Stability Evaluation through Distributional Perturbation Analysis

Jose Blanchet [1 *] **Peng Cui** [2 *] **Jiajin Li** [1 *] **Jiashuo Liu** [1 2 *]

Abstract

The performance of learning models often deteriorates when deployed in out-of-sample environments. To ensure reliable deployment, we propose a stability evaluation criterion based on distributional perturbations. Conceptually, our stability evaluation criterion is defined as the minimal perturbation required on our observed dataset to induce a prescribed deterioration in risk evaluation. In this paper, we utilize the optimal transport (OT) discrepancy with moment constraints on the *(sample, density)* space to quantify this perturbation. Therefore, our stability evaluation criterion can address both data corruptions and sub-population shifts—the two most common types of distribution shifts in real-world scenarios. To further realize practical benefits, we present a series of tractable convex formulations and computational methods tailored to different classes of loss functions. The key technical tool to achieve this is the strong duality theorem provided in this paper. Empirically, we validate the practical utility of our stability evaluation criterion across a host of real-world applications. These empirical studies showcase the criterion's ability not only to compare the stability of different learning models and features but also to provide valuable guidelines and strategies to further improve models.

1. Introduction

The issue of poor out-of-sample performance frequently arises, particularly in high-stakes applications such as healthcare (Bandi et al., 2018; Wynants et al., 2020; Roberts et al., 2021), economics (Hand, 2006; Ding et al., 2021), self-driving (Malinin et al., 2021; Hell et al., 2021). This phenomenon can be attributed to discrepancies between the training and test datasets, influenced by various factors. Some of these factors include measurement errors during data collection (Jacobucci & Grimm, 2020; Elmes et al., 2020), deployment in dynamic, non-stationary environments (Camacho & Conover, 2011; Conger et al., 2023), and the under-representativeness of marginalized groups in the training data (Corbett-Davies et al., 2023), among others. The divergence between training and test data presents substantial challenges to the reliability, robustness, and fairness of machine learning models in practical settings. Recent empirical studies have shown that algorithms intentionally developed for addressing distribution shifts—such as distributionally robust optimization (Blanchet et al., 2019; Sagawa et al., 2019; Kuhn et al., 2019; Duchi & Namkoong, 2021; Rahimian & Mehrotra, 2022; Blanchet et al., 2024), domain generalization (Zhou et al., 2022), and causally invariant learning (Arjovsky et al., 2019; Krueger et al., 2021) — experience a notable performance degradation when faced with real-world scenarios (Gulrajani & Lopez-Paz, 2020; Frogner et al., 2021; Yang et al., 2023; Liu et al., 2023).

Instead of providing a robust training algorithm, we shift focus towards a more fundamental (in some sense even simpler) question:

Q: How do we evaluate the stability of a learning model when subjected to data perturbations?

To answer this question, our initial step is to gain a comprehensive understanding of various types of data perturbations. In this paper, we categorize data perturbations into two classes: (i) *Data corruptions*, which encompass changes in the distribution support (i.e., observed data samples). These changes can be attributed to measurement errors in data collection or malicious adversaries. Typical examples include factors like street noises in speech recognition (Kinoshita et al., 2020), rounding errors in finance (Li & Mykland, 2015), adversarial examples in vision (Goodfellow et al., 2020) and, the Goodhart's law empirically observed in government assistance allocation (Camacho & Conover, 2011). (ii) *Sub-population shifts*, refer to perturbation on the probability density or mass function while keeping the same support. For example, model performances substantially degrade under demographic shifts in recommender

[*]Equal contribution [1]Department of Management Science and Engineering, Stanford University [2]Department of Computer Science and Technology, Tsinghua University. Correspondence to: Jiajin Li <jiajinli@stanford.edu>, Jiashuo Liu <liujiashuo77@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

systems (Blodgett et al., 2016; Sapiezynski et al., 2017); under temporal shifts in medical diagnosis (Pasterkamp et al., 2017); and under spatial shifts in wildlife conservation (Beery et al., 2021).

Recent investigation on the question **Q** predominantly centers around sub-population shifts, see (Li et al., 2021; Namkoong et al., 2022; Gupta & Rothenhaeusler, 2023). However, in practical scenarios, it is common to encounter both types of data perturbation. Studies such as Gokhale et al. (2022) and Zou & Liu (2023) have documented that models demonstrating robustness against sub-population shifts can still be vulnerable to data corruptions. This underscores the importance of adopting a more holistic approach when evaluating model stability, one that addresses both sub-population shifts and data corruptions.

To fully answer the question **Q**, we frame the model stability as a projection problem over probability space under the OT discrepancy with moment constraints. Specifically, we seek the minimum perturbation necessary on our reference measure (i.e., observed data) to guarantee that the model's risk remains below a specified threshold. The crux of our approach is to conduct this projection within the joint *(sample, density)* space. Consequently, our stability metric is capable of addressing both data corruptions on the sample space and sub-population shifts on the density or probability mass space. To enhance the practical utility of our approach, we present a host of tractable convex formulations and computational methods tailored to different learning models. The key technical tool for this is the strong duality theorem provided in this paper.

To offer clearer insights, we visualize the most sensitive distribution in stylized examples. Our approach achieves a balanced and reasoned stance by avoiding overemphasis on specific samples or employing overly aggressive data corruptions. Moreover, we demonstrate the practical effectiveness of our proposed stability evaluation criterion by applying it to tasks related to income prediction, health insurance prediction, and COVID-19 mortality prediction. These real-world scenarios showcase the framework's capacity to assess stability across various models and features, uncover potential biases and fairness issues, and ultimately enhance decision-making.

Notations. Throughout this paper, we let $\mathbb{R}$ denote the set of real numbers, $\mathbb{R}_+$ denote the subset of non-negative real numbers. We use capitalized letters for random variables, e.g., X, Y, Z, and script letters for the sets, e.g., $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$. For any close set $\mathcal{Z} \subset \mathbb{R}^d$, we define $\mathcal{P}(\mathcal{Z})$ as the family of all Borel probability measures on $\mathcal{Z}$. For $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$, we use the notation $\mathbb{E}_{\mathbb{P}}[\cdot]$ to denote expectation with respect to the probability distribution $\mathbb{P}$. For the prediction problem, the random variable of data points is denoted by $Z = (X, Y) \in \mathcal{Z}$, where $X \in \mathcal{X}$ denotes the input covariates, $Y \in \mathcal{Y}$ denotes the target. $f_\beta : \mathcal{X} \to \mathcal{Y}$ denotes the prediction model parameterized by β. The loss function is denoted as $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, and $\ell(f_\beta(X), Y)$ is abbreviated as $\ell(\beta, Z)$. We use $(\cdot)_+ = \max(\cdot, 0)$. We adopt the conventions of extended arithmetic, whereby $\infty \cdot 0 = 0 \cdot \infty = 0/0 = 0$ and $\infty - \infty = -\infty + \infty = 1/0 = \infty$.

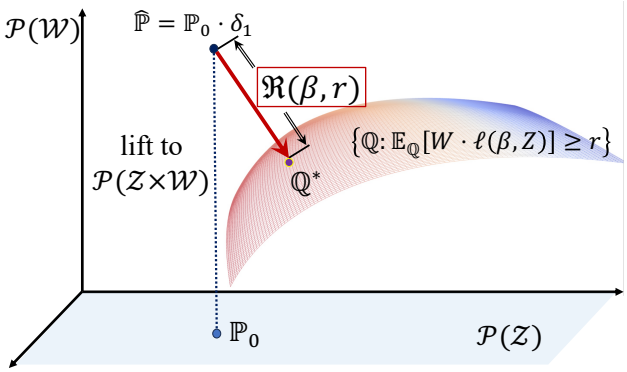


Figure 1: Data Distribution Projection

2. Model Evaluation Framework

In this section, we present a stability evaluation criterion based on OT discrepancy with moment constraints, capable of considering both types of data perturbation — data corruptions and sub-population shifts — in a unified manner. The key insight lies in computing the projection distance, as shown in Figure 1, which involves minimizing the probability discrepancy between the most sensitive distribution denoted as $\mathbb{Q}^\star$ and the lifted training distribution $\mathbb{P}_0 \otimes \delta_1$ in the joint (sample, density) space, while maintaining the constraint that the model performance falls below a specific threshold. This threshold refer to a specific level of risk, error rate, or any other relevant performance metrics. The projection type methodology has indeed been employed in the literature for statistical inference, particularly in tasks like constructing confidence regions (Owen, 2001; Blanchet et al., 2019). However, this application is distinct from our current purpose.

2.1. OT-based stability evaluation criterion

We begin by presenting the OT discrepancy with moment constraints, as proposed in Blanchet et al. (2023, Definition 2.1). This serves as a main technical tool for our further discussions.

Definition 1 (OT discrepancy with moment constraints)**.** If $\mathcal{Z} \subseteq \mathbb{R}^d$ and $\mathcal{W} \subseteq \mathbb{R}_+$ are convex and closed sets, $c : (\mathcal{Z} \times \mathcal{W})^2 \to \mathbb{R}_+$ is a lower semicontinuous function, and $\mathbb{Q}, \mathbb{P} \in \mathcal{P}(\mathcal{Z} \times \mathcal{W})$, then the OT discrepancy with moment constraints induced by c, $\mathbb{Q}$ and $\mathbb{P}$ is the function $\mathbb{M}_c$:

$\mathcal{P}(\mathcal{Z} \times \mathcal{W})^2 \to \mathbb{R}_+$ defined through

$$\mathbb{M}_c(\mathbb{Q}, \mathbb{P}) = \left\{ \begin{array}{ll} \inf & \mathbb{E}_\pi[c((Z,W),(\hat{Z},\hat{W}))] \\ \text{s. t.} & \pi \in \mathcal{P}((\mathcal{Z} \times \mathcal{W})^2) \\ & \pi_{(Z,W)} = \mathbb{Q},\ \pi_{(\hat{Z},\hat{W})} = \mathbb{P} \\ & \mathbb{E}_\pi[W] = 1 \quad \pi\text{-a.s,} \end{array} \right.$$

where $\pi_{(Z,W)}$ and $\pi_{(\hat{Z},\hat{W})}$ are the marginal distributions of (Z, W) and $(\hat{Z}, \hat{W})$ under π.

Remark 1. The core idea is to lift the original sample space $\mathcal{Z}$ to a higher dimensional space $\mathcal{Z} \times \mathcal{W}$ —- a joint (sample, density) space. Here, we treat the additional random variable W as the "density " or "probability mass", making it also amenable to perturbations through optimal transport methods. However, these perturbations are subject to the constraint that the expectation of the density must remain equal to one. Thus, the transportation cost function $c((z, w), (\hat{z}, \hat{w}))$ can measure the changes in both samples ($\hat{z} \to z$) and their probability densities ($\hat{w} \to w$).

To evaluate the stability of a given learning model f_β trained on the distribution $\mathbb{P}_0 \in \mathcal{P}(\mathcal{Z})$, we formally introduce the OT-based stability evaluation criterion as

$$\mathfrak{R}(\beta, r) = \left\{ \begin{array}{ll} \inf\limits_{\mathbb{Q} \in \mathcal{P}(\mathcal{Z} \times \mathcal{W})} & \mathbb{M}_c(\mathbb{Q}, \hat{\mathbb{P}}) \\ \text{s.t.} & \mathbb{E}_\mathbb{Q}[W \cdot \ell(\beta, Z)] \geq r. \end{array} \right. \tag{P}$$

Here, the reference measure $\hat{\mathbb{P}}$ is selected as $\mathbb{P}_0 \otimes \delta_1$, with δ_1 denoting the Dirac delta function,[1] $\mathbb{M}_c(\mathbb{Q}, \hat{\mathbb{P}})$ represents the OT discrepancy with moment constraints between the projected distribution $\mathbb{Q}$ and the reference distribution $\hat{\mathbb{P}}$, $\ell(\beta, z)$ denotes the prediction risk of model f_β on sample z, and $r > 0$ is the pre-defined risk threshold.

To sum up, we evaluate a model's stability under distribution shifts by quantifying the minimum level of perturbations required for the model's performance to degrade to a predetermined risk threshold. The magnitude of perturbations is determined through the use of the OT discrepancy with moment constraints and the cost function c, see definition 1.

