner value function defined in* (13)*, where $\mathcal{L}(\cdot, \theta) : \mathcal{S}_{n-1} \to \mathbb{R}$ is continuous.*

Assume that the inner maximization problem

$$\max_q \mathcal{L}(q, \theta) \quad \textit{s.t.} \quad \sum_{i=1}^n q_i = 1, \ q_i \ge 0, \ q_i \le \tfrac{1}{\alpha} q'_i$$

admits an optimal solution $q^\star$, and that a constraint qualification holds so that the KKT conditions are satisfied with multipliers

$$\nu^\star \in \mathbb{R}, \quad \lambda^\star \in \mathbb{R}^n_+, \quad \mu^\star \in \mathbb{R}^n_+,$$

where $\nu^\star$, $\lambda^\star$, and $\mu^\star$ are the Lagrange multipliers associated with the constraints $\sum_{i=1}^n q_i = 1$, $q_i \ge 0$, and $q_i \le \frac{1}{\alpha} q'_i$, respectively. Assume also that $\mathcal{L}(\cdot, \theta)$ is continuously differentiable on a neighborhood of $q^\star$.

Let $g^\star := \nabla_q \mathcal{L}(q^\star, \theta)$. Then the following statements hold:

(i) Stationarity and complementary slackness

$$g^\star - \nu^\star \mathbf{1} + \lambda^\star - \mu^\star = 0,$$
$$\lambda^\star_i q^\star_i = 0, \qquad \mu^\star_i \left(q^\star_i - \tfrac{1}{\alpha} q'_i \right) = 0, \qquad \forall i.$$

(ii) Interior characterization *For any i such that $0 < q^\star_i < \frac{1}{\alpha} q'_i$,*

$$\nu^\star = g^\star_i.$$

(iii) Multiplier characterization *For each nondegenerate coordinate with $q'_i > 0$,*

$$\mu^\star_i = \begin{cases} (g^\star_i - \nu^\star)_+, & \text{if } q^\star_i = \frac{1}{\alpha} q'_i, \\ 0, & \text{if } q^\star_i < \frac{1}{\alpha} q'_i. \end{cases}$$

(iv) Supergradient *If, in addition, $0 < \alpha < 1$, $q'_i > 0$ for all i, and the local stability condition* (14) *holds at $q^\star = q^\star(b)$ with $b = q'/\alpha$, then*

$$\frac{1}{\alpha} \mu^\star \in \partial^+ \Phi_\alpha(q'; \theta).$$

Proof. We first use stationarity and complementary slackness to identify the upper-bound multipliers. Applying Proposition 1 under the additional assumptions in part (iv) then gives the supergradient characterization.

Set

$$b := \frac{1}{\alpha} q', \qquad \mathcal{F}(b) := \left\{ q \in \mathbb{R}^n : \sum_{i=1}^{n} q_i = 1,\ q \geq 0,\ q \leq b \right\}.$$

By the definition of Φ_α,

$$\Phi_\alpha(q'; \theta) = \max_{q \in \mathcal{F}(b)} \mathcal{L}(q, \theta).$$

By assumption, an optimal solution exists and a constraint qualification holds, and hence the KKT conditions are satisfied.

Step 1: KKT conditions. Let $q^\star$ be an optimal solution of the inner problem, and let

$$\nu^\star \in \mathbb{R}, \qquad \lambda^\star \in \mathbb{R}^n_+, \qquad \mu^\star \in \mathbb{R}^n_+$$

be the corresponding Lagrange multipliers associated with the constraints $\sum_{i=1}^n q_i = 1$, $q_i \geq 0$, and $q_i \leq b_i$, respectively. Define

$$g^\star := \nabla_q \mathcal{L}(q^\star, \theta).$$

Then the stationarity condition is

$$g^\star - \nu^\star \mathbf{1} + \lambda^\star - \mu^\star = 0,$$

where $\mathbf{1} \in \mathbb{R}^n$ denotes the vector of all ones. The complementary slackness conditions are

$$\lambda_i^\star q_i^\star = 0, \qquad \mu_i^\star (q_i^\star - b_i) = 0, \qquad \forall i \in [n].$$

Step 2: Identification of the multipliers for the upper-bound constraints. Consider an index i such that $0 < q_i^\star < b_i$. Then both the lower-bound and upper-bound constraints are inactive, so

$$\lambda_i^\star = 0, \qquad \mu_i^\star = 0.$$

Hence, by the stationarity condition,

$$g_i^\star = \nu^\star.$$

If $q_i^\star < b_i$, then complementary slackness gives

$$\mu_i^\star = 0.$$

If $0 < q_i^\star = b_i$, then the lower-bound constraint is inactive, so

$$\lambda_i^\star = 0.$$

The stationarity condition then gives

$$\mu_i^\star = g_i^\star - \nu^\star.$$

Since $\mu_i^\star \geq 0$, we obtain

$$\mu_i^\star = \begin{cases} (g_i^\star - \nu^\star)_+, & \text{if } 0 < q_i^\star = b_i, \\ 0, & \text{if } q_i^\star < b_i. \end{cases}$$

The degenerate case $q_i^\star = b_i = 0$ is not included in this characterization. In this case, the lower-bound and upper-bound constraints are both active, and the multipliers $\lambda_i^\star$ and $\mu_i^\star$ are not separately identifiable from the stationarity condition alone.

Step 3: Supergradient characterization. Under the additional assumptions in part (iv), Proposition 1 gives

$$\frac{1}{\alpha} \mu^\star \in \partial^+ \Phi_\alpha(q'; \theta),$$

which completes the proof. □

Supergradient estimation. At an exact inner maximizer, Theorem 2(ii) gives $g_i^\star = \nu^\star$ on interior coordinates. For an approximate maximizer $\hat{q}$, let $\hat{g} = \nabla_q \mathcal{L}(\hat{q}, \theta)$ and estimate this common value by

$$\nu := \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \hat{g}_i, \qquad \mathcal{I} = \left\{ i \in [n] : 0 < \hat{q}_i < \tfrac{1}{\alpha} q'_i \right\}.$$

Substituting $\hat{q}$, $\hat{g}$, and ν into the multiplier formula gives the supergradient estimate used in the q'-update. The exact characterization requires the assumptions of part (iv) at an inner optimum. If $\mathcal{I}$ is empty, q' is left unchanged. We use small positive lower bounds on sample weights and weighted risk-set sums for numerical stability.

Theorem 3 below establishes local convergence of the outer variables (q', θ) under exact inner maximization and the stated assumptions. For computational efficiency, we approximate the inner maximizer using a finite number of gradient steps and estimate the KKT multipliers from the resulting iterate for the outer update.

A.1.3 Computational complexity

Let n be the sample size, d the covariate dimension, T the number of inner weight updates, and K the number of outer iterations in Algorithm 3. For the linear Cox model, evaluating the risk scores and their parameter derivatives requires $O(nd)$ time. Weighted risk-set sums are computed cumulatively after sorting observations by time. The two capped-simplex projections can be computed exactly in $O(n \log n)$ time by sorting; our implementation finds the multiplier λ by bisection, which costs $O(n)$ per bisection step. Consequently, the total computational complexity is

$$O(KTn(d + \log n)).$$

For fixed K, T, and d, this cost is nearly linear in the sample size. The additional memory required for the sample weights and cumulative risk-set quantities is $O(n)$, beyond storage for the data and model parameters. Empirical model-fitting times are reported in Appendix A.3.11.

