Learning from a Biased Sample

Roshni Sahoo
rsahoo@stanford.edu

Lihua Lei
lihualei@stanford.edu

Stefan Wager
swager@stanford.edu

Stanford University

Abstract

The empirical risk minimization approach to data-driven decision making assumes that we can learn a decision rule from training data drawn under the same conditions as the ones we want to deploy it in. However, in a number of settings, we may be concerned that our training sample is biased, and that some groups (characterized by either observable or unobservable attributes) may be under- or over-represented relative to the general population; and in this setting empirical risk minimization over the training set may fail to yield rules that perform well at deployment. We propose a model of sampling bias called Γ-biased sampling, where observed covariates can affect the probability of sample selection arbitrarily much but the amount of unexplained variation in the probability of sample selection is bounded by a constant factor. Applying the distributionally robust optimization framework, we propose a method for learning a decision rule that minimizes the worst-case risk incurred under a family of test distributions that can generate the training distribution under Γ-biased sampling. We apply a result of Rockafellar and Uryasev to show that this problem is equivalent to an augmented convex risk minimization problem. We give statistical guarantees for learning a model that is robust to sampling bias via the method of sieves, and propose a deep learning algorithm whose loss function captures our robust learning target. We empirically validate our proposed method in simulations and a case study on ICU length of stay prediction.

1 Introduction

Empirical risk minimization is a practical and popular approach to learning data-driven decision rules [Bertsimas and Kallus, 2020, Kitagawa and Tetenov, 2018, Vapnik, 1995]. Formally, suppose that we observe $i = 1, \ldots, n$ samples (X_i, Y_i) independently drawn from a distribution P, where $X \in \mathcal{X}$ are covariates and $Y \in \mathcal{Y}$ is a target outcome, and we want to learn a decision rule h that minimizes a loss L under P:

$$h^* = \operatorname*{argmin}_h \mathbb{E}_{(X,Y)\sim P}\left[L(h(X), Y)\right]. \tag{1}$$

Then, empirical risk minimization involves choosing a decision rule $\hat{h}$ that is a (potentially penalized) minimizer of the in-sample loss $n^{-1}\sum_{i=1}^n L(h(X_i), Y_i)$; and the learned decision rule is deemed to perform well if the loss of $\hat{h}$ approaches the minimum possible loss that could be attained using h^* [Vapnik, 1995].

Formal justifications for empirical risk minimization crucially rely on the assumption that the target distribution we want to deploy our decision rule on, i.e., the one used to define the objective in (1), is the same as the distribution P from which we drew the training samples (X_i, Y_i) used for learning. In several important application areas, however, sampling bias in the data collection process may prevent practitioners from accessing training data from the distribution that they intend to deploy the rule on; and such sampling bias may hurt the target distribution performance of decision rules learned via empirical risk minimization.

One setting where sampling bias can be a concern is in designing risk predictors in a medical setting. Various risk predictors are widely used to guide both clinical practice and hospital logistics: Goff et al. [2014] discuss how risk predictors for cardiovascular disease are used to inform clinical guidelines, while Gul and

Draft version: September 2023. We are grateful for helpful comments from seminar participants at Harvard, Stanford, UC Berkeley, UC Irvine and UW Madison, and for advice from Alyssa Chen regarding the experimental setup of our MIMIC-III case study. Code available at https://github.com/roshni714/ru_regression.

Celik [2020] present a number of approaches to predicting emergency room admissions that can be used to anticipate hospital staffing needs. We may be concerned about sampling bias if these risk models are trained using data from a handful of hospitals (e.g., university hospitals participating in a study), and if the patients in these hospitals are not representative of the general patient population.

Another major situation where sampling bias may matter is in studies run on volunteers. In randomized trials for estimating treatment effects, participants often volunteer or apply to be a part of the study: Attanasio et al. [2011] measures the effect of a vocational training program on labor market outcomes in a randomized trial where participants needed to apply to be a part of the study; and the effectiveness of antidepressants is typically assessed in randomized trials involving volunteers [Wang et al., 2018]. In such studies, participants may differ from non-participants in fundamental ways, and a decision rule based on trial data may perform poorly when deployed on non-participants. Furthermore, vulnerable populations may be less exposed to study recruitment, so failure to generalize under sampling bias may cause these populations to be disproportionately impacted by errors of the decision rule.

The goal of this paper is to develop an alternative to empirical risk minimization that is robust to potential sampling bias. We still assume that we get to work with n i.i.d. samples from P; however, we now define the optimal decision rule in terms of a different distribution Q,

$$h^* = \operatorname*{argmin}_h \mathbb{E}_{(X,Y)\sim Q}\left[L(h(X), Y)\right], \tag{2}$$

and allow for the prospect that P may be biased relative to our target distribution Q. For example, in the context of medical risk prediction, Q could be the nationwide patient distribution, whereas P is the patient distribution in the hospitals we have data from.