Then, a natural question arises: How do we select the cost function c to effectively quantify the various types of perturbations? We aim for this cost function to be capable of quantifying changes in both the support of the distribution and the probability density or mass function. One possible candidate cost function is

$$c((z, w), (\hat{z}, \hat{w})) = \theta_1 \cdot w \cdot d(z, \hat{z}) + \theta_2 \cdot (\phi(w) - \phi(\hat{w}))_+. \tag{1}$$

Here, $d(z, \hat{z}) = \|x - \hat{x}\|_2^2 + \infty \cdot |y - \hat{y}|$ quantifies the cost associated with the different data samples z and $\hat{z}$ in the set $\mathcal{Z}$, with the label measurement's reliability considered infinite; $(\phi(w) - \phi(\hat{w}))_+$ denotes the cost related to differences in probability mass, where $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a convex function satisfying $\phi(1) = 0$; $\theta_1, \theta_2 \geq 0$ serve as two hyperparameters, satisfying $1/\theta_1 + 1/\theta_2 = C$ for some *constant* C, to control the trade-off between the cost of perturbing the distribution's support and the probability density or mass on the observed data points. This cost function was originally proposed in Blanchet et al. (2023, Section 5) within the framework of distributionally robust optimization.

Remark 2 (Effect of θ_1 and θ_2). (i) When $\theta_1 = +\infty$, the stability criterion $\mathfrak{R}(\beta, r)$ only counts the sub-population shifts, as any data sample corruptions are not allowed. In this scenario, our proposed stability criterion can be reduced to the one recently introduced in Gupta & Rothenhaeusler (2023) and Namkoong et al. (2022). (ii) When $\theta_2 = +\infty$, the stability criterion $\mathfrak{R}(\beta, r)$ only takes the data corruptions into account instead. (iii) The most intriguing scenario arises when both θ_1 and θ_2 have finite values. These parameters, θ_1 and θ_2, hold a pivotal role in adjusting the balance between data corruptions and sub-population shifts within our stability criterion, which allows us to simultaneously consider both types of distribution shifts. By manipulating the values of θ_1 and θ_2, we can achieve a versatile representation of a model's resilience across a wide range of distributional perturbation directions. This adaptability carries significant implications when evaluating the robustness of models in diverse and ever-evolving real-world environments.

2.2. Dual reformulation and its interpretation

Problem (P) constitutes an infinite-dimensional optimization problem over probability distributions and thus appears to be intractable. However, we will now demonstrate that by first establishing a strong duality result, problem (P) can be reformulated as a finite-dimensional optimization problems and discuss the structure of the most sensitive distribution from problem (P).

Theorem 1 (Strong duality for problem (P))**.** *Suppose that (i) The set $\mathcal{Z} \times \mathcal{W}$ is compact, (ii) $\ell(\beta, \cdot)$ is upper semi-continuous for all β, (iii) the cost function $c : (\mathcal{Z} \times \mathcal{W})^2 \to \mathbb{R}_+$ is continuous; and (iv) the risk level r is less than the worst case value $\bar{r} := \max_{z \in \mathcal{Z}} \ell(\beta, z)$. Then, we have*

$$\mathfrak{R}(\beta, r) = \sup_{h \in \mathbb{R}_+, \alpha \in \mathbb{R}} hr + \alpha + \mathbb{E}_{\hat{\mathbb{P}}}\left[\tilde{\ell}_c^{\alpha,h}(\beta, (\hat{Z}, \hat{W}))\right] \tag{D}$$

where the surrogate function $\tilde{\ell}_c^{\alpha,h}(\beta, (\hat{z}, \hat{w}))$ equals to

$$\min_{(z,w) \in \mathcal{Z} \times \mathcal{W}} c((z, w), (\hat{z}, \hat{w})) + \alpha w - h \cdot w \cdot \ell(\beta, z),$$

for all $\hat{z} \in \mathcal{Z}$ and $\hat{w} \in \mathcal{W}$.

For a detailed proof, we direct interested readers to the Appendix A.1.

[1] This implies that the sample weights are almost surely equal to one with respect to the reference distribution, as we lack any prior information about them.

Remark 3. When the reference measure $\mathbb{P}_0$ is a discrete measure, some technical conditions in Theorem 1 (e.g., compactness, (semi)-continuity) can be eliminated by utilizing the abstract semi-infinite duality theory for conic linear programs. Please refer to Shapiro (2001, Proposition 3.4) and our proof in Appendix A.1 for more detailed information.

If we adopt the cost function in the form of (1) for two commonly used ϕ functions, we can simplify the surrogate function further by obtaining the closed form of w. Here, we explore the following cases: (i) Selecting $\phi(t) = t\log t - t + 1$, which is associated with the Kullback–Leibler (KL) divergence. (ii) Choosing $\phi(t) = (t-1)^2$, which is linked to the χ^2-divergence.

Proposition 1 (Dual reformulations)**.** *Suppose that $\mathcal{W} = \mathbb{R}_+$. (i) If $\phi(t) = t\log t - t + 1$, then the dual problem (D) admits:*

$$\sup_{h\geq 0} hr - \theta_2 \log \mathbb{E}_{\mathbb{P}_0}\left[\exp\left(\frac{\ell_{h,\theta_1}(\hat{Z})}{\theta_2}\right)\right]; \tag{2}$$

(ii) If $\phi(t) = (t-1)^2$, then the dual problem (D) admits:

$$\sup_{h\geq 0,\alpha\in\mathbb{R}} hr + \alpha + \theta_2 - \theta_2\mathbb{E}_{\mathbb{P}_0}\left[\left(\frac{\ell_{h,\theta_1}(\hat{Z}) + \alpha}{2\theta_2} + 1\right)_+^2\right], \tag{3}$$

where the d-transform of $h \cdot \ell(\beta, \cdot)$ with the step size θ_1 is defined as

$$\ell_{h,\theta_1}(\hat{z}) := \max_{z\in\mathcal{Z}} h \cdot \ell(\beta, z) - \theta_1 \cdot d(z, \hat{z}).$$

When the reference measure $\mathbb{P}_0$ is represented as the empirical measure $\mathbb{P}_0 = \frac{1}{n}\sum_{i=1}^n \delta_{\hat{z}_i}$, the characterization of the most sensitive distribution $\mathbb{Q}^\star$, can be elucidated through the dual formulation provided in (2) and (3).

Remark 4 (Structure of the most sensitive distribution). We express $\mathbb{Q}^\star$ as follows: $\mathbb{Q}^\star = \frac{1}{n}\sum_{i=1}^n \delta_{(z_i^\star, w_i^\star)}$, where each $(z_i^\star, w_i^\star) \in \mathcal{Z} \times \mathbb{R}_+$ satisfies the conditions:

$$z_i^\star = \arg\max_{z\in\mathcal{Z}} h^\star \ell(\beta; z) - \theta_1 \cdot d(z, \hat{z}_i), \quad \forall i \in [n].$$

Using various ϕ functions requires adjusting the weight in a distinct manner:
(i) If $\phi(t) = \log t - t + 1$, then we have:

$$w_i^\star \propto \exp\left(\frac{\ell_{h^\star,\theta_1}(\hat{z}_i)}{\theta_2}\right), \quad \forall i \in [n];$$

(ii) If $\phi(t) = (t-1)^2$, then we have:

$$w_i^\star \propto \left(\frac{\ell_{h^\star,\theta_1}(\hat{z}_i) - \alpha^*}{2\theta_2} + 1\right)_+, \quad \forall i \in [n],$$

where $h^\star$ and $\alpha^\star$ are the optimal solution of problem (D). Therefore, it becomes evident that the most sensitive distribution encompasses both aspects of shifts: the transformation from $\hat{z}_i$ to $z_i^\star$ and the reweighting from $\frac{1}{n}$ to $w_i^\star$. Our cost function enables a versatile evaluation of model stability across a range of distributional perturbation directions. This approach yields valuable insights into the behavior of a model in different real-world scenarios and underscores the importance of incorporating both types of distributional perturbation in stability evaluation.

2.3. Computation

In this subsection, our emphasis lies in addressing problems (2) and (3) with varying types of loss functions, specifically when the reference measure $\mathbb{P}_0$ takes the form of the empirical distribution.

Convex piecewise linear loss functions. If the loss function $\ell(\beta, \cdot)$ is piecewise linear (e.g., linear SVM), we can show that (2) admits a tractable finite convex program.

Theorem 2 (KL divergence)**.** *Suppose that $\mathcal{Z} = \mathbb{R}^d \times \{+1, -1\}$ and $\ell(\{(a_k, b_k)\}_{k\in[K]}, z) = \max_{k\in[K]} y \cdot a_k^\top x + b_k$. The negative optimal value of problem (2) is equivalent to the optimal value of the finite convex program:*

$$\begin{array}{ll} \min & -hr + t \\ \text{s.t.} & \lambda \in \mathbb{R}_+, t \in \mathbb{R}, \eta \in \mathbb{R}_+^n, p \in \mathbb{R}^n \\ & (\eta_i, \theta_2, p_i - t) \in \mathcal{K}_{\exp} \\ & \frac{\|a_k\|_2^2}{4\theta_1} h^2 + \hat{y}_i \cdot a_k^T \hat{x}_i \cdot h + b_k \leq p_i \\ & \frac{1}{n}\sum_{i=1}^n \eta_i \leq \theta_2, \end{array}$$

where the set $\mathcal{K}_{\exp}$ is the exponential cone.

Theorem 3 (χ^2 Divergnce)**.** *Suppose that $\mathcal{Z} = \mathbb{R}^d \times \{+1, -1\}$ and $\ell(\{(a_k, b_k)\}_{k\in[K]}, z) = \max_{k\in[K]} y \cdot a_k^\top x + b_k$. The negative optimal value of problem (2) is equivalent to the optimal value of the finite convex program*

$$\begin{array}{ll} \min & -hr + t \\ \text{s.t.} & h \in \mathbb{R}_+, \alpha \in \mathbb{R}, t \in \mathbb{R}, \eta \in \mathbb{R}_+^n \\ & \frac{\|a_k\|_2^2}{4\theta_1} \cdot h^2 + \hat{y}_i \cdot a_k^T \hat{x}_i \cdot h + b_k + 2\theta_2\alpha + 2\theta_2 \leq 2\theta_2\eta_i \\ & \frac{\theta_2}{n}\sum_{i=1}^n \eta_i^2 \leq t. \end{array}$$

For a detailed proof, we direct interested readers to the Appendix A.3 and A.4 for more detailed information. Equipped with Theorem 2 and 3, we can calculate our evaluation criterion by general purpose conic optimization solvers such as MOSEK and GUROBI.

0/1 loss function. In practical applications, employing a 0/1 loss function offers users a simpler method to set up the risk level r, which corresponds to a pre-defined acceptable level of *error rate*. That is, given a trained model β, we

define the loss function on the sample (x, y) as

$$\ell(\beta, (x, y)) = \mathbb{I}_{y \neq f_\beta(x)},$$

where $\mathbb{I}$ is the indicator function defined as $\mathbb{I}_{y \neq f_\beta(x)} = 0$ if $y \neq f_\beta(x)$; $= 0$ otherwise. In this scenario, the d-transform of $h \cdot \ell_\beta(\cdot)$ can be expressed in a closed form. Conceptually, this loss function promotes long-haul transportation, as it encourages either minimal perturbation or no movement at all, i.e.,

$$\ell_{h,\theta_1}(\hat{z}) = (h - \theta_1 \cdot d^\star(\hat{z}))_+,$$

where $d^\star(\hat{z}) := \min_{z \in Z}\{d(z, \hat{z}) : \ell(\beta, z) = 1\}$. This distance quantifies the minimal adjustment needed to fool or mislead the classifier's prediction for the sample $\hat{z}$. A similar formulation has been employed in Si et al. (2021) to assess group fairness through optimal transport projections. Finally, the dual formulation (2) is reduced to an one-dimensional convex problem w.r.t h.