A.1.4 Local convergence analysis

We analyze local convergence of the outer variables q' and θ, assuming that before each q'-update, we obtain the exact global inner maximizer $q^\star$ given the current q' and θ. Under the assumptions below, we show that the outer iterates of the specified alternating scheme with KL-projected q'-updates and gradient updates of θ converge R-linearly in the Euclidean norm to a strict local minimizer of the resulting outer value function, for a sufficiently small common step size and sufficiently close feasible initialization.

Write $x = (q', \theta)$, $\mathcal{C} = \mathcal{Q}_{1-\epsilon}(\bar{p})$, and $F(x) = \Phi_\alpha(q'; \theta)$. Let $q^\star(x)$ denote the global inner maximizer and $\mu^\star(x)$ its upper-bound KKT multipliers. With $D_{\mathrm{KL}}(u\|v) = \sum_i u_i \log(u_i/v_i)$, the outer weight update takes the entropic mirror-descent form of Beck & Teboulle (2003):

$$q'^{+} = \arg\min_{u \in \mathcal{C}} \left\{ \left\langle \frac{\mu^\star(x)}{\alpha}, u - q' \right\rangle + \frac{1}{\eta} D_{\mathrm{KL}}(u\|q') \right\}.$$

This is equivalent to the exponentiated update (7) with KL projection. We retain the update order of Algorithm 1: T weight steps with θ fixed, followed by a parameter step using the same inner solution as the last weight step. For the analysis, set $\eta_{q'} = \eta_\theta = \eta$ and fix $T \geq 1$.

Theorem 3 (Local convergence)**.** *Let $0 < \alpha < 1$, $0 < \epsilon < 1$, and $x^\star = (q'^\star, \theta^\star)$ be feasible, with $q'^\star > 0$ and $\theta^\star$ interior to the parameter space. Assume:*

- *(i)* ***Smooth inner solution.*** *$\mathcal{L}$ is twice continuously differentiable near $(q^\star(x^\star), \theta^\star)$. For every x near $x^\star$, the global inner maximizer $q^\star(x)$ is unique, at least one coordinate satisfies $0 < q_i^\star(x) < q'_i/\alpha$, and both $q^\star(x)$ and $\mu^\star(x)$ are continuously differentiable in x.*
- *(ii)* ***Outer stationarity.*** *The outer KKT conditions hold at $x^\star$, and every active upper bound of q' has a strictly positive outer KKT multiplier.*
- *(iii)* ***Positive local curvature.*** *At $x^\star$, the Hessian of F is positive definite along directions that preserve the simplex constraint and keep the active bound coordinates of q' fixed, with θ free to vary.*

Then $(q'^\star, \theta^\star)$ is a strict local minimizer of the outer value function F. For a sufficiently small fixed $\eta > 0$ and feasible initialization sufficiently close to $(q'^\star, \theta^\star)$, there exist constants $C > 0$ and $\rho \in (0, 1)$ such that the outer iterates of the specified KL-projected alternating scheme satisfy

$$\left\| \begin{pmatrix} q'^{(k)} - q'^\star \\ \theta^{(k)} - \theta^\star \end{pmatrix} \right\|_2 \leq C\rho^k, \qquad k = 0, 1, 2, \ldots.$$

Proof. We first show that F is smooth and $x^\star$ is a strict local minimizer. Next, we show that the KL update fixes the weight coordinates whose upper bounds are active at $x^\star$. With these coordinates fixed, we prove that a full update cycle contracts the error in a weighted norm, yielding the stated Euclidean bound.

Step 1: Smoothness and local minimality. By assumption (i), the inner optimizer varies smoothly with q' and θ. Differentiating the optimized loss and applying the inner KKT conditions gives, for perturbations along the simplex,

$$\mathrm{d}F = \left\langle \frac{\mu^\star(x)}{\alpha}, \mathrm{d}q' \right\rangle + \langle \nabla_\theta \mathcal{L}(q^\star(x), \theta), \mathrm{d}\theta \rangle \,.$$

The derivatives in this expression are continuously differentiable by assumption (i), so F is twice continuously differentiable locally. Outer stationarity and strict complementarity in (ii), together with the positive curvature in (iii), give the second-order sufficient conditions for $x^\star$ to be a strict local minimizer.

Step 2: Identification of active upper bounds. We next show that nearby KL updates place the same coordinates of q' at their upper bounds as at $x^\star$. We adapt the active-constraint argument of Bertsekas (1976) to the KL update. At $x^\star$, its subproblem has a positive definite objective Hessian, independent active constraint gradients, and strictly positive active multipliers by (ii). Its solution therefore varies smoothly with nearby inputs while retaining the same active upper bounds. These weight coordinates can thus be held fixed when analyzing subsequent nearby updates.

Step 3: Contraction and the Euclidean error bound. With the active coordinates fixed, we analyze the remaining weight coordinates jointly with θ. Let U have orthonormal columns spanning the directions that preserve $\sum_i q'_i = 1$ and keep q'_i fixed whenever $q'^\star_i = \bar{p}_i/(1-\epsilon)$. Write $q' = q'^\star + Uz$, so z describes changes in the remaining weight coordinates. By assumption (iii), the Hessian H of $F(q'^\star + Uz, \theta)$ at $(0, \theta^\star)$ is positive definite. Set

$$M_q = U^\top \operatorname{diag}(1/q'^\star) U, \qquad M = \operatorname{diag}(M_q, I), \qquad D = \operatorname{diag}(M_q^{-1}, 0).$$

Zero-dimensional weight blocks are omitted. Differentiating the KL optimality equations gives $R = I - \eta DH$ for a weight-only step and $J = I - \eta M^{-1} H$ for the final weight and parameter step. The derivative of a full cycle is therefore JR^{T-1}.

For sufficiently small $\eta > 0$, $H \succ 0$ and $0 \preceq D \preceq M^{-1}$ imply $\|J\|_H < 1$ and $\|R\|_H \leq 1$, where $\|v\|_H^2 = v^\top H v$ and the corresponding matrix norm is induced by this vector norm. By continuity of the cycle map's derivative, the full cycle is therefore a local contraction in this norm after these upper-bound coordinates have been fixed. Iterating the contraction gives geometric decay of the error in the H-norm. Norm equivalence and the orthonormality of U yield the stated Euclidean bound $C\rho^k$, with all fixed factors and the finite initial segment absorbed into C. Continuity of $q^\star(x)$ also gives convergence of the exact inner maximizers. □

A.1.5 ALTERNATING GRADIENT ALGORITHM

Algorithm 3 alternates updates of the adversarial weights q and nominal weights q', followed by a model-parameter update.