Of course, if there is no link between our sampling distribution P and our target distribution Q, then learning data-driven rules is not possible. Throughout this paper, we will assume that P is biased—but not too biased—relative to Q, in the sense formalized by the Γ-biased sampling model given below. Here, $\Gamma \geq 1$ captures the allowed strength of sampling bias, and larger values of Γ allow for more bias.

Definition 1. Let $\Gamma \geq 1$. For any pair of distributions P and Q over (X, Y), we say that Q can generate P under Γ-biased sampling if there exists a distribution $\tilde{Q}$ over (X, Y, S), where $S \in \{0, 1\}$ is a "selection indicator" that satisfies the following properties: The (X, Y)-marginal of $\tilde{Q}$ is equal to Q, then (X, Y)-marginal of $\tilde{Q}$ conditionally on $S = 1$ is equal to P, and

$$\frac{\mathbb{P}_{\tilde{Q}}\left[S = 1 \mid X = x, Y = y\right]}{\mathbb{P}_{\tilde{Q}}\left[S = 1 \mid X = x\right]} \in [\Gamma^{-1}, \Gamma] \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}. \tag{3}$$

This model of sampling bias is an extension of the model adopted in Aronow and Lee [2013] and Miratrix et al. [2018] to the setting where there are covariates X that may affect whether a sample is selected ($S = 1$). In Γ-biased sampling, the covariates X may affect the probability of sample selection arbitrarily much, but the amount of unexplained variation in the probability of sample selection is limited due to (3). On the other hand, the model with $\Gamma = 1$ corresponds to "unconfounded sample selection" that is widely studied in the literature on generalizability [e.g., Stuart et al., 2011, Tipton, 2013, 2014].

One challenge of learning under Γ-biased sampling is that the true test distribution is unknown and there are many possible distributions that can generate the observed training distribution P under Γ-biased sampling. To address this problem, we use distributionally robust optimization (DRO) [Ben-Tal et al., 2013] to learn a decision rule that is robust to all distributions that can generate our observed training distribution P under Γ-biased sampling. The goal of DRO is to minimize the worst-case risk over a family of plausible test distributions $\mathcal{S}$ (the robustness set), i.e.

$$\operatorname*{argmin}_h \sup_{Q\in\mathcal{S}} \mathbb{E}_Q\left[L(h(X), Y)\right]. \tag{4}$$

To learn decision rules that are robust to Γ-biased sampling, we consider robustness sets $\mathcal{S}_\Gamma(P, Q_X)$ for $\Gamma > 1$, where $Q \in \mathcal{S}_\Gamma(P, Q_X)$ if Q can generate P via Γ-biased sampling and Q has covariate distribution equal to Q_X.

The main contribution of this work is a method for learning

$$h^*_\Gamma = \operatorname*{argmin}_h \sup_{Q\in\mathcal{S}_\Gamma(P,Q_X)} \mathbb{E}_Q\left[L(h(X), Y)\right] \tag{5}$$

for any distribution Q_X that is absolutely continuous to P_X. We show that there is a convex loss function L^{Γ}_{RU} (given below), defined over an augmented feature space, such that the solution to the following risk minimization problem

$$\begin{aligned}(h^*_\Gamma, \alpha^*_\Gamma) &= \underset{h,\alpha}{\operatorname{argmin}}\, \mathbb{E}_P\left[L^{\Gamma}_{\text{RU}}(h(X), \alpha(X), Y)\right], \\ L^{\Gamma}_{\text{RU}}(z, a, y) &= \Gamma^{-1} L(z, y) + (1 - \Gamma^{-1})a + (\Gamma - \Gamma^{-1})(L(z,\, y) - a)_+,\end{aligned} \tag{6}$$

with data drawn from the training distribution P, also solves (5) for any distribution Q_X that is absolutely continuous with respect to P_X. We call the minimization problem in (6) Rockafellar-Uryasev (RU) Regression and L^{Γ}_{RU} the RU loss because results of Rockafellar and Uryasev [2000] play a key role in our derivation of this loss function. A notable aspect of our proposed method is that it does not require any knowledge of Q_X because it relies on the fact that the minimization of the worst-case risk over a sufficiently flexible class of functions is equivalent to minimization of the conditional worst-case risk for every $x \in \mathcal{X}$.