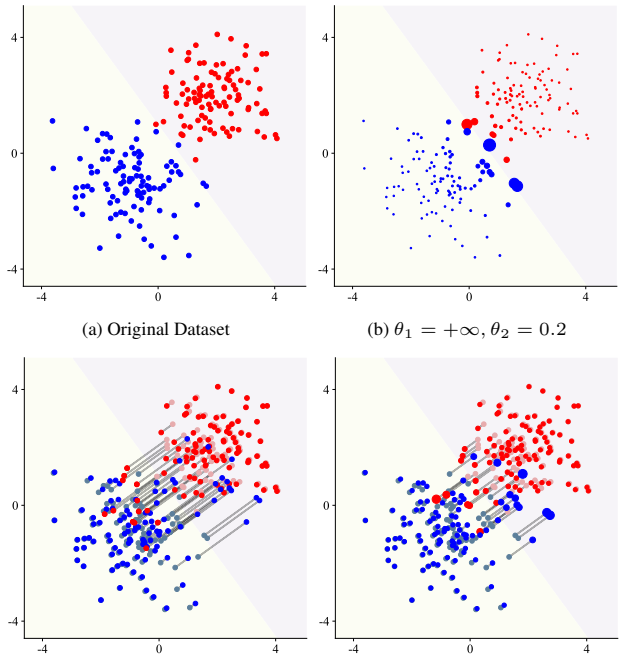
Nonlinear loss functions. For general nonlinear loss functions, such as those encountered in deep neural networks, the dual formulation (2) retains its one-dimensional convex nature with respect to h. However, the primary computational challenge lies in solving the inner maximization problem concerning the sample z. In essence, this dual maximization problem (2) for nonlinear loss functions is closely associated with adversarial training (Nouiehed et al., 2019; Yi et al., 2021). All algorithms available in the literature for this purpose can be applied to our problem as well. The key distinction lies in the outer loop. In our case, we optimize over $h \in \mathbb{R}_+$ to perturb the sample weights, whereas in adversarial training, this outer loop is devoted to the training of model parameters.

For simplicity, we adopt a widely-used approach in our paper: Performing multiple gradient ascent steps to generate adversarial examples, followed by an additional gradient ascent step over h. For a more thorough understanding, please see Algorithm 1. If we can solve the inner maximization problem nearly optimally, then we can ensure that the sequence generated by Algorithm 1 converges to the global optimal solution. You can find further details in Sinha et al. (2018, Theorem 2).

2.4. Feature stability analysis

As an additional benefit, if we select an alternative cost function, different from the one proposed in (1), our evaluation criterion $\mathfrak{R}(\beta, r)$ can serve as an effective metric for assessing feature stability within machine learning models. If we want to evaluate the stability of the i-th feature, we can modify the distance function d in (1) as

$$d(z, \hat{z}) = \|z_{(i)} - \hat{z}_{(i)}\|_2^2 + \infty \cdot \|z_{(,-i)} - \hat{z}_{(,-i)}\|_2^2,$$



(a) Original Dataset (b) $\theta_1 = +\infty, \theta_2 = 0.2$ (c) $\theta_1 = 0.2, \theta_2 = +\infty$ (d) $\theta_1 = \theta_2 = 0.4$

Figure 2: Visualizations of the original dataset and the most sensitive distribution $\mathbb{Q}^\star$ produced by cross-entropy loss function under different θ_1, θ_2. The original prediction error is 0.1, and the risk threshold is 0.5.

where $z_{(i)}$ represents the i-th feature of z, while $z_{(,-i)} = z \backslash z_{(i)}$ denotes all variables in z except for the i-th one. This implies that during evaluation, we are only permitted to perturb the i-th feature while keeping all other features unchanged.

Substituting $d(z, \hat{z})$ in problem (2), we could obtain the corresponding feature stability criterion $\mathfrak{R}_i(\beta, r)$, which provides a quantitative stability evaluation of how robust the model is with respect to changes in the i-th feature. Specifically, a higher value of $\mathfrak{R}_i(\beta, r)$ indicates greater stability of the corresponding feature against potential shifts.

3. Visualizations on stylized / toy examples

In this section, we use a toy example to visualize the most sensitive distribution $\mathbb{Q}^\star$ based on Remark 4, which provides intuitive insights into the structure of $\mathbb{Q}^\star$.

We consider a two-dimensional binary classification problem. We generate 100 samples for $Y = 0$ from distribution $\mathcal{N}([2, 2]^T, I_2)$, and 100 samples for $Y = 1$ from distribution $\mathcal{N}([-1, -1]^T, I_2)$. The model $f_\beta(\cdot)$ under evaluation is logistic regression (LR). To explore the effects of varying the adjustment parameters, we fix $1/\theta_1 + 1/\theta_2 = 5$. We use the cross-entropy loss function, set the risk threshold to be 0.5 (the original loss was 0.1), and solve the problem (2). In Figure 2c-2d, we visualize the most sensitive distribution

$\mathbb{Q}^\star$ in each setting, where the decision boundary of $f_\beta(\cdot)$ is indicated by the boundary line, colored points represent the perturbed samples, shallow points represent the original samples, and the size of each point is proportional to its sample weight in $\mathbb{Q}^\star$. Corresponding with the analysis in Section 2.1, we have the following observations:

- When $\theta_1 = +\infty$, our stability criterion only considers sub-population shifts. From Figure 2b, we observe an increased concentration on several samples near the decision boundary. This corresponds with Namkoong et al. (2022); Gupta & Rothenhaeusler (2023) that focus on the tail performances.
- When $\theta_2 = +\infty$, the stability criterion only considers data corruptions. From Figure 2c, a significant number of samples are severely perturbed to adhere to the predefined risk threshold.
- When $\theta_1 = \theta_2 = 0.4$, in Figure 2d, a more balanced $\mathbb{Q}^\star$ is observed, reflecting the incorporation of both data corruptions and sub-population shifts. This showcases a scenario where samples undergo moderate and reasonable perturbations, and the sensitive distribution is not disproportionately concentrated on a limited number of samples. Such a distribution is a more holistic and reasonable approach to evaluating stability in practice, taking into account a broader range of potential shifts.

Besides, the visualizations of $\mathbb{Q}^\star$ with 0/1 loss function are provided in Figure 6 in Appendix C.1. Additionally, Figure 8 further shows that the risk $\mathbb{E}_{\mathbb{Q}^{(t)}}[W \cdot \ell(\beta, Z)]$ always converges to the pre-defined threshold r when the number of epoch number t goes to infinity.

4. Experiments

In this section, we explore real-world applications to show

the practical effectiveness of our stability evaluation criterion, including how this criterion can be utilized to compare the stability of both models and features, and to inform strategies for further enhancements.

Datasets. We use three real-world datasets. ACS Income and ACS Public Coverage are based on the American Community Survey (ACS) Public Use Microdata Sample (PUMS) (Ding et al., 2021). COVID-19 dataset (Baqui et al., 2020) is based on SIVEP-Gripe data.

- **ACS Income**. The task is to predict whether an individual's income is above $50,000 based on his/her demographic features and occupational information.
- **ACS Public Coverage** (PubCov). The task is to predict whether an individual has public health insurance.
- **COVID-19**. The dataset consists of 6,882 patients from Brazil recorded between Feb 27-May 4, 2020, which captures risk factors including comorbidities, symptoms, and demographic characteristics. The task is to predict the mortality of each patient.

Throughout the experiments, we set $1/\theta_1 + 1/\theta_2 = 5$ for adjustment parameters θ_1 and θ_2. Details of experiments can be found in Appendix D.

4.1. Model stability analysis

In this section, we first provide more in-depth empirical analyses of our proposed criterion, and demonstrate how it can reflect a model's stability with respect to data corruptions and sub-population shifts. We focus on the income prediction task for individuals from California, using the ACS Income dataset, where we sample 2,000 data points for training and another 2,000 points for evaluation.

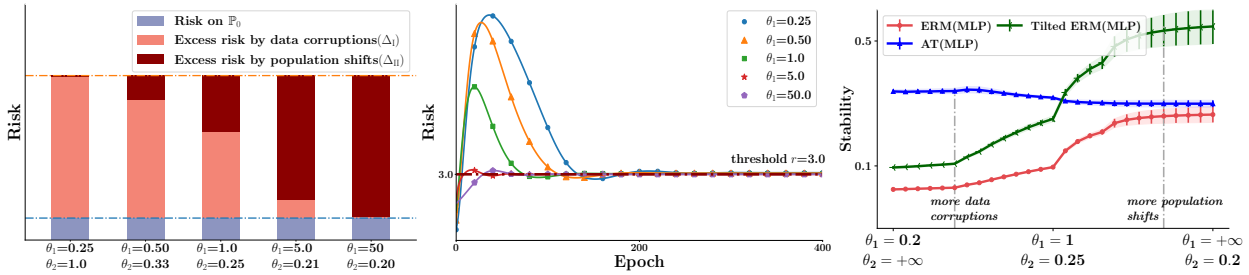
Excess risk decomposition. Recall that our stability evaluation misleads the model to a pre-defined risk threshold by perturbing the original distribution $\mathbb{P}_0$ in two ways, i.e. data corruptions and sub-population shifts. Based on the optimal solutions $\mathbb{Q}^\star \in \mathcal{P}(\mathcal{Z} \times \mathcal{W})$ of problem (P), we can compute the excess risk $\Delta = \mathbb{E}_{\mathbb{Q}^\star}[W\ell(\beta, Z)] - \mathbb{E}_{\mathbb{P}_0}[\ell(\beta, Z)]$ into two parts satisfying $\Delta = \Delta_{\mathrm{I}} + \Delta_{\mathrm{II}}$:

$$\begin{aligned} \Delta_{\mathrm{I}} &:= \mathbb{E}_{\mathbb{Q}^\star_Z}[\ell(\beta, Z)] - \mathbb{E}_{\mathbb{P}_0}[\ell(\beta, Z)], \\ \Delta_{\mathrm{II}} &:= \mathbb{E}_{\mathbb{Q}^\star}[W \cdot \ell(\beta, Z)] - \mathbb{E}_{\mathbb{Q}^\star_Z}[\ell(\beta, Z)], \end{aligned} \tag{4}$$

where Δ_{I} denotes the excess risk induced by data corruptions (data samples $\hat{z} \to z$), and Δ_{II} denotes that induced by sub-population shifts (probability density $1 \to w$). In this experiment, for a MLP model trained with empirical risk minimization (ERM), we use the cross-entropy loss and set the risk threshold to be 3.0. In Figure 3a, we vary the θ_1 and θ_2 and plot the $\Delta_{\mathrm{I}}, \Delta_{\mathrm{II}}$ in each setting. The results align with our theoretical understanding that a decrease in θ_1 leads our evaluation method to place greater emphasis on data corruptions. Conversely, a reduction in θ_2 shifts the focus of our evaluation towards sub-population shifts. This observation confirms the adaptability of our approach in weighing different types of distribution shifts based on the values of θ_1 and θ_2.

Convergence of our optimization algorithm. In Figure 3b, we plot the curve of the risk on $\mathbb{Q}^{(t)}$ w.r.t. the epoch number t throughout the optimization process. For different values of θ_1 and θ_2, we observe that the risk consistently converges to the pre-defined risk threshold of $r = 3.0$. This empirical observation is in agreement with our theoretical investigation, demonstrating the reliability and effectiveness of our optimization approach.

Reflection of stability. We then proceed to compare the



(a) Excess risk decomposition. (b) Convergence of $\mathbb{E}_{\mathbb{Q}(t)}[W \cdot \ell(\beta, Z)]$ w.r.t. epoch t. (c) Stability measure.

Figure 3: Results of the income prediction task. (a): The excess risk decomposition under different values of θ_1 and θ_2 according to (4). (b): The curve of the risk on the most sensitive distribution $\mathbb{Q}^\star$ during optimization for different choices of θ_1 and θ_2, which converge to the pre-defined risk threshold. The models under evaluation in (a) and (b) are both ERM (MLP). (c): The stability measure for MLP models trained with ERM, AT, and Tilted ERM, under varying θ_1 and θ_2. Error bars denote the standard deviations over multiple runs.

stability of MLP models trained with three well-established methods, including ERM, adversarial training (AT (Sinha et al., 2018)), and Tilted ERM (Li et al., 2023). Adversarial training is specifically designed to enhance the model's resilience to data corruptions, whereas Tilted ERM, through its use of the log-sum-exp loss function, aims to prioritize samples with elevated risks, potentially enhancing stability in the presence of sub-population shifts. For our analysis, we set the risk threshold r to 3.0, vary θ_1 and θ_2, and plot the resulting stability measure $\mathfrak{R}(\beta, 3.0)$ for each method.