Algorithm 3 Alternating gradient algorithm for the coupled DRO problem

Input: step sizes $\eta_q, \eta_{q'}, \eta_\theta$; update steps T, K; initial values $\theta^{(0)}, q^{(0)}, q'^{(0)}$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  Set $q^{(k,0)} \leftarrow q^{(k)}, \quad q'^{(k,0)} \leftarrow q'^{(k)}$
  ▷ Update (q, q')
  for $t = 0, \ldots, T-1$ **do**
    Compute $g^{(k,t)} \leftarrow \nabla_q \mathcal{L}(q^{(k,t)}, \theta^{(k)})$
    Set

$$\tilde{q}_i^{(k,t+1)} \leftarrow \frac{q_i^{(k,t)} \exp\left(\eta_q g_i^{(k,t)}\right)}{\sum_{j=1}^n q_j^{(k,t)} \exp\left(\eta_q g_j^{(k,t)}\right)}$$

    Project

$$q^{(k,t+1)} \leftarrow \Pi_{\mathcal{Q}_\alpha(q'^{(k,t)})}(\tilde{q}^{(k,t+1)})$$

    Compute $g^{(k,t+1)} \leftarrow \nabla_q \mathcal{L}(q^{(k,t+1)}, \theta^{(k)})$
    Define

$$\mathcal{I}^{(k,t+1)} \leftarrow \left\{ i \in [n] : 0 < q_i^{(k,t+1)} < \tfrac{1}{\alpha} q_i'^{(k,t)} \right\}$$

    if $\mathcal{I}^{(k,t+1)} = \emptyset$ **then**
      Set $q'^{(k,t+1)} \leftarrow q'^{(k,t)}$
    else
      Compute

$$\nu^{(k,t+1)} \leftarrow \frac{1}{|\mathcal{I}^{(k,t+1)}|} \sum_{i \in \mathcal{I}^{(k,t+1)}} g_i^{(k,t+1)}$$

      Set

$$s_i^{(k,t+1)} \leftarrow \begin{cases} \frac{1}{\alpha}(g_i^{(k,t+1)} - \nu^{(k,t+1)})_+, & \text{if } 0 < q_i^{(k,t+1)} = \frac{1}{\alpha} q_i'^{(k,t)}, \\ 0, & \text{otherwise.} \end{cases}$$

      Set

$$\tilde{q}_i'^{(k,t+1)} \leftarrow \frac{q_i'^{(k,t)} \exp\left(-\eta_{q'} s_i^{(k,t+1)}\right)}{\sum_{j=1}^n q_j'^{(k,t)} \exp\left(-\eta_{q'} s_j^{(k,t+1)}\right)}$$

      Project

$$q'^{(k,t+1)} \leftarrow \Pi_{\mathcal{Q}_{1-\epsilon}(\bar{p})}(\tilde{q}'^{(k,t+1)})$$

    end if
  end for
  Set $q^{(k+1)} \leftarrow q^{(k,T)}, \quad q'^{(k+1)} \leftarrow q'^{(k,T)}$
  ▷ Update θ
  Update

$$\theta^{(k+1)} \leftarrow \theta^{(k)} - \eta_\theta \nabla_\theta \mathcal{L}(q^{(k+1)}, \theta^{(k)})$$

end for

A.2 Sensitivity analysis

A.2.1 Robustness parameters

We examine sensitivity to α and ϵ, which control robustness to subpopulation shift and outlier contamination, respectively. We consider $\alpha \in \{0.05, 0.2, 0.4, 1\}$ and $\epsilon \in \{0, 0.01, 0.05, 0.1\}$. When varying α, we fix $\epsilon = 0$; when varying ϵ, we fix $\alpha = 0.1$.

We use the data-generating process in Section 6.1 with adjusted mixture proportions and $N_{\text{train}} = N_{\text{test}} = 500$. The training distribution is $P_{\text{train}} = 0.95P_1 + 0.05P_2$ for subpopulation shift and $\widetilde{P}_{\text{train}} = 0.7P_1 + 0.2P_2 + 0.1\widetilde{P}$ for outlier contamination. Both settings use the clean test distribution $P_{\text{test}} = 0.8P_1 + 0.2P_2$.

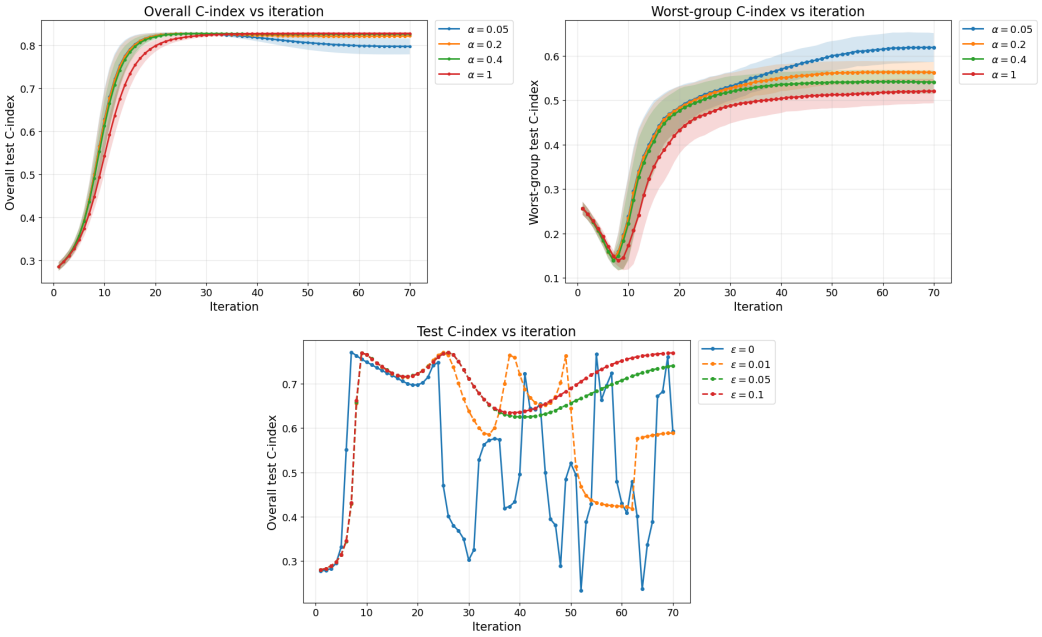

Figure 2: Sensitivity to α and ϵ. (Top-left and top-right) Overall and worst-group test C-index for varying α under subpopulation shift. (Bottom) Overall test C-index for varying ϵ under outlier contamination.

In Figure 2, smaller α improves worst-group C-index, with a slight decrease in overall C-index at $\alpha = 0.05$. Under contamination, positive ϵ reduces training oscillations, with smoother trajectories at $\epsilon = 0.05$ and 0.1.

A.3 Additional experiments and experimental details

We examine robustness in additional simulated and real-data settings and describe the experimental procedures. All experiments were conducted on a server equipped with Intel Xeon Gold 6230R CPUs and NVIDIA GeForce RTX 3090 GPUs.

A.3.1 Simulation: Subpopulation shift

We isolate subpopulation shift by using uncontaminated mixtures of two latent groups. The training distribution is $P_{\text{train}} = \pi_1^{\text{tr}} P_1 + \pi_2^{\text{tr}} P_2$, with $(\pi_1^{\text{tr}}, \pi_2^{\text{tr}}) = (0.95, 0.05)$ and $N_{\text{train}} = 3{,}000$. The test distribution is $P_{\text{test}} = \pi_1^{\text{te}} P_1 + \pi_2^{\text{te}} P_2$, with $(\pi_1^{\text{te}}, \pi_2^{\text{te}}) = (0.8, 0.2)$ and $N_{\text{test}} = 5{,}000$. The remaining data-generation settings follow Section 6.1.

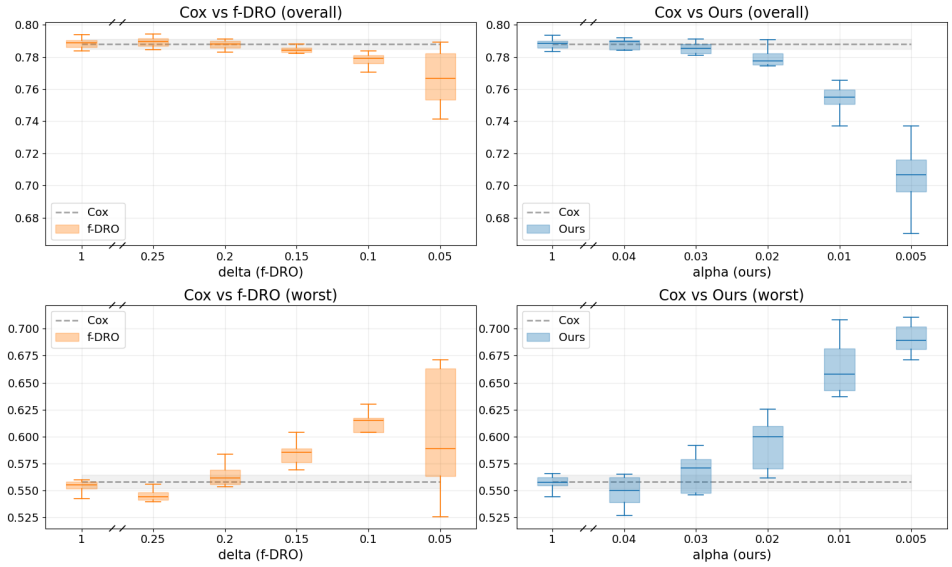

Figure 3: Overall and worst-group C-index under subpopulation shift for the Cox model, f-DRO (left panels), and the proposed method (right panels), across robustness parameters. Box plots summarize 20 independent runs.

Figure 3 shows that both robust methods can improve worst-group C-index at a cost in overall C-index, depending on the robustness parameters.

A.3.2 Simulation: Nonlinear risk functions

We extend the simulation in Section 6.1 to nonlinear group-specific risk functions with linear, quadratic, and interaction terms. We index the majority, minority, and outlier components by $g = 0, 1, 2$, respectively, and define

$$r_0(x) = x_1 + x_2 + 0.15x_1^2 + 0.15x_2^2 + 0.15x_1x_2,$$
$$r_1(x) = x_1 - x_2 + 0.15x_1^2 - 0.15x_2^2 - 0.15x_1x_2,$$
$$r_2(x) = -3x_1 - 3x_2 - 0.45x_1^2 - 0.45x_2^2 - 0.45x_1x_2.$$

Survival times follow the group-specific hazard $h_g(t \mid x) = h_0 \exp(r_g(x))$. The training mixture proportions are $(0.90, 0.08, 0.02)$, and the clean test proportions are $(0.80, 0.20, 0)$. We use a single-hidden-layer neural Cox model with width 5; the training and test sample sizes, covariate generation, and censoring setup are otherwise the same as in Section 6.1.

Training trajectories. Figure 4 summarizes the test C-index trajectories, including the substantial variation of f-DRO and the proposed method with $\epsilon = 0$. With $\epsilon = 0.05$, the proposed method attains higher late-stage worst-group C-index than Cox-NN and f-DRO, with lower overall C-index than Cox-NN. The corresponding IBS trajectories are provided in Appendix A.3.4.

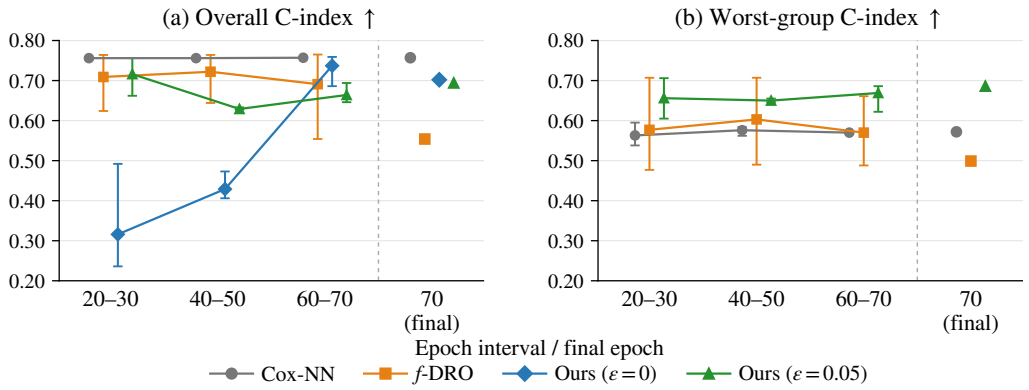

Figure 4: Test C-index under nonlinear risks and outlier contamination. Points show interval means, whiskers show minima and maxima over epochs, and lines connect means. Epoch 70 is shown separately.

Repeated-run performance. Table 4 reports a separate experiment averaged over five independent repetitions. The proposed method with $\epsilon = 0.05$ outperforms f-DRO on all four metrics and improves worst-group C-index and IBS relative to Cox-NN, at a cost in overall performance.

Table 4: Test performance under nonlinear risks and outlier contamination, averaged over five independent repetitions. Bold and underlined entries denote the best and second-best results, respectively.

Method	Overall C-index ↑	Worst-group C-index ↑	Overall IBS ↓	Worst-group IBS ↓
Cox-NN	**0.753**	<u>0.562</u>	**0.125**	0.241
f-DRO	0.551	0.522	0.161	<u>0.217</u>
Ours ($\epsilon = 0.05$)	<u>0.698</u>	**0.686**	<u>0.152</u>	**0.199**

A.3.3 Simulation: Higher dimension, smaller samples, and heavier censoring

We evaluate the methods with $d = 10$ covariates, training sample sizes $n \in \{500, 1000\}$, 5,000 test samples, and 50% censoring. Without contamination, the training mixture proportions of the two clean groups are $(0.9, 0.1)$; with contamination, the majority, minority, and outlier proportions are $(0.90, 0.08, 0.02)$. The clean test proportions remain $(0.80, 0.20)$ in both settings.