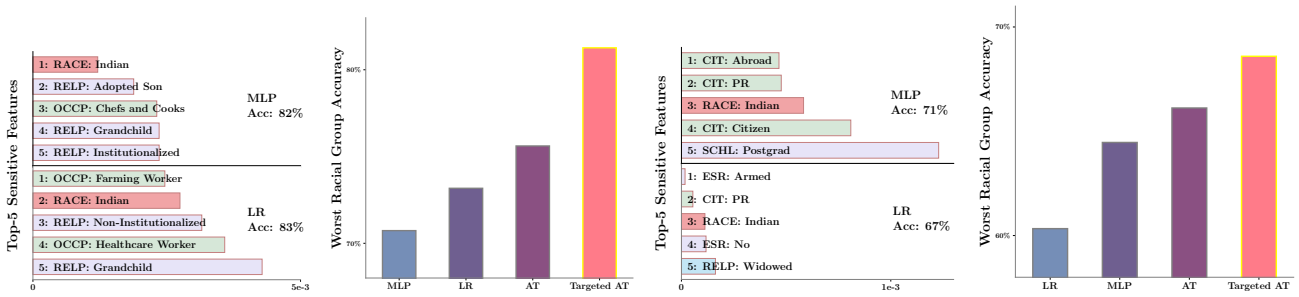
From Figure 3c, we have the following observations: (i) both robust learning methods exhibit markedly higher stability compared to ERM; (ii) AT exhibits greater stability in the context of data corruptions, while Tilted ERM shows superior performance in scenarios involving sub-population shifts. These findings align with our initial hypotheses regarding the strengths of these methods; (iii) Furthermore, the results suggest that robust learning methods tailored to specific types of distribution shifts *may face challenges in generalizing to other contexts*. Therefore, accurately identifying the types of shifts to which a model is most sensitive is crucial in practice, as it can inform machine learning engineers on strategies to further refine and improve the model's robustness and efficacy. This insight underscores the significance of our proposed stability evaluation framework. It offers a comprehensive and unified perspective on a model's stability across various types of distribution shifts, enabling a more holistic understanding and strategic approach to enhancing model robustness and reliability. Additional results on ACS PubCov and COVID-19 datasets are shown in Figure 9 in Appendix D.4.

4.2. Feature stability analysis

Building upon our previous findings, we further investigate the applicability of feature stability analysis across multiple prediction tasks, including income, insurance, and COVID-19 mortality prediction. By examining feature stability, we gain valuable insights into the specific attributes that significantly influence model performance. It provides a principle approach to enhance our understanding of the risky factors contributing to overall model instability, and thereby helps to *discover potential discriminations and improve model robustness and fairness*. Throughout all the experiments, we use 0/1 loss function and set the error rate threshold r to be 40%. The adjustment parameter θ_1 is set to 1.0, and θ_2 is 0.25.

Income prediction

We sample 2,000 data points from ACS Income dataset for training, an additional 2,000 points for the evaluation set, and a further 5,000 points to test the effectiveness of algorithmic interventions. For both the LR model and the MLP model, trained using ERM, we use the evaluation set to compute the feature sensitivity measure $\mathfrak{R}_i(\beta, r)$ for each feature as outlined in Section 2.4. The top-5 most sensitive features for each model – MLP and LR – are displayed in Figure 4a. In these visualizations, distinct colors are assigned to different types of features for clarity; for example, red is used to denote racial features, while green indicates occupation features. From the results, we observe that: (i) When the performances are similar (82% v.s. 83%), the LR model is less sensitive to input features, compared with the MLP model, which corresponds with the well-known Occam's Razor. (ii) Interestingly, our stability criterion reveals that both the MLP and LR models exhibit a notable sensitivity to the racial feature "American Indian". This raises concerns regarding potential *racial discrimination* and *unfairness* towards this specific demographic group. It is important to highlight that an individual's race should not be a determinant factor in predicting their income, and the heightened sensitivity to this feature suggests a need for careful examination and potential mitigation of biases in the models before deployment.



(a) *Income* Prediction: Feature Stability (b) *Income* Prediction: Worst Group Acc (c) *PubCov* Prediction: Feature Stability (d) *PubCov* Prediction: Worst Group Acc

Figure 4: Feature sensitivity analysis for income prediction and public coverage prediction. Figure (a) and (c): the top-5 sensitive feature scores for MLP and LR in the income prediction and the public coverage (PubCov) prediction tasks, where a smaller score means the corresponding feature is more sensitive. Figure (b) and (d): the worst racial group accuracy for MLP, LR, AT, and targeted AT in the income prediction and the public prediction tasks.

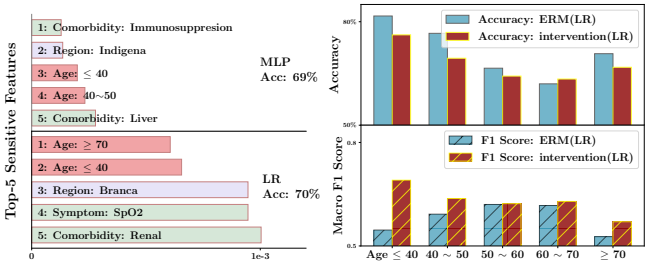
Building on our initial observations, we conduct an in-depth analysis of the accuracy across different racial groups for both the LR and MLP models. The findings, as shown in Figure 4b, align with our earlier feature stability results. Notably, the accuracy for the worst-performing racial group is significantly lower compared to other groups (for instance, a decrease from 82% to 72% in the case of the MLP model). Such findings indicate that both the LR and MLP models, when trained using ERM, exhibit unfairness towards minority racial groups. In light of these insights, our feature stability analysis serves as a valuable tool to identify and prevent the deployment of models that may perpetuate such disparities in practice.

Subsequently, we use adversarial training as an algorithmic intervention to enhance model performance. Figure 4b illustrates the results of this intervention: AT denotes adversarial training that perturbs *all* racial features, whereas targeted AT specifically perturbs the *identified* sensitive racial feature "American Indian". The results indicate that targeted AT markedly outperforms all baseline models, achieving a significant improvement in accuracy for the worst-performing racial group. This outcome effectively demonstrates the utility of our feature stability analysis in guiding targeted improvements to model performance and fairness.

Public coverage prediction

We replicated the aforementioned experiment on the ACS PubCov dataset, which involves predicting an individual's public health insurance status. Following the previous setup, we identify and display the top-5 most sensitive features for both LR and MLP models in Figure 4c. Additionally, Figure 4d presents the accuracy for the worst-performing racial group for each method.

The findings reveal several key insights: (i) The MLP model outperforms the LR model in this context (71% vs. 67%), and it exhibits less sensitivity to input features. This observation suggests that feature sensitivity is influenced by both the nature of the task and the characteristics of the model. (ii) Consistent with previous results, the "American Indian" racial feature is identified as sensitive in both models. The accuracy of the worst-performing racial group further underscores the presence of discrimination against minority groups. (iii) Leveraging our feature stability analysis, targeted AT achieves the most notable improvement. This again underscores the effectiveness of our evaluation method in enhancing model performance and fairness.



(a) Feature Stability (b) Accuracy & Macro F1 Score

Figure 5: Results of the COVID-19 mortality prediction task. (a): The top-5 most sensitive features for MLP and LR, respectively. (b): The prediction accuracy (upper sub-figure) and macro F1 score (lower sub-figure) before and after algorithmic intervention on the LR model.

COVID-19 mortality prediction

We use the COVID-19 dataset, and the task is to predict the mortality of a patient based on features including comorbidities, symptoms, and demographic characteristics. For the LR and MLP models trained with ERM, we follow the outlines in Section 2.4 and identify the top-5 most sensitive features, as shown in Figure 5a. From the results, we observe that: (i) Consistent with the trends observed in the income prediction task, the LR model demonstrates lower

sensitivity to input features compared to the MLP model when their performance levels are comparable; (ii) Notably, both LR and MLP models are quite sensitive to the "Age" feature. Given the variety of risk factors for COVID-19, such as comorbidities and symptoms, it is concerning that these models might overemphasize age, which is not the sole determinant of mortality. This highlights a critical need to ensure models effectively account for diverse age groups and do not rely excessively on age as a predictive factor.

Building on these insights, we further evaluate the accuracy and macro F1 score across different age groups for the LR model. As illustrated in Figure 5, the accuracy for younger individuals (age < 40) and older individuals (age ≥ 70) is notably high (the blue bars in the upper sub-figure). However, their corresponding macro F1 scores are significantly lower (as shown by the blue bars in the lower sub-figure). This suggests that the LR model may overly rely on the age feature for making predictions. For example, it tends to predict survival for younger individuals and mortality for older individuals with high probability, irrespective of other relevant clinical indicators. Such a simplistic approach raises concerns about the model's ability to provide nuanced predictions for these age groups.

Considering the possibility of varied mortality prediction mechanisms among different age groups, we propose a targeted algorithmic intervention: training *distinct* LR models for each age group. From the lower sub-figure in Figure 5, we see a substantial improvement in macro F1 scores for both younger and older populations.

From these three real-world experiments, we demonstrate how the proposed feature stability analysis can help discover potential discrimination and inform targeted algorithmic interventions to improve the model's reliability and fairness.

5. Closing Remarks

This work proposes an OT-based stability criterion that allows both data corruptions and sub-population shifts within a single framework. Applied to three real-world datasets, our method yields insightful observations into the robustness and reliability of machine learning models, and suggests potential algorithmic interventions for further enhancing model performance. The utility of our stability evaluation criterion to modern model architectures (e.g., Transformer, tree-based ensembles) and popular real-world applications (e.g., LLMs) is natural to further explore.

Acknowledgements

Jose Blanchet and Jiajin Li are supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0397 and NSF 1915967, 2118199, 2229012, 2312204. Peng Cui is supported in part by National Natural Science Foundation of China (No. 62141607).

Impact Statement

In this paper, we propose an OT-based stability criterion that addresses the challenges posed by both data corruptions and sub-population shifts, offering a comprehensive approach to evaluating the robustness of machine learning models. The potential broader impact of this work is significant, particularly in providing a principle approach to evaluate fairness and reliability of models deployed in real-world scenarios (based on specified criteria which we take as given). By enabling more nuanced assessments of model stability, our criterion can help prevent the deployment of biased or unreliable models, thereby contributing to more equitable outcomes, especially in high-stakes applications like healthcare, finance, and social welfare. Furthermore, our work underscores the necessity of considering and mitigating potential biases and unfairness in automated decision-making systems. As machine learning continues to play an increasingly integral role in societal functions, the tools and methodologies developed in this study provide crucial steps towards ensuring that these technologies are used responsibly and ethically.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B. E., Lee, B., Paeng, K., Zhong, A., et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2018. 1

Baqui, P., Bica, I., Marra, V., Ercole, A., and Van Der Schaar, M. Ethnic and regional variation in hospital mortality from covid-19 in brazil. *The Lancet Global Health*, 8(8): e1018–e1026, 2020. 6, 19

Beery, S., Agarwal, A., Cole, E., and Birodkar, V. The iWildCam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021. 2

Blanchet, J. and Murthy, K. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019. 13

Blanchet, J., Kang, Y., and Murthy, K. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019. 1, 2

Blanchet, J., Kuhn, D., Li, J., and Taskesen, B. Unifying distributionally robust optimization via optimal transport theory. *arXiv preprint arXiv:2308.05414*, 2023. 2, 3

Blanchet, J., Li, J., Lin, S., and Zhang, X. Distributionally robust optimization and robust statistics. *arXiv preprint arXiv:2401.14655*, 2024. 1

Blodgett, S. L., Green, L., and O'Connor, B. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pp. 1119–1130, 2016. 2