We define a shared risk component and a component controlling group heterogeneity as

$$C(x) = 0.6x_1 + 0.5x_5, \qquad S(x) = 1.2x_2 + \frac{0.7}{\sqrt{2}}(x_3 + x_4) + 0.35x_6.$$

The majority and minority log-risk functions are

$$r_0(x) = C(x) + 0.7S(x), \qquad r_1(x) = C(x) - 0.7S(x),$$

and the outlier log-risk function is $r_2(x) = -2(C(x) + 0.7S(x))$. The covariates $x_7, \ldots, x_{10}$ are noise variables and do not affect the true risk. The robust subpopulation parameters are fixed at $\delta = \alpha = 0.1$ for f-DRO and the proposed method, respectively. For the proposed method, we compare $\epsilon \in \{0, 0.01, 0.02, 0.05\}$ under contamination.

Tables 5 and 6 report overall and worst-group C-index, IBS, and tail loss. Without contamination, the robust methods improve worst-group C-index and IBS relative to Cox, with lower overall C-index and higher overall IBS. Under contamination, increasing ϵ improves overall C-index and IBS, while intermediate values achieve the best worst-group C-index and IBS among the evaluated settings.

Table 5: Simulation results with $d = 10$, $n = 500$ training samples, and 50% censoring. Values are mean (standard deviation) over 10 runs. Bold and underlined entries denote the best and second-best results within each contamination setting, respectively.

Method ($n = 500$)	**Overall C-index** ↑	**Worst-group C-index** ↑	**Overall IBS** ↓	**Worst-group IBS** ↓	**Overall tail loss** ↓	**Worst-group tail loss** ↓
			Without Outliers			
Cox	**0.787** (0.007)	0.458 (0.025)	**0.150** (0.004)	0.357 (0.022)	11.315 (0.373)	11.850 (0.581)
f-DRO	0.716 (0.043)	**0.537** (0.063)	0.194 (0.016)	**0.230** (0.015)	8.711 (0.130)	8.309 (0.059)
Ours ($\epsilon = 0$)	0.708 (0.038)	0.526 (0.049)	0.207 (0.012)	**0.230** (0.008)	**8.568** (0.068)	**8.245** (0.019)
			With Outliers			
Cox	**0.787** (0.005)	0.456 (0.022)	**0.150** (0.003)	0.351 (0.012)	10.957 (0.164)	11.298 (0.286)
f-DRO	0.628 (0.044)	0.469 (0.088)	0.212 (0.012)	0.246 (0.016)	9.030 (0.148)	8.694 (0.244)
Ours ($\epsilon = 0$)	0.599 (0.045)	0.469 (0.104)	0.224 (0.002)	0.242 (0.008)	8.844 (0.154)	8.397 (0.025)
Ours ($\epsilon = 0.01$)	0.732 (0.019)	0.475 (0.051)	0.189 (0.005)	0.248 (0.009)	8.858 (0.121)	**8.340** (0.059)
Ours ($\epsilon = 0.02$)	0.737 (0.038)	**0.522** (0.061)	0.182 (0.014)	**0.237** (0.024)	**8.819** (0.096)	8.359 (0.078)
Ours ($\epsilon = 0.05$)	0.764 (0.015)	0.458 (0.050)	0.169 (0.010)	0.287 (0.026)	9.275 (0.248)	8.574 (0.153)

Table 6: Simulation results with $d = 10$, $n = 1000$ training samples, and 50% censoring. Values are mean (standard deviation) over 10 runs. Bold and underlined entries denote the best and second-best results within each contamination setting, respectively.

Method ($n = 1000$)	**Overall C-index** ↑	**Worst-group C-index** ↑	**Overall IBS** ↓	**Worst-group IBS** ↓	**Overall tail loss** ↓	**Worst-group tail loss** ↓
			Without Outliers			
Cox	**0.787** (0.007)	0.454 (0.017)	**0.148** (0.005)	0.344 (0.013)	10.877 (0.169)	11.169 (0.233)
f-DRO	0.701 (0.061)	0.523 (0.081)	0.210 (0.010)	0.228 (0.011)	8.722 (0.102)	8.425 (0.135)
Ours ($\epsilon = 0$)	0.693 (0.047)	**0.525** (0.089)	0.217 (0.005)	**0.227** (0.013)	**8.535** (0.075)	**8.264** (0.025)
			With Outliers			
Cox	**0.782** (0.008)	0.449 (0.024)	**0.153** (0.008)	0.340 (0.028)	10.464 (0.368)	10.468 (0.577)
f-DRO	0.557 (0.091)	0.422 (0.077)	0.224 (0.019)	0.246 (0.017)	8.827 (0.168)	8.545 (0.227)
Ours ($\epsilon = 0$)	0.592 (0.104)	0.441 (0.083)	0.223 (0.008)	0.238 (0.008)	**8.503** (0.011)	8.270 (0.021)
Ours ($\epsilon = 0.01$)	0.716 (0.050)	**0.522** (0.060)	0.202 (0.008)	**0.227** (0.008)	8.593 (0.013)	8.249 (0.078)
Ours ($\epsilon = 0.02$)	0.746 (0.025)	0.455 (0.063)	0.188 (0.008)	0.244 (0.016)	8.755 (0.093)	**8.224** (0.017)
Ours ($\epsilon = 0.05$)	0.774 (0.009)	0.460 (0.045)	0.166 (0.006)	0.271 (0.020)	9.135 (0.126)	8.317 (0.119)

A.3.4 Simulation: Integrated Brier score trajectories

Figures 5 and 6 report the integrated Brier score (IBS) across training for the linear and nonlinear simulations, respectively. The settings follow Section 6.1 and Appendix A.3.2. Worst-group IBS is the maximum IBS across the two clean test groups.

At epoch 70, Cox and Cox-NN achieve the lowest overall IBS in their respective settings, while the proposed method with $\epsilon = 0.05$ achieves the lowest worst-group IBS.

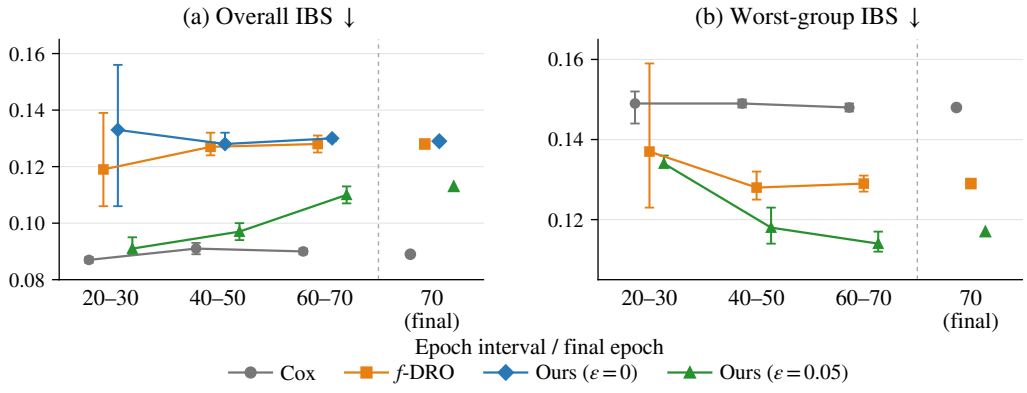

Figure 5: Test IBS under linear risks and outlier contamination. Points show interval means, whiskers show minima and maxima over epochs, and lines connect means. Epoch 70 is shown separately.

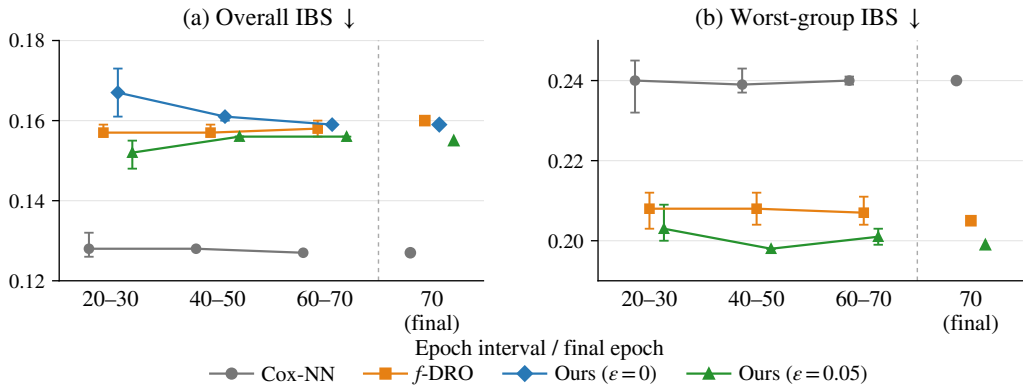

Figure 6: Test IBS under nonlinear risks and outlier contamination, using the plotting conventions in Figure 5.

A.3.5 Benchmark datasets and contamination setup

We describe the benchmark populations, data splits, and contamination settings used in the real-data experiments.

METABRIC The METABRIC dataset contains breast cancer survival data with genomic, transcriptomic, and clinical information (Curtis et al., 2012; Liu et al., 2014). We use 13 clinical covariates. The dataset contains 1,793 samples, among which 764 observations are censored, resulting in a censoring rate of 42.61%.

We define four groups by estrogen receptor (ER) status and age at a threshold of 60 years. Group 0 contains ER-negative patients with age ≤ 60, Group 1 contains ER-negative patients with age > 60, Group 2 contains ER-positive patients with age ≤ 60, and Group 3 contains ER-positive patients with age > 60. The group sizes are 281, 134, 525, and 853, respectively. The corresponding censoring rates are 48.40%, 35.82%, 62.67%, and 29.43%.

FLC The FLC dataset (Dispenzieri et al., 2012) comes from a community cohort of Olmsted County, Minnesota residents aged 50 years or older and was used to study serum free light chains and mortality. We use 6 covariates following Hu & Chen (2024). The dataset contains 7,874 samples, among which 5,705 observations are censored, resulting in a censoring rate of 72.45%.

We define four groups by sex and age at a threshold of 60 years. Group 0 consists of females with age ≤ 60, Group 1 consists of males with age ≤ 60, Group 2 consists of females with age > 60, and Group 3 consists of males with age > 60. The group sizes are 1,768, 1,625, 2,582, and 1,899, respectively. The corresponding censoring rates are 92.82%, 90.09%, 59.80%, and 55.61%.

Data split Each dataset is split into training and test sets at an 80:20 ratio, stratified by group labels and censoring indicators. The training set is then split into training and validation subsets at the same ratio, stratified only by censoring indicators. Group labels are used only for the initial split and worst-group evaluation, not for training or model selection.

Outlier construction Outliers are introduced only into the training data; validation and test sets remain clean.

For linear models, we use a risk-based adversarial construction. We first train a standard Cox model on the training data and compute sample-wise risk scores

$$r_i = f_{\hat{\theta}_{\mathrm{Cox}}}(x_i). \tag{15}$$

We then select the top 2% highest-risk training samples and construct perturbed risks by sign inversion,

$$\widetilde{r}_i = -r_i. \tag{16}$$

We generate replacement survival times from the perturbed hazard model

$$\widetilde{T}_i \sim \widetilde{P}(\cdot \mid x_i), \qquad \widetilde{h}(t \mid x_i) = h_0(t)\exp(\widetilde{r}_i). \tag{17}$$

The generated times replace the selected samples' original times, creating outcomes that conflict with the fitted risk ordering. The Cox model is fitted with an elastic-net penalty (0.01, mixing ratio 0.5) on standardized covariates, so the risk scores are centered; the ranking includes censored and uncensored samples, and the event indicators of the selected samples are unchanged. The baseline hazard is constant, $h_0(t) = \lambda_0$ with $\lambda_0 = \sum_i \delta_i / \sum_i y_i e^{r_i}$, the maximum-likelihood rate given the fitted scores.

For nonlinear models, we use the same risk-based construction with the top 5% highest-risk training samples on both datasets.

A.3.6 Baselines

We compare methods that require no group labels for training or model selection.

- **Elastic-Cox**: A Cox proportional hazards model with elastic-net regularization. We tune the elastic-net regularization strength over $\{0, 0.1, 0.2, 0.4, 0.8\}$.
- **Cox-NN**: A neural-network Cox model with one hidden layer and hidden dimension 2. We tune the Cox-NN regularization strength over $\{0, 0.1, 0.2, 0.4, 0.8\}$.
- **f-DRO** (Hu & Chen, 2024): A survival DRO method based on the χ^2-divergence and sample splitting. We tune the DRO robustness parameter over $\{0.1, 0.2, 0.4, 0.8\}$.
- **Exact-Cox** (Hu & Chen, 2024): The exact DRO-Cox variant, which reparameterizes the Cox model with a piecewise-constant baseline hazard so that the full likelihood decomposes across observations, and applies the χ^2-DRO objective to the resulting per-observation losses without sample splitting. This construction is specific to the Cox model. We tune the DRO robustness parameter over $\{0.1, 0.2, 0.4, 0.8\}$, initialize the baseline-hazard parameters at their maximum-likelihood values for a zero risk score, and use the same training budget and validation criterion as f-DRO.
- **Robust-Cox** (Minder & Bednarski, 1996): A robust proportional hazards estimator for outlier contamination, applied to linear and nonlinear Cox models. We use the R package `coxrobust` for linear models. For nonlinear models, we implement an extension of the original estimator by replacing its linear predictor with the same one-hidden-layer network. Starting from a standard neural Cox fit, we optimize the truncated-weight partial likelihood over three refinement rounds. We tune its truncation parameter over $\{0.90, 0.95, 0.975, 0.99\}$ and hold the remaining settings fixed.

A.3.7 Training and hyperparameter selection

Neural-network training. Cox-NN, f-DRO, Exact-Cox, and the proposed method share a one-hidden-layer network of width 2 (ReLU hidden layer, linear output), except in the network configurations evaluated in Table 3, and all are trained with Adam for 100 full-batch parameter updates ($K = 100$ outer iterations for the proposed method) and evaluated at the final iterate. They differ only in the step size (0.01 for the baselines, 0.1 for the proposed method) and in regularization: Cox-NN tunes an elastic-net penalty over $\{0, 0.1, 0.2, 0.4, 0.8\}$, f-DRO and Exact-Cox use none, and the proposed method fixes it at 0.01. For linear models, f-DRO and Exact-Cox run 100 Adam steps with step size 0.05, whereas the proposed method runs $K = 200$ gradient steps with step size 0.1 and a fixed elastic-net penalty of 0.1. The nonlinear simulation uses the architecture in Appendix A.3.2.