Camacho, A. and Conover, E. Manipulation of social program eligibility. *American Economic Journal: Economic Policy*, 3(2):41–65, 2011. 1

Conger, L. E., Hoffman, F., Mazumdar, E., and Ratliff, L. J. Strategic distribution shift of interacting agents via coupled gradient flows. In *Advances in Neural Information Processing Systems 36*, 2023. 1

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., and Goel, S. The measure and mismeasure of fairness. *The Journal of Machine Learning Research*, 24(1):14730–14846, 2023. 1

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. In *Advances in neural information processing systems 34*, pp. 6478–6490, 2021. 1, 6, 18, 19

Duchi, J. C. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021. 1

Elmes, A., Alemohammad, H., Avery, R., Caylor, K., Eastman, J. R., Fishgold, L., Friedl, M. A., Jain, M., Kohli, D., Laso Bayas, J. C., et al. Accounting for training data error in machine learning applied to earth observations. *Remote Sensing*, 12(6):1034, 2020. 1

Frogner, C., Claici, S., Chien, E., and Solomon, J. Incorporating unlabeled data into distributionally robust learning. *Journal of Machine Learning Research*, 22(56): 1–46, 2021. 1

Gokhale, T., Mishra, S., Luo, M., Sachdeva, B. S., and Baral, C. Generalized but not robust? comparing the effects of data modification methods on out-of-domain generalization and adversarial robustness. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2705–2718, 2022. 2

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, 2020. 1

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*, 2020. 1

Gupta, S. and Rothenhaeusler, D. The s-value: evaluating stability with respect to distributional shifts. In *Advances in Neural Information Processing Systems 37*, 2023. 2, 3, 6

Hand, D. J. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–15, 2006. 1

Hell, F., Hinz, G., Liu, F., Goyal, S., Pei, K., Lytvynenko, T., Knoll, A., and Yiqiang, C. Monitoring perception reliability in autonomous driving: Distributional shift detection for estimating the impact of input data on prediction accuracy. In *Proceedings of the 5th ACM Computer Science in Cars Symposium*, pp. 1–9, 2021. 1

Jacobucci, R. and Grimm, K. J. Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, 15(3): 809–816, 2020. 1

Kinoshita, K., Ochiai, T., Delcroix, M., and Nakatani, T. Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2020)*, pp. 7009–7013. IEEE, 2020. 1

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In *Proceedings of the 38th of International Conference on Machine Learning (ICML 2021)*, pp. 5815–5826. PMLR, 2021. 1

Kuhn, D., Esfahani, P. M., Nguyen, V. A., and Shafieezadeh-Abadeh, S. 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. 1

Li, M., Namkoong, H., and Xia, S. Evaluating model performance under worst-case subpopulations. In *Advances in Neural Information Processing Systems 34*, pp. 17325–17334, 2021. 2

Li, T., Beirami, A., Sanjabi, M., and Smith, V. On tilted losses in machine learning: Theory and applications. *Journal of Machine Learning Research*, 24(142):1–79, 2023. 7, 19, 20

Li, Y. and Mykland, P. A. Rounding errors and volatility estimation. *Journal of Financial Econometrics*, 13(2): 478–504, 2015. 1

Liu, J., Shen, Z., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021. 18, 19

Liu, J., Wang, T., Cui, P., and Namkoong, H. On the need for a language describing distribution shifts: Illustrations on tabular datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 1

Malinin, A., Band, N., Gal, Y., Gales, M., Ganshin, A., Chesnokov, G., Noskov, A., Ploskonosov, A., Prokhorenkova, L., Provilkov, I., et al. Shifts: A dataset of real distributional shift across multiple large-scale tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 1

Namkoong, H., Ma, Y., and Glynn, P. W. Minimax optimal estimation of stability under distribution shift. *arXiv preprint arXiv:2212.06338*, 2022. 2, 3, 6, 20

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., and Razaviyayn, M. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems 32*, 2019. 5

Owen, A. B. *Empirical Likelihood*. CRC press, 2001. 2

Pasterkamp, G., Den Ruijter, H. M., and Libby, P. Temporal shifts in clinical presentation and underlying mechanisms of atherosclerotic disease. *Nature Reviews Cardiology*, 14(1):21–29, 2017. 2

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32*, 2019. 20

Rahimian, H. and Mehrotra, S. Frameworks and Results in Distributionally Robust Optimization. *Open Journal of Mathematical Optimization*, 2022. URL https://ojmo.centre-mersenne.org/articles/10.5802/ojmo.15/. 1

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217, 2021. 1

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*, 2019. 1

Sapiezynski, P., Kassarnig, V., Wilson, C., Lehmann, S., and Mislove, A. Academic performance prediction in a gender-imbalanced environment. In *FATREC Workshop on Responsible Recommendation Proceedings*, 2017. 2

Shapiro, A. On duality theory of conic linear problems. In *Semi-infinite programming*, pp. 135–165. Springer, 2001. 4, 13

Si, N., Murthy, K., Blanchet, J., and Nguyen, V. A. Testing group fairness via optimal transport projections. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, pp. 9649–9659, 2021. 5

Sinha, A., Namkoong, H., and Duchi, J. Certifying some distributional robustness with principled adversarial training. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, 2018. 5, 7, 19, 20

Van der Vaart, A. W. *Asymptotic Statistics*, volume 3. Cambridge university press, 2000. 12

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Albu, E., Arshi, B., Bellou, V., Bonten, M. M., et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj*, 369, 2020. 1

Yang, Y., Zhang, H., Katabi, D., and Ghassemi, M. Change is hard: A closer look at subpopulation shift. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, pp. 39584–39622, 2023. 1

Yi, M., Hou, L., Sun, J., Shang, L., Jiang, X., Liu, Q., and Ma, Z. Improved ood generalization via adversarial training and pretraing. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, pp. 11987–11997, 2021. 5

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

Zou, X. and Liu, W. On the adversarial robustness of out-of-distribution generalization models. In *Advances in Neural Information Processing Systems 37*, 2023. 2

A. Proofs

A.1. Proof of Theorem 1

Proof. To start with, we first reformulation the primal problem (P) into an infinite-dimensional linear program:

$$\begin{array}{ll} \inf\limits_{\pi} & \mathbb{E}_{\pi}[c((Z,W),(\hat{Z},\hat{W}))] \\ \text{s. t.} & \pi \in \mathcal{P}((\mathcal{Z}\times\mathcal{W})^2) \\ & r - \mathbb{E}_{\pi}[W\cdot\ell(\beta,Z)] \le 0 \\ & \mathbb{E}_{\pi}[W] = 1 \\ & \pi_{(\hat{Z},\hat{W})} = \hat{\mathbb{P}}. \end{array} \tag{Primal}$$

We aim to apply Sion's minimax theorem to the Lagrangian function

$$L(\pi;h,\alpha) = hr + \alpha + \mathbb{E}_{\pi}[c((Z,W),(\hat{Z},\hat{W})) - h\cdot W\cdot\ell(\beta,Z) - \alpha\cdot W],$$

where $h\in\mathbb{R}_+$, $\alpha\in\mathbb{R}$, and π belongs to the primal feasible set

$$\Pi_{\hat{\mathbb{P}}} = \left\{\pi\in\mathcal{P}((\mathcal{Z}\times\mathcal{W})^2)\ :\ \pi_{(\hat{Z},\hat{W})} = \hat{\mathbb{P}}\right\}.$$

Since $\mathcal{Z}\times\mathcal{W}$ is compact, it follows that $\mathcal{P}(\mathcal{Z}\times\mathcal{W})$ is tight. Furthermore, as a subset of a tight set is also tight, we conclude that $\Pi_{\hat{\mathbb{P}}}$ is tight as well. Consequently, according to Prokhorov's theorem (Van der Vaart, 2000, Theorem 2.4), $\Pi_{\hat{\mathbb{P}}}$ has a compact closure. By taking the limit in the marginal equation, we observe that $\Pi_{\hat{\mathbb{P}}}$ is weakly closed, establishing that $\Pi_{\hat{\mathbb{P}}}$ is indeed compact. Moreover, it can be readily demonstrated that $\Pi_{\hat{\mathbb{P}}}$ is convex.

The Lagrangian function $L(\pi;h,\alpha)$ is linear in both π and (h,α). To employ Sion's minimax theorem, we will now prove that (i) $L(\pi;h,\alpha)$ is lower semicontinuous in π under the weak topology and (ii) continuous in (h,α) under the uniform topology in $\mathbb{R}_+\times\mathbb{R}$.

(i) Suppose that π_n converges weakly to π. Then, Portmanteau theorem states that for any lower semicontinuous function g that is bounded below, we have

$$\liminf_{n\to+\infty}\int g\,\mathrm{d}\pi_n \ge \int g\,\mathrm{d}\pi.$$

Since $\ell(\beta,\cdot)$ is upper semicontinuous for all β and $w,h\ge 0$, we can conclude that $h\cdot w\cdot\ell(\beta,z)$ is upper semicontinuous w.r.t (z,w). Moreover, armed with the lower semicontinuity of the function $c((z,w),(\hat{z},\hat{w}))$, we know the following candidate function

$$c((z,w),(\hat{z},\hat{w})) - h\cdot w\cdot\ell(\beta,z) - \alpha\cdot w$$

is lower semicontinuous with respect to (z,w) for any $(\hat{z},\hat{w})\in\mathcal{Z}\times\mathcal{W}$. As $\mathcal{Z}\times\mathcal{W}$ is compact, the above candidate function is also bounded below. Thus, we have

$$\liminf_{n\to+\infty} L(\pi_n;h,\alpha) \ge L(\pi;h,\alpha).$$

It follows that $L(\pi;h,\alpha)$ is lower semicontinuous in π under the weak topology.

(ii) Suppose now that $\lim_{n\to+\infty} h_n = h$ in the Euclidean topology and $\lim_{n\to\infty}\alpha_n = \alpha$ in the Euclidean topology. There exists $\bar{h}\in\mathbb{R}_+$ and $\bar{\alpha}\in\mathbb{R}$ with $\sup_{n\to\infty}|h_n|\le\bar{h}$ and $\sup_{n\to\infty}|\alpha_n|<\bar{\alpha}$ for all $n\ge 1$. Thus, by the dominated convergence theorem, we have

$$\lim_{n\to+\infty} L(\pi;h_n,\alpha_n) = L(\pi;h,\alpha).$$

We then conclude that $L(\pi;h,\alpha)$ is continuous in (h,α) under the Ecludiean topology in $\mathbb{R}_+\times\mathbb{R}$.