Optimization details of the proposed method. Algorithm 1 uses $T = 1$ inner weight update per outer iteration and $K = 200$ outer iterations. Step sizes are $\eta_q = 0.01$, $\eta_{q'} = 0.1$, and $\eta_\theta = 0.1$; the θ-step is plain gradient descent for the linear model and Adam for the nonlinear model. An elastic-net penalty (strength 0.1 linear, 0.01 nonlinear, mixing ratio 0.5) is added to the weighted objective. Projections onto $\mathcal{Q}_{1-\epsilon}(\bar{p})$ and $\mathcal{Q}_\alpha(q')$ are KL projections onto the capped simplex, $[\Pi(x)]_i = \min\{c_i, \lambda x_i\}$ with c the coordinate-wise upper bounds and λ found by bisection so that the result sums to one.

Validation criteria. Validation and test losses are Cox negative partial log-likelihoods. Elastic-Cox and Cox-NN are selected by maximizing validation log-likelihood; Robust-Cox and the DRO methods by minimizing validation tail loss, the average of the largest 5% or 10% of validation losses (Duchi & Namkoong, 2021). The grids for the proposed method are $\alpha \in \{0.1, 0.2, 0.4, 0.6, 0.8\}$ and $\epsilon \in \{0, 0.05, 0.1\}$; simulation grids are given in the corresponding subsections.

Practical choice of search ranges. Prior knowledge of the smallest meaningful subgroup and of the expected contamination proportion can guide the ranges for α and ϵ; otherwise clustering, anomaly detection, or data-quality diagnostics provide heuristic ranges, and the final pair is selected by validation tail loss.

IBS evaluation. We compute IBS from Breslow survival curves with inverse-probability-of-censoring weights estimated by Kaplan–Meier on the training set. Integration uses 80 equally spaced points from the smallest observed time to the 90th percentile of test times, capped at the maximum training time; test samples beyond training follow-up are excluded only from IBS evaluation.

A.3.8 Real data analysis: Subpopulation shift

Table 7 reports results under subpopulation shift without outliers.

Table 7: Results on METABRIC and FLC under subpopulation shift without outliers. The DRO-based methods use the 10% validation tail loss for model selection; Elastic-Cox and Cox-NN use validation log-likelihood. Values are mean (standard deviation) over 10 independent runs.

METABRIC						
Model	**Method**	**IBS ↓**	**Test log-likelihood ↑**	**Overall C-index ↑**	**Worst-group C-index ↑**	**Test tail loss ↓**
Linear	Elastic-Cox	**0.167** (0.004)	**-5.003** (0.037)	0.664 (0.014)	0.575 (0.047)	6.432 (0.162)
	f-DRO	0.183 (0.005)	-5.117 (0.033)	0.671 (0.014)	0.583 (0.034)	**5.829** (0.017)
	Exact-Cox	0.173 (0.007)	-5.044 (0.064)	0.656 (0.034)	0.550 (0.053)	6.109 (0.085)
	Ours	0.186 (0.004)	-5.128 (0.034)	**0.683** (0.015)	**0.607** (0.029)	**5.829** (0.009)
Nonlinear	Cox-NN	**0.167** (0.005)	**-5.004** (0.043)	**0.670** (0.015)	**0.584** (0.027)	6.537 (0.103)
	f-DRO	0.182 (0.005)	-5.106 (0.032)	0.666 (0.028)	0.574 (0.041)	**5.841** (0.054)
	Exact-Cox	0.179 (0.006)	-5.102 (0.040)	0.627 (0.034)	0.525 (0.083)	6.056 (0.053)
	Ours	0.179 (0.004)	-5.078 (0.028)	0.655 (0.026)	0.575 (0.042)	5.912 (0.034)
FLC						
Model	**Method**	**IBS ↓**	**Test log-likelihood ↑**	**Overall C-index ↑**	**Worst-group C-index ↑**	**Test tail loss ↓**
Linear	Elastic-Cox	**0.093** (0.004)	**-6.467** (0.071)	0.797 (0.004)	0.629 (0.039)	8.087 (0.262)
	f-DRO	0.111 (0.012)	-6.801 (0.200)	**0.798** (0.005)	0.629 (0.038)	7.400 (0.091)
	Exact-Cox	0.115 (0.007)	-6.875 (0.112)	0.792 (0.007)	0.620 (0.038)	7.323 (0.065)
	Ours	0.120 (0.002)	-6.956 (0.022)	**0.798** (0.004)	**0.631** (0.036)	**7.283** (0.006)
Nonlinear	Cox-NN	**0.092** (0.001)	**-6.452** (0.028)	**0.797** (0.005)	0.608 (0.026)	8.268 (0.036)
	f-DRO	0.109 (0.011)	-6.765 (0.183)	**0.797** (0.006)	**0.623** (0.032)	7.434 (0.097)
	Exact-Cox	0.116 (0.003)	-6.890 (0.041)	0.774 (0.021)	0.556 (0.078)	**7.330** (0.024)
	Ours	0.114 (0.003)	-6.839 (0.048)	0.780 (0.028)	0.605 (0.051)	7.357 (0.061)

The proposed method reduces test tail loss relative to the standard Cox models in all settings, while these baselines achieve better IBS and test log-likelihood. It attains the highest overall and worst-

group C-index on METABRIC with linear models; on FLC, its worst-group C-index is within 0.003 of the standard Cox models.

Rationale for controlled contamination. Luo et al. (2022) model survival outliers through individual changes in hazard rates, allowing survival times to be unusually long or short relative to predictions based on covariates. Our constructions introduce such departures at a known contamination proportion while preserving the covariates and clinically defined evaluation groups; their effects can differ across datasets and model classes. All methods use the same contamination construction within each setting. Appendix A.3.9 evaluates an alternative that appends perturbed observations while preserving the original training samples.

A.3.9 Additional outlier setting

We evaluate linear models on FLC after appending perturbed copies of short-survival event samples, preserving the original training data. Let

$$\mathcal{E} = \{i : z_i \in \mathcal{D}_{\text{train}},\ \delta_i = 1\} \tag{18}$$

denote the index set of uncensored event samples in the training data. We sort the samples indexed by $\mathcal{E}$ by their observed survival times y_i in ascending order and select the lowest 2%. Specifically, we define

$$\mathcal{O}_{\text{low}} = \{i \in \mathcal{E} : y_i \le c_{0.02}\}\,. \tag{19}$$

Here, $c_{0.02}$ is the empirical 2nd percentile of the observed event times among samples in $\mathcal{E}$.

Each selected sample $i \in \mathcal{O}_{\text{low}}$ is duplicated with its covariates and event indicator unchanged, and its survival time resampled near the median event time:

$$\widetilde{x}_i = x_i, \qquad \widetilde{\delta}_i = \delta_i, \qquad \widetilde{y}_i \sim \mathcal{U}(m_E - \tau, m_E + \tau), \tag{20}$$

where m_E is the median observed survival time among event samples and $\tau = 0.1 m_E$. The resulting duplicated sample is denoted by

$$\widetilde{z}_i = (\widetilde{x}_i, \widetilde{y}_i, \widetilde{\delta}_i). \tag{21}$$

The contaminated training set is

$$\widetilde{\mathcal{D}}_{\text{train}} = \mathcal{D}_{\text{train}} \cup \{\widetilde{z}_i : i \in \mathcal{O}_{\text{low}}\}. \tag{22}$$

The appended samples retain covariates from short-survival cases but have times near the median, creating inconsistencies between covariates and outcomes.

Table 8: Linear-model results on FLC with appended outliers. The 10% and 5% settings specify the validation tail loss for the DRO methods; Elastic-Cox uses validation log-likelihood. Values are mean (standard deviation) over 10 independent runs.

		FLC				
Model	**Method**	**IBS ↓**	**Test log-likelihood ↑**	**Overall C-index ↑**	**Worst-group C-index ↑**	**Test tail loss ↓**
Linear (10%)	Elastic-Cox	**0.091** (0.001)	**-6.430** (0.025)	**0.798** (0.004)	0.618 (0.030)	7.520 (0.017)
	f-DRO	0.115 (0.016)	-7.377 (1.157)	0.728 (0.106)	0.579 (0.092)	8.933 (2.887)
	Ours	0.113 (0.002)	-6.842 (0.031)	0.797 (0.004)	**0.644** (0.028)	**7.180** (0.009)
Linear (5%)	Elastic-Cox	**0.091** (0.001)	**-6.430** (0.025)	**0.798** (0.004)	0.618 (0.030)	7.520 (0.017)
	f-DRO	0.129 (0.006)	-7.629 (0.011)	0.688 (0.089)	0.563 (0.087)	10.177 (4.925)
	Ours	0.120 (0.002)	-6.970 (0.018)	0.797 (0.005)	**0.646** (0.031)	**7.283** (0.004)

In Table 8, Elastic-Cox achieves the best IBS, test log-likelihood, and overall C-index. The proposed method achieves the best worst-group C-index and test tail loss under both validation criteria, with overall C-index close to Elastic-Cox. For f-DRO, the 5% validation criterion yields lower overall C-index and higher test tail loss than the 10% criterion.

A.3.10 Contamination sensitivity

Tables 1 and 2 use the same risk-inversion construction (Appendix A.3.5) on both datasets, with 2% contamination for linear models and 5% for nonlinear models. To examine sensitivity to the

contamination rate, Table 9 evaluates all four model–dataset combinations at both rates, using the 10% validation setting of Table 1. The (linear, 2%) and (nonlinear, 5%) entries coincide with the 10% validation results in Tables 1 and 2. At 5%, the proposed method attains the highest worst-group C-index in all four settings. At 2%, it leads in the linear settings but trails the best nonlinear baselines. Its test tail loss is the lowest or second lowest in every setting.

Table 9: Worst-group C-index and 5% test tail loss under the risk-based contamination applied to all model classes and datasets at contamination rates of 2% and 5%. Hyperparameters are selected by the 10% validation tail loss for Robust-Cox, f-DRO, Exact-Cox, and the proposed method and by validation log-likelihood for Elastic-Cox and Cox-NN. Values are mean (standard deviation).

		2% contamination		5% contamination	
Model	**Method**	**Worst-group C-index ↑**	**Test tail loss ↓**	**Worst-group C-index ↑**	**Test tail loss ↓**
		METABRIC			
Linear	Elastic-Cox	0.501 (0.058)	6.305 (0.111)	0.466 (0.064)	6.195 (0.178)
	Robust-Cox	0.562 (0.042)	6.229 (0.048)	0.520 (0.052)	6.117 (0.091)
	f-DRO	0.574 (0.030)	**5.825** (0.015)	0.567 (0.031)	**5.822** (0.015)
	Exact-Cox	0.459 (0.057)	6.179 (0.164)	0.393 (0.032)	6.267 (0.176)
	Ours	**0.613** (0.028)	5.835 (0.011)	**0.609** (0.028)	5.835 (0.008)
Nonlinear	Cox-NN	0.572 (0.038)	6.404 (0.084)	0.508 (0.091)	6.561 (0.310)
	Robust-Cox	**0.586** (0.040)	6.351 (0.077)	0.545 (0.049)	6.295 (0.161)
	f-DRO	0.559 (0.043)	**5.832** (0.049)	0.533 (0.047)	**5.839** (0.041)
	Exact-Cox	0.419 (0.105)	6.139 (0.100)	0.395 (0.078)	6.226 (0.178)
	Ours	0.571 (0.039)	5.901 (0.029)	**0.585** (0.040)	5.902 (0.014)
		FLC			
Linear	Elastic-Cox	0.618 (0.038)	7.650 (0.114)	0.602 (0.035)	7.560 (0.128)
	Robust-Cox	0.611 (0.040)	7.854 (0.031)	0.594 (0.040)	7.626 (0.030)
	f-DRO	0.626 (0.037)	7.409 (0.035)	0.618 (0.035)	7.384 (0.048)
	Exact-Cox	0.609 (0.037)	7.581 (0.127)	0.584 (0.044)	7.506 (0.173)
	Ours	**0.631** (0.036)	**7.283** (0.006)	**0.632** (0.036)	**7.284** (0.006)
Nonlinear	Cox-NN	0.614 (0.037)	7.917 (0.183)	0.623 (0.029)	7.801 (0.208)
	Robust-Cox	**0.626** (0.041)	8.071 (0.173)	0.602 (0.030)	8.134 (0.324)
	f-DRO	**0.626** (0.031)	7.457 (0.035)	0.624 (0.028)	7.447 (0.043)
	Exact-Cox	0.582 (0.057)	7.603 (0.209)	0.561 (0.081)	7.622 (0.188)
	Ours	0.613 (0.040)	**7.359** (0.064)	**0.630** (0.033)	**7.371** (0.067)

A.3.11 COMPUTATIONAL RUNTIME

Table 10 compares model-fitting times for linear Cox models on METABRIC and FLC. For this comparison, all optimization-based methods, including the proposed method, are run for 100 iterations, separately from the training schedules in Appendix A.3.7. The proposed method is faster than f-DRO on both datasets, although standard Cox and Exact-Cox are faster. With the $K = 200$ schedule of Appendix A.3.7, it takes 0.631 and 1.283 seconds on METABRIC and FLC, respectively. The corresponding complexity analysis is given in Appendix A.1.3.

Table 10: Model-fitting time in seconds, reported as mean (standard deviation) over 10 repetitions on a single CPU thread. All optimization-based methods are run for 100 iterations in this comparison.

Dataset	**Cox**	**Elastic-Cox**	**f-DRO**	**Exact-Cox**	**Ours**
METABRIC	0.104 (0.005)	1.239 (0.040)	1.129 (0.281)	0.195 (0.004)	0.368 (0.008)
FLC	0.356 (0.133)	2.210 (0.796)	18.122 (0.531)	0.236 (0.007)	0.817 (0.135)