We are now prepared to utilize Sion's minimax theorem, and thus, we have:

$$\inf_{\pi\in\Pi_{\hat{\mathbb{P}}}}\ \sup_{h\in\mathbb{R}_+,\alpha\in\mathbb{R}} L(\pi;h,\alpha) = \sup_{h\in\mathbb{R}_+,\alpha\in\mathbb{R}}\ \inf_{\pi\in\Pi_{\hat{\mathbb{P}}}} L(\pi;h,\alpha). \tag{5}$$

Our subsequent task involves demonstrating the equivalence between the left-hand side of (5) and the primal problem (Primal). To achieve this, we will re-express the function L as follows:

$$L(\pi; h, \alpha) = \mathbb{E}_\pi[c((Z, W), (\hat{Z}, \hat{W})] + h\left(r - \mathbb{E}_\pi[W \cdot \ell(\beta, Z)]\right) + \alpha(1 - \mathbb{E}_\pi[W]).$$

Then, we can see $\inf_{\pi \in \Pi_{\hat{\mathbb{P}}}} \sup_{h \in \mathbb{R}_+, \alpha \in \mathbb{R}} L(\pi; h, \alpha)$ is bounded above. To start with, we construct a single support distribution as follows: $\mathbb{Q}_0 = \delta_{(z^\star, 1)}$ where $z^\star = \arg\max_{z \in \mathcal{Z}} \ell(\beta, z)$. Then, we have

$$\begin{aligned}\inf_{\pi \in \Pi_{\hat{\mathbb{P}}}} \sup_{h \in \mathbb{R}_+, \alpha \in \mathbb{R}} L(\pi; h, \alpha) &\leq \sup_{h \in \mathbb{R}_+, \alpha \in \mathbb{R}} L(\mathbb{Q}_0 \otimes \hat{\mathbb{P}}; h, \alpha), \\ &= \mathbb{E}_{\mathbb{Q}_0 \otimes \hat{\mathbb{P}}}[c((Z, W), (\hat{Z}, \hat{W}))] + \sup_{h \in \mathbb{R}_+} h(r - \bar{r}) < +\infty,\end{aligned}$$

where the second inequality follows from $\mathbb{E}_{\mathbb{Q}_0}[W] = 1$ and the last equality holds as we know $r \leq \bar{r} = \mathbb{E}_{\mathbb{Q}_0}[\ell(\beta, Z)] = \max_{z \in Z} \ell(\beta, Z)$ and c is continuous and hence bounded on a compact domain $\mathcal{Z} \times W$. For any feasible point $\pi \in \Pi_{\hat{\mathbb{P}}}$, let us consider the inner supremum of the left-hand-side of (5), ensuring it doesn't go to infinity. In this case, we find that

$$\begin{aligned}&r - \mathbb{E}_\pi[W \cdot \ell(\beta, Z)] \leq 0 \\ &\mathbb{E}_\pi[W] = 1.\end{aligned}$$

It remains to be shown that the sup-inf part is equivalent to the dual problem (D). To do this, we rewrite the dual problem as

$$\begin{aligned}&\sup_{h \in \mathbb{R}_+, \alpha \in \mathbb{R}} \inf_{\pi \in \Pi_{\hat{\mathbb{P}}}} L(\pi; h, \alpha). \\ =& \sup_{h \in \mathbb{R}_+, \alpha \in \mathbb{R}} hr + \alpha + \inf_{\pi \in \Pi_{\hat{\mathbb{P}}}} \mathbb{E}_\pi[c((Z, W), (\hat{Z}, \hat{W})) - h \cdot W \cdot \ell(\beta, Z) - \alpha \cdot W].\end{aligned}$$

The last step is to take the supremum of L over $\pi \in \Pi_{\hat{\mathbb{P}}}$. That is,

$$\begin{aligned}&\inf_{\pi \in \Pi_{\hat{\mathbb{P}}}} \mathbb{E}_\pi[c((Z, W), (\hat{Z}, \hat{W})) - h \cdot W \cdot \ell(\beta, Z) - \alpha \cdot W] \\ =& \mathbb{E}_{\hat{\mathbb{P}}}\left[\min_{(z,w) \in \mathcal{Z} \times \mathcal{W}} c((z, w), (\hat{Z}, \hat{W})) - h \cdot w \cdot \ell(\beta, z) - \alpha \cdot w\right],\end{aligned}$$

due to the measurability of functions of the form $\min_{(z,w) \in \mathcal{Z} \times \mathcal{W}} c((z, w), (\hat{Z}, \hat{W})) - h \cdot w \cdot \ell(\beta, z) - \alpha \cdot w$, following the similar argument in (Blanchet & Murthy, 2019). □

When the reference measure is discrete, i.e., $\mathbb{P}_0 = \frac{1}{n}\sum_{i=1}^n \delta_{\hat{z}_i}$, we can get the strong duality result under some mild conditions.

Theorem 4 (Strong duality for problem (P))**.** *Suppose that $r < \mathbb{E}_{\mathbb{P}_0}[\ell(\beta, Z)]$ holds. Then we have,*

$$\mathfrak{R}(\beta, r) = \sup_{h \in \mathbb{R}_+, \alpha \in \mathbb{R}} hr + \alpha + \mathbb{E}_{\hat{\mathbb{P}}}\left[\tilde{\ell}_c^{\alpha,h}(\beta, (\hat{Z}, \hat{W}))\right]. \tag{D}$$

Proof. To start, we have the primal problem (P) admits

$$\begin{aligned}\mathfrak{R}(\beta, r) = \inf_{\substack{\pi : \pi_{(\hat{Z}, \hat{W})} = \hat{\mathbb{P}}, \pi \in \mathcal{P}((\mathcal{Z} \times \mathcal{W})^2) \\ \text{s. t.}}} \quad & \mathbb{E}_\pi[c((Z, W), (\hat{Z}, \hat{W}))] \\ & r - \mathbb{E}_\pi[W \cdot \ell(\beta, Z)] \leq 0 \\ & \mathbb{E}_\pi[W] = 1.\end{aligned} \tag{6}$$

Due to the condition $r < \mathbb{E}_{\mathbb{P}_0}[\ell(\beta, Z)]$, we know the Slater condition, i.e., $r - \mathbb{E}_{\mathbb{P}_0 \otimes \delta_1 \times \mathbb{P}_0 \otimes \delta_1}[W \cdot \ell(\beta, Z)] < 0$, holds. Thus, we can apply Shapiro (2001, Proposition 3.4) to get the strong duality result directly. That is,

$$\mathfrak{R}(\beta, r) = \sup_{h \in \mathbb{R}_+, \alpha \in \mathbb{R}} hr + \alpha + \mathbb{E}_{\hat{\mathbb{P}}}\left[\inf_{(z,w) \in \mathcal{Z} \times \mathcal{W}} c((z, w), (\hat{Z}, \hat{W})) + \alpha w - h \cdot w \cdot \ell(\beta, z)\right].$$

□

A.2. Proof of Proposition 1

Proof. Now, we are trying to calculate the surrogate function with our proposed cost function c in (1) . Then, we have

$$\begin{aligned}\tilde{\ell}_c^{\alpha,h}(\beta,(\hat{z},\hat{w})) &= \min_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \theta_1\cdot w\cdot d(z,\hat{z})+\theta_2(\phi(w)-\phi(\hat{w}))_+ - \alpha w - h\cdot w\cdot \ell(\beta,z)\\ &= \min_{z\in\mathcal{Z}} \theta_2\cdot \min_{w\in\mathbb{R}} -w\frac{h\cdot\ell(\beta,z)-\theta_1\cdot d(z,\hat{z})+\alpha}{\theta_2}+\phi(w)+\mathbb{I}_{\mathcal{W}}(w)\\ &= \min_{z\in\mathcal{Z}} -\theta_2\cdot(\phi+\mathbb{I}_{\mathcal{W}})^*\left(\frac{h\cdot\ell(\beta,z)-\theta_1\cdot d(z,\hat{z})+\alpha}{\theta_2}\right).\end{aligned}$$

where the first equality follows as $\hat{W}=1$ almost surely and $\phi(1)=0$, and the second equality holds due to the definition of conjugate functions.

(i) When $\mathcal{W}=\mathbb{R}_+$ and $\phi(t)=t\log t-t+1$, we know its conjugate function $(\phi+\mathbb{I}_{\mathbb{R}_+})^*=\exp(t)-1$. Consequently, we obtain the following:

$$\begin{aligned}\tilde{\ell}_c^{\alpha,h}(\beta,(\hat{z},\hat{w})) &= \min_{z\in\mathcal{Z}} -\theta_2\cdot\exp\left(\frac{h\cdot\ell(\beta,z)-\theta_1\cdot d(z,\hat{z})+\alpha}{\theta_2}\right)+\theta_2\\ &= -\theta_2\cdot\exp\left(\frac{\max_{z\in\mathcal{Z}} h\cdot\ell(\beta,z)-\theta_1\cdot d(z,\hat{z})+\alpha}{\theta_2}\right)+\theta_2\\ &= -\theta_2\cdot\exp\left(\frac{\ell_{h,\theta_1}(\hat{z})+\alpha}{\theta_2}\right)+\theta_2.\end{aligned}$$

where the second equality follows from the fact the function $\exp(\cdot)$ is monotonically increasing. Hence, we can reformulate the dual problem (D) as

$$\mathfrak{R}(\beta,r)=\sup_{h\in\mathbb{R}_+,\alpha\in\mathbb{R}} hr+\alpha-\theta_2\mathbb{E}_{\mathbb{P}_0}\left[\exp\left(\frac{\ell_{h,\theta_1}(\hat{Z})+\alpha}{\theta_2}\right)\right]+\theta_2.$$

Next, we will solve the supremum problem via α and the first-order condition reads

$$1-\exp\left(\frac{\alpha}{\theta_2}\right)\mathbb{E}_{\mathbb{P}_0}\left[\exp\left(\frac{\ell_{h,\theta_1}(\hat{Z})}{\theta_2}\right)\right]=0$$

and $\alpha^\star=-\theta_2\log\left(\mathbb{E}_{\mathbb{P}_0}\left[\frac{\ell_{h,\theta_1}(\hat{Z})}{\theta_2}\right]\right)$. Put all of them together, we get

$$\mathfrak{R}(\beta,r)=\sup_{h\in\mathbb{R}_+} hr-\theta_2\log\left(\mathbb{E}_{\mathbb{P}_0}\left[\exp\left(\frac{\ell_{h,\theta_1}(\hat{Z})}{\theta_2}\right)\right]\right).$$

(ii) When $\mathcal{W}=\mathbb{R}+$ and $\phi(t)=(t-1)^2$, the conjugate function can be computed as $(\phi+\mathbb{I}_{\mathbb{R}_+})^*(t)=(\frac{t}{2}+1)_+^2-1$. Additionally, it is straightforward to demonstrate that $(\phi+\mathbb{I}_{\mathbb{R}_+})^*(t)$ is a monotonically increasing function. Hence, we have:

$$\begin{aligned}\tilde{\ell}_c^{\alpha,h}(\beta,(\hat{z},\hat{w})) &= \min_{z\in\mathcal{Z}} -\theta_2\cdot(\phi+\mathbb{I}_{\mathcal{W}})^*\left(\frac{h\cdot\ell(\beta,z)-\theta_1\cdot d(z,\hat{z})+\alpha}{\theta_2}\right)\\ &= \min_{z\in\mathcal{Z}} -\theta_2\cdot\left(\frac{h\cdot\ell(\beta,z)-\theta_1\cdot d(z,\hat{z})+\alpha}{2\theta_2}+1\right)_+^2+\theta_2\\ &= -\theta_2\cdot\left(\frac{\ell_{h,\theta_1}(\hat{z})+\alpha}{2\theta_2}+1\right)_+^2+\theta_2\end{aligned}$$

where the third equality holds as the monotonicity of $(\phi+\mathbb{I}_{\mathbb{R}_+})^*$. Then, we can reduce the dual problme (D) as

$$\sup_{h\geq 0,\alpha\in\mathbb{R}} hr+\alpha+\theta_2-\theta_2\mathbb{E}_{\mathbb{P}_0}\left[\left(\frac{\ell_{h,\theta_1}(\hat{Z})+\alpha}{2\theta_2}+1\right)_+^2\right].$$

□

Remark 5. We want to highlight the distinction between the KL and χ^2-divergence cases. In the latter case, we are unable to derive a closed-form expression for the optimal $\alpha^\star$. Instead, we must reduce it to a solution of a piecewise linear equation as follows:

$$\mathbb{E}_{\mathbb{P}_0}\left[\left(\frac{\ell_{h,\theta_1}(\hat{Z})+\alpha}{2\theta_2}+1\right)_+\right]=1. \tag{7}$$

A.3. Proof of Theorem 2

Proof. By introducing epigraphical auxiliary variable $t\in\mathbb{R}$, we know problem (2) is equivalent to

$$\begin{aligned}
&\min_{h\geq 0} -hr+\theta_2\log\mathbb{E}_{\mathbb{P}_0}\left[\exp\left(\frac{\ell_{h,\theta_1}(\hat{Z})}{\theta_2}\right)\right]\\
&=\left\{\begin{array}{ll}\min & -hr+t\\ \text{s.t.} & h\in\mathbb{R}_+, t\in\mathbb{R}\\ & \theta_2\log\mathbb{E}_{\mathbb{P}_0}\left[\exp\left(\frac{\ell_{h,\theta_1}(\hat{Z})}{\theta_2}\right)\right]\leq t\end{array}\right. \qquad (8)\\
&=\left\{\begin{array}{ll}\min & -hr+t\\ \text{s.t.} & \lambda\in\mathbb{R}_+, t\in\mathbb{R}, \eta\in\mathbb{R}^n_+\\ & (\eta_i,\theta_2,\ell_{h,\theta_1}(\hat{z}_i)-t)\in\mathcal{K}_{\exp}\quad \forall i\in[n]\\ & \frac{1}{n}\sum_{i=1}^n\eta_i\leq\theta_2\end{array}\right.\\
&=\left\{\begin{array}{lll}\min & -hr+t & \\ \text{s.t.} & \lambda\in\mathbb{R}_+, t\in\mathbb{R}, \eta\in\mathbb{R}^n_+, p\in\mathbb{R}_n & \\ & (\eta_i,\theta_2,p_i-t)\in\mathcal{K}_{\exp} & \forall i\in[n]\\ & \ell_{h,\theta_1}(\hat{z}_i)\leq p_i & \forall i\in[n]\\ & \frac{1}{n}\sum_{i=1}^n\eta_i\leq\theta_2. & \end{array}\right. \qquad (9)
\end{aligned}$$

Here, the second equality can be derived from the fact that the second inequality in problem (8) can be reformulated as

$$\mathbb{E}_{\mathbb{P}_0}\left[\exp\left(\frac{\ell_{h,\theta_1}(\hat{Z})-t}{\theta_2}\right)\right]\leq 1.$$

To handle this constraint, we introduce an auxiliary variable $\eta\in\mathbb{R}^n_+$, allowing us to further decompose it into n exponential cone constraints and one additional linear constraint. Specifically, we have

$$\begin{cases}\dfrac{1}{n}\displaystyle\sum_{i=1}^n\eta_i\leq\theta_2\\ \theta_2\exp\left(\dfrac{\ell_{h,\theta_1}(\hat{z}_i)-t}{\theta_2}\right)\leq\eta_i,\quad\forall i\in[n]\end{cases}$$

The third constraint can be further reduced to (9) by considering the fact that the set $\mathcal{K}_{\exp}$ corresponds to the exponential cone, which is defined as

$$K_{\exp}=\left\{(x_1,x_2,x_3)\in\mathbb{R}^3: x_1\geq x_2\cdot\exp\left(\tfrac{x_3}{x_2}\right), x_2>0\right\}\cup\{(x_1,0,x_3)\in\mathbb{R}^3: x_1\geq 0, x_3\leq 0\}.$$

The fourth equality is due to $\ell_{h,\theta_1}(\hat{z}_i)\leq p_i$ when we introduce auxiliary variables p_i.

Next, we show that $\ell_{h,\theta_1}(\hat{z}_i)\leq p_i$ admits the following equivalent forms

$$\begin{aligned}
&\ell_{h,\theta_1}(\hat{z}_i)\leq p_i\\
\iff&\sup_{z\in\mathcal{Z}}\left\{h\cdot\max_{k\in[K]}y\cdot a_k^\top x+b_k-\theta_1 d(z,\hat{z}_i)\right\}\leq p_i\\
\iff&\sup_{z\in\mathcal{Z}}\left\{h\cdot y\cdot a_k^\top x+b_k-\theta_1 d(z,\hat{z}_i)\right\}\leq p_i\ \forall k\in[K]\\
\iff&\sup_{x\in\mathbb{R}^d}\left\{h\cdot\hat{y}_i\cdot a_k^\top x+b_k-\theta_1\|x-\hat{x}_i\|_2^2\right\}\leq p_i\ \forall k\in[K]\\
\iff&\frac{\|a_k\|_2^2}{4\theta_1}\cdot h^2+\hat{y}_i\cdot a_k^T\hat{x}_i\cdot h+b_k\leq p_i,\ \forall k\in[K]
\end{aligned} \tag{10}$$

where the second equivalence arises from the non-negativity of h, while the third one can be derived from the nature of the cost function, which is defined as $d(z, \hat{z}_i) = \|x - \hat{x}_i\|_2^2 + \infty \cdot |y - \hat{y}_i|$. The second term in the cost function prevents us from perturbing the label due to the imposed budget limit.

Put everthing together, we have

$$\begin{array}{ll} \min & -hr + t \\ \text{s.t.} & \lambda \in \mathbb{R}_+, t \in \mathbb{R}, \eta \in \mathbb{R}_+^n, p \in \mathbb{R}^n \\ & (\eta_i, \theta_2, p_i - t) \in \mathcal{K}_{\exp} \qquad \forall i \in [n] \\ & \frac{\|a_k\|_2^2}{4\theta_1} \cdot h^2 + \hat{y}_i \cdot a_k^T \hat{x}_i \cdot h + b_k \leq p_i, \qquad \forall k \in [K], \forall i \in [n] \\ & \frac{1}{n} \sum_{i=1}^n \eta_i \leq \theta_2. \end{array}$$

□

A.4. Proof of Theorem 3

Proof. By introducing epigraphical auxiliary variable $t \in \mathbb{R}$, we know problem (2) is equivalent to

$$\min_{h \geq 0, \alpha \in \mathbb{R}} -hr - \alpha + \theta_2 + \theta_2 \mathbb{E}_{\mathbb{P}_0} \left[\left(\frac{\ell_{h,\theta_1}(\hat{Z}) + \alpha}{2\theta_2} + 1 \right)_+^2 \right]$$

$$= \left\{ \begin{array}{ll} \min & -hr - \alpha + t \\ \text{s.t.} & h \in \mathbb{R}_+, \alpha \in \mathbb{R}, t \in \mathbb{R} \\ & \theta_2 \mathbb{E}_{\mathbb{P}_0} \left[\left(\frac{\ell_{h,\theta_1}(\hat{Z}) + \alpha}{2\theta_2} + 1 \right)_+^2 \right] \leq t \end{array} \right.$$

$$= \left\{ \begin{array}{ll} \min & -hr + t \\ \text{s.t.} & h \in \mathbb{R}_+, \alpha \in \mathbb{R}, t \in \mathbb{R}, \eta \in \mathbb{R}_+^n \\ & \ell_{h,\theta_1}(\hat{z}) + 2\theta_2 \alpha + 2\theta_2 \leq 2\theta_2 \eta_i \qquad \forall i \in [n] \\ & \frac{\theta_2}{n} \sum_{i=1}^n \eta_i^2 \leq t \end{array} \right.$$

$$= \left\{ \begin{array}{ll} \min & -hr + t \\ \text{s.t.} & h \in \mathbb{R}_+, \alpha \in \mathbb{R}, t \in \mathbb{R}, \eta \in \mathbb{R}_+^n \\ & \frac{\|a_k\|_2^2}{4\theta_1} \cdot h^2 + \hat{y}_i \cdot a_k^T \hat{x}_i \cdot h + b_k + 2\theta_2 \alpha + 2\theta_2 \leq 2\theta_2 \eta_i \qquad \forall k \in [K], \forall i \in [n] \\ & \frac{\theta_2}{n} \sum_{i=1}^n \eta_i^2 \leq t \end{array} \right.$$

Here, the second equality follows from the fact that the constraint can be reformulated as

$$\begin{cases} \dfrac{\theta_2}{n} \displaystyle\sum_{i=1}^n \eta_i^2 \leq t \\ \ell_{h,\theta_1}(\hat{z}) + 2\theta_2 \alpha + 2\theta_2 \leq 2\theta_2 \eta_i, \eta_i \in \mathbb{R}_+. \end{cases}$$

as the function $(\cdot)_+^2$ is monotonically increasing. The last equality holds due to (10) . □

B. Pseudo-code for Algorithms

In this section, we provide the pseudo-code of our algorithms. For $\phi(t) = t \log t - t + 1$, please refer to Algorithm 1, and for $\phi(t) = (t-1)^2$, please see Algorithm 2.

C. Stylized / Toy Example

In this section, we provide more results for our toy example in Section 3.

C.1. Visualization

For the toy example in Section 3, we showcase the most sensitive distributions with continuous loss function (cross-entropy loss). Here we choose $\phi(t) = t \log t - t + 1$, and provide the results with 0/1 loss. We set the error rate threshold r to be 30%. The results are shown in Figure 6. From the results, we have the following observations:

Algorithm 1 Stability evaluation with general nonlinear loss functions ($\phi(t) = t\log t - t + 1$).

Input: trained model $f_\beta(\cdot)$, samples $\{\hat{z}_i\}_{i=1}^n$, adjustment parameters θ_1, θ_2, pre-defined threshold r;
Hyper-parameters: outer iteration number T_{out}, inner iteration number T_{in}, learning rates η, γ;
Initialize for $i \in [n]$, set $z_i^{(0)} \leftarrow \hat{z}_i$, and $h^{(0)} = 1$;
for $t = 0$ to $T_{\text{out}} - 1$ **do**
  for $k = 0$ to $T_{\text{in}} - 1$ **do**
    For $i \in [n]$, $z_i^{(k+1)} \leftarrow z_i^{(k)} + \eta \cdot \nabla_Z \left(h^{(t)} \ell(\beta, z_i^{(k)}) - \theta_1 d(z_i^{(k)}, \hat{z}_i) \right)$ (update samples using ADAM optimizer)
  end for
  Update the dual parameter using ADAM optimizer as:

$$h^{(t+1)} \leftarrow h^{(t)} + \gamma \cdot \nabla_h \left(h^{(t)} r - \theta_2 \log \sum_{i=1}^n \left[\exp(\frac{h^{(t)} \ell(\beta, z_i^{(T_{\text{in}})}) - \theta_1 d(z_i^{(T_{\text{in}})}, \hat{z}_i)}{\theta_2}) \right] \right)$$

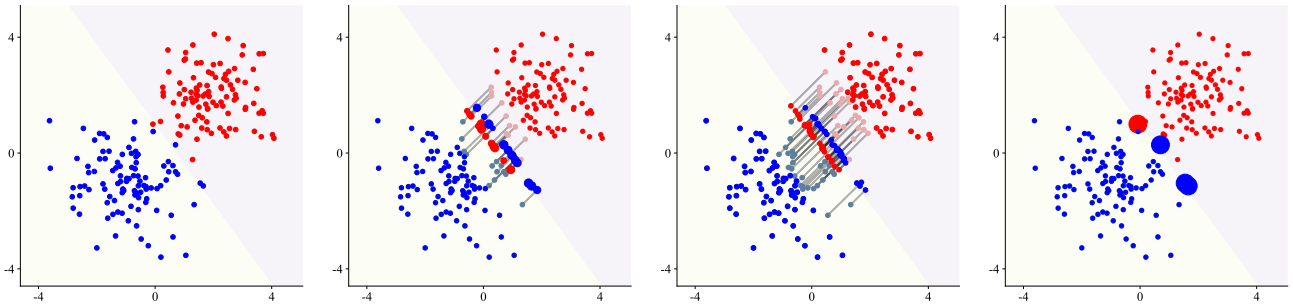
end for
Output: stability criterion $\mathfrak{R}(\beta, r)$ (Equation (2)), the most sensitive distribution $\hat{\mathbb{Q}}^*$ (according to Remark 4).

Algorithm 2 Stability evaluation with general nonlinear loss functions ($\phi(t) = (t-1)^2$)

Input: trained model $f_\theta(\cdot)$, samples $\{\hat{z}_i\}_{i=1}^n$, adjustment parameters θ_1, θ_2, mis-classification threshold r;
Hyper-parameters: outer iteration number T_{out}, inner iteration number T_{in}, learning rates $\eta, \gamma_h, \gamma_\alpha$;
Initialize for $i \in [n]$, set $z_i^{(1)} \leftarrow \hat{z}_i$, and $h^{(1)} = 1$;
for $t = 1$ to T_{out} **do**
  for $k = 1$ to T_{in} **do**
    For $i \in [n]$, $z_i^{(k+1)} \leftarrow z_i^{(k)} + \eta \cdot \nabla_Z \left(h^{(t)} \ell(\beta; z_i^{(k)}) - \theta_1 d(z_i^{(k)}, \hat{z}_i) \right)$;
    (update samples using ADAM optimizer)
  end for
  Compute α^* via Equation 7;
  Update the dual parameter using ADAM optimizer as:

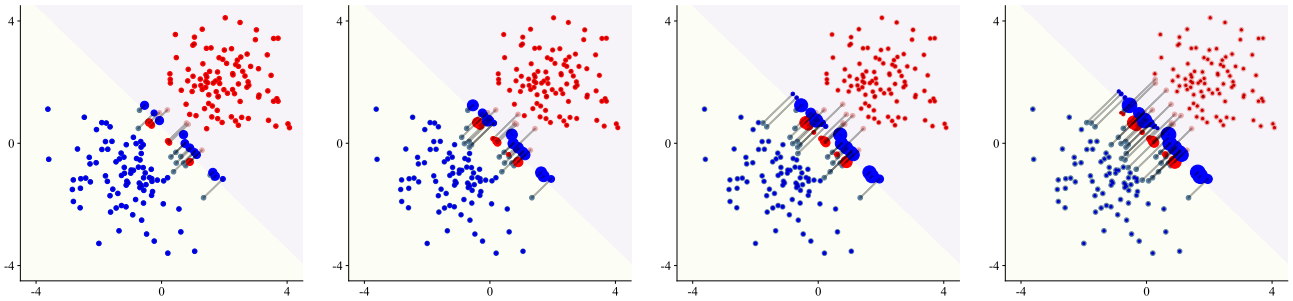
$$h^{(t+1)} \leftarrow h^{(t)} + \gamma \cdot \nabla_h \left(hr + \alpha^* + \theta_2 - \theta_2 \sum_{i=1}^n \left(\frac{\ell_{h,\theta_1}(\beta, z_i^{(T_{\text{in}})}) + \alpha^*}{2\theta_2} + 1 \right)_+^2 \right); \tag{11}$$

end for
Output: stability criterion $\mathfrak{R}(\beta, r)$ (Equation (2)), the most sensitive distribution $\hat{\mathbb{Q}}^*$ (according to Remark 4).



(a) Original Dataset (b) $\theta_1 = 1.0, \theta_2 = 0.25$ (c) $\theta_1 = 0.2, \theta_2 = +\infty$ (d) $\theta_1 = +\infty, \theta_2 = 0.2$

Figure 6: Visualizations of the original dataset and the most sensitive distribution $\mathbb{Q}^\star$ with *0/1* loss function under different θ_1, θ_2. The original prediction error rate is 1%, and the error rate threshold r is set to 30%.



(a) Error rate $r = 20\%$ (b) Error rate $r = 40\%$ (c) Error rate $r = 60\%$ (d) Error rate $r = 80\%$

Figure 7: Visualizations of the most sensitive distribution $\mathbb{Q}^\star$ with *0/1* loss function under different error rate threshold. We set $\theta_1 = 1.0$ and $\theta_2 = 0.25$ here.

- Similar to the phenomenon in Section 3, when $\theta_2 = +\infty$, the stability criterion only considers data corruptions; and when $\theta_1 = +\infty$, it only considers sub-population shifts.
- Different from Figure 2, since we use 0/1 loss here, the perturbed samples are all near the boundary.

Furthermore, for fixed θ_1 and θ_2, we vary the error rate threshold r and visualize the most sensitive distribution $\mathbb{Q}^\star$ in Figure 7. We set $\theta_1 = 1.0$ and $\theta_2 = 0.25$ and our stability criterion will consider both data corruptions and sub-population shifts.

C.2. Convergence of risk

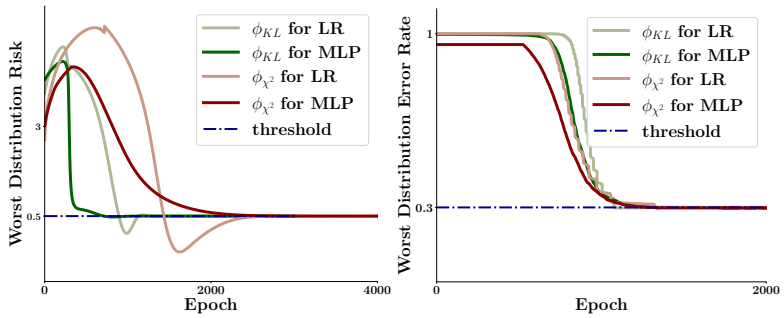
In Figure 8, we plot the curve of $\mathbb{E}_{\mathbb{Q}^{(t)}}[W \cdot \ell(\beta, Z)]$ with respect to the epoch number t. From the results, we could see that our solution could converge to the pre-defined threshold r.

D. Experiments

D.1. Datasets

Here we provide more details about our datasets used in Section 4.

- **ACS Income dataset**. The dataset is based on the American Community Survey (ACS) Public Use Microdata Sample (PUMS) (Ding et al., 2021). The task is to predict whether an individual's income is above \$50,000. We filter the dataset to only include individuals above the age of 16, usual working hours of at least 1 hour per week in the past year, and an income of at least \$100. The dataset contains individuals from all American states, and we focus on California (CA) in our experiments. We follow the data pre-processing procedures in (Liu et al., 2021). The dataset comprises a total of 76



(a) Continuous loss function. (b) 0/1 loss function.

Figure 8: The convergence of $\mathbb{E}_{\mathbb{Q}^{(t)}}[W \cdot \ell(\beta, Z)]$ w.r.t. epoch t. (a): Use continuous loss function with $r = 0.5$. (b): Use 0/1 loss function with $r = 30\%$. Here ϕ_{KL} denotes $\phi(t) = t \log t - t + 1$, and ϕ_{χ^2} denotes $\phi(t) = (t-1)^2$.

features, with the majority of categorical features being one-hot encoded to facilitate analysis. In our experiments, we sample 2,000 data points from CA for model training, and another 2,000 for evaluation. When involving algorithmic interventions in Section 4.2, we further sample 5,000 points to compare the performances of different algorithms.

- **ACS Public Coverage dataset**. The dataset is based on the American Community Survey (ACS) Public Use Microdata Sample (PUMS) (Ding et al., 2021). The task is to predict whether an individual has public health insurance. We focus on low-income individuals who are not eligible for Medicare by filtering the dataset to only include individuals under the age of 65 and with an income of less than $30,000. Similar to the ACS Income dataset, we focus on individuals from California in our experiments. We follow the data pre-processing procedures in (Liu et al., 2021). The dataset comprises a total of 42 features, with the majority of categorical features being one-hot encoded to facilitate analysis. In our experiments, we sample 2,000 data points from CA for model training, and another 2,000 for evaluation. When involving algorithmic interventions in Section 4.2, we further sample 5,000 points to compare the performances of different algorithms.
- **COVID-19 dataset**. The COVID-19 dataset contains COVID patients from Brazil, which is based on SIVEP-Gripe data (Baqui et al., 2020). It has 6882 patients from Brazil recorded between Februrary 27-May 4, 2020. There are 29 features in total, including comorbidities, symptoms, and demographic characteristics. The task is to predict the mortality of a patient, which is a binary classification problem. In our experiments, we split the dataset with a ratio of 1:1 for training and evaluation sets.

D.2. Algorithms under evaluation

In Section 4.1, we evaluate AT (Sinha et al., 2018) and Tilted ERM (Li et al., 2023). In Section 4.2, we introduce the Targeted AT as a simple algorithmic intervention. We provide their formulations here:

- AT:
$$\min_{\beta} \left\{ \mathbb{E}_{\mathbb{P}_0}[\phi_\gamma(\beta, Z)] := \mathbb{E}_{\mathbb{P}_0}\left[\sup_{z \in \mathcal{Z}} \ell(\beta, Z) - \gamma c(Z, \hat{Z}) \right] \right\}, \tag{12}$$
where $c(z, \hat{z}) = \|x - \hat{x}\|_2^2 + \infty \cdot |y - \hat{y}|$, and γ is the penalty hyper-parameter.
- Tilted ERM:
$$\min_{\beta} t \log \left(\mathbb{E}_{\mathbb{P}_0}\left[\exp\left(\frac{\ell(\beta, Z)}{t} \right) \right] \right), \tag{13}$$
where t is the temperature hyper-parameter.
- Targeted AT:
$$\min_{\beta} \left\{ \mathbb{E}_{\mathbb{P}_0}[\phi_\gamma(\beta, Z)] = \mathbb{E}_{\mathbb{P}_0}[\sup_{z \in \mathcal{Z}} \ell(\beta, Z) - \gamma c(Z, \hat{Z})] \right\}. \tag{14}$$
In this case, $c(z, \hat{z}) = \|z_{(i)} - \hat{z}_{(i)}\|_2^2 + \infty \cdot \|z_{(,-i)} - \hat{z}_{(,-i)}\|_2^2$, where $z_{(i)}$ denotes the target feature of z, and $z_{(,-i)}$ denotes all the other features. γ is the penalty hyper-parameter. By choosing this $c(z, \hat{z})$, the targeted AT will only perturb the target feature while keeping the others unchanged.

D.3. Training Details.

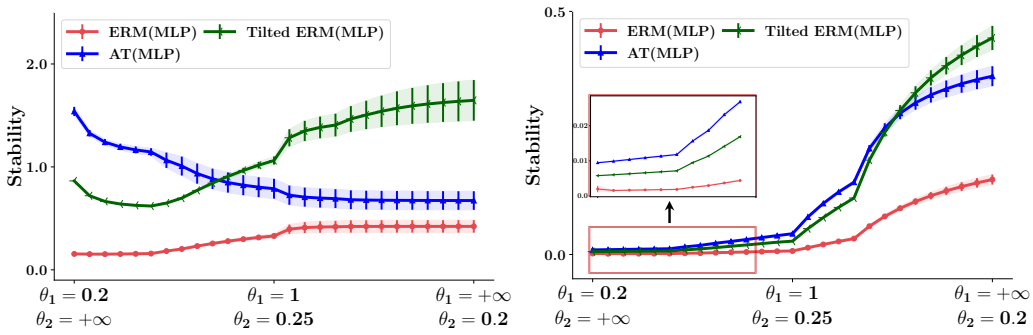
In our experiments, we use Logistic Regression (LR) for linear model and a two-layer MLP for neural network. We use PyTorch Library (Paszke et al., 2019) throughout our experiments. The number of hidden units of MLP is set to 16. As for the models under evaluation in Section 4, (1) for AT (Sinha et al., 2018), we vary the penalty parameter $\gamma \in \{0.1, 0.2, \ldots, 1.0\}$ and select the best γ according to the validation accuracy, and the inner optimization step is set to 20; (2) for Tilted ERM (Li et al., 2023), we vary the temperature parameter $t \in \{0.1, 0.2, \ldots, 1.0\}$ and select the best t according to the validation accuracy. Throughout all experiments, the ADAM optimizer with a learning rate of $1e^{-3}$ is used. All experiments are performed using a single NVIDIA GeForce RTX 3090.

D.4. More results on model sensitivity analysis

Here we provide more results on model sensitivity analysis in Section 4.1 in Figure 9. Similar to the setup in Section 4.1, for ACS PubCov and COVID-19 datasets, we vary the values of θ_1 and θ_2 and calculate the stability measure for MLP models trained with ERM, AT, and Tilted ERM. From Figure 9, the observations are consistent with those found in the main body (on ACS Income dataset):

- When θ_1 is small, our stability measure pays more attention to data corruptions. Therefore, AT performs better than Tilted ERM and ERM.
- When θ_2 is small, the main focus shifts to population shifts, where Tilted ERM is more preferred.

Furthermore, it is noteworthy that the standard deviation of the stability measure increases as θ_1 approaches infinitely (we set it to 100 in our experiments), especially for Tilted ERM. Recall that our stability criterion relies on both the evaluation data and the properties of the model under evaluation. The variability observed can be attributed to two primary sources of randomness. Firstly, there is the inherent randomness associated with the sampling process of the evaluation dataset. Secondly, the algorithmic randomness introduced by our proposed evaluation algorithm also contributes to this variability. When fixing the evaluation data, the standard derivations that reflect the randomness of our computing algorithm are relatively small. Therefore, sampling randomness is identified as the principal factor. Besides, the introduction of $\theta_1 = +\infty$ will incur a statistical cost in calculating the stability measure, as detailed in (Namkoong et al., 2022). This aspect leads to an increase in the randomness of our proposed algorithm.



(a) PubCov dataset. (b) COVID-19 dataset.

Figure 9: The stability measure for MLP models trained with ERM, AT, and Tilted ERM on ACS PubCov dataset and COVID-19 dataset.