The remainder of the paper investigates RU Regression theoretically and empirically. In Section 3.1, we demonstrate useful properties of the population RU risk, including convexity, differentiability, existence and uniqueness of the minimizer, and strong convexity around the minimizer. In Section 3.2, these properties enable us to derive formal guarantees for learning via empirical minimization of L^{Γ}_{RU} using the method of sieves [Geman and Hwang, 1982]. Furthermore, the useful properties of the population RU risk also suggest that for practical implementation, the problem in (6) can be solved via stochastic gradient descent. As a result, we propose to perform the optimization in (6) by joint-training of neural networks, one for each of h and α, with the RU loss as the objective. In Section 4, we validate our approach in simulations and a case study with the MIMIC-III dataset [Johnson et al., 2016a].

1.1 Related Work

Our proposed model of sampling bias, Γ-biased sampling, builds on previous models for sampling bias [Aronow and Lee, 2013, Miratrix et al., 2018], where samples Y_i are drawn i.i.d. from the target distribution Q but only included in the training dataset with a latent probability $\pi_i \in [\alpha, \beta]$, for $\alpha, \beta \in (0, 1]$. Under this model, these works focus on partial identification of the population mean outcome $\mathbb{E}_Q[Y]$. If we interpret $\pi_i := \mathbb{P}_{\tilde{Q}}\left[S_i \,\middle|\, X_i, Y_i\right]$, then our Γ-biased sampling model as specified in Definition 1 is statistically equivalent to an extension of the model used in Aronow and Lee [2013] and Miratrix et al. [2018] to include covariates, in such a way that we allow the unobserved probability of sample selection π to be arbitrarily affected by the covariates X but place bounds on the amount of unexplained variation in π_i. Also, unlike Aronow and Lee [2013] and Miratrix et al. [2018], we focus on the problem of learning a robust decision rule rather than on partial identification of moments of Q.

Our model is also connected to the broader literature on sensitivity analysis in causal inference [Andrews and Oster, 2019, Dorn et al., 2021, Jin et al., 2022, Nie et al., 2021, Yadlowsky et al., 2018], the goal of which is to understand how causal analyses justified by assuming randomized or unconfounded treatment assignment could be affected by a failure of these assumptions. In particular, our Γ-biased sampling model has a similar statistical structure as the Γ-marginal sensitivity model used by Tan [2006] to quantify failures of unconfoundedness. However, in these sensitivity analyses, the concern is typically regarding threats to internal validity (i.e., failures of unconfoundedness), whereas here we model sampling bias as a threat to external validity.

To learn a decision rule that is robust to Γ-biased sampling, we apply the DRO framework, which is widely used for learning models that are robust to unknown distribution shift [Duchi and Namkoong, 2021, Duchi et al., 2020, Hu et al., 2018, Michel et al., 2022, Mohajerin Esfahani and Kuhn, 2018, Oberst et al., 2021, Oren et al., 2019, Sagawa et al., 2019, Thams et al., 2022]. Previous works that apply DRO for learning robust models typically specify a robustness set of interest and provide a method for either evaluating the worst-case risk over the set, learning the solution that minimizes the worst-case risk over the set, or both. These works vary in how they define the robustness set and whether they consider robustness sets over the conditional distribution of Y given X, the marginal distribution over X, or the joint distribution over (X, Y). While learning under our Γ-sampling bias model is conceptually a DRO problem, our problem setting has many crucial differences from the most widely studied DRO setting, e.g., the one considered in Duchi et al.

[2020]; and these differences require us to develop new learning algorithms and new analysis techniques to prove formal results. We discuss connections to the DRO literature in more detail in Section 2.2.

Finally, our contribution is related to the broader literature on data-driven decision making. This literature has been active in recent years, including contributions from Athey and Wager [2021], Bertsimas and Kallus [2020], Elmachtoub and Grigas [2022], Foster and Syrgkanis [2019], Kallus and Zhou [2021], Kitagawa and Tetenov [2018], Manski [2004], Nie and Wager [2021], Stoye [2009], Swaminathan and Joachims [2015], Zhao et al. [2012] and Zhou et al. [2022]. A recurring theme of this line of work is in choosing loss functions $L(\cdot)$ that capture relevant aspects of various decision tasks [Bertsimas and Kallus, 2020]. Our results pair naturally with this line of work, in that our approach can be applied with generic loss functions to learn decision rules that are robust to potential sampling bias. We also draw attention to Kallus and Zhou [2021], who consider learning optimal treatment rules from confounded data, i.e., where the "treated" and "control" samples available for training may be biased according to unobservable attributes. Our work is related to that of Kallus and Zhou [2021] in that we both consider using robust optimization techniques to learn from data potentially corrupted via biased sampling; however, the type of bias we consider (test/train vs. treatment/control), and resulting algorithmic and conceptual remedies, are different.

2 Rockafellar-Uryasev Regression

We consider the following general loss minimization setting. We have access to $i = 1, \ldots, n$ samples (X_i, Y_i) drawn from a training distribution P over $\mathcal{X} \times \mathcal{Y}$. We seek to learn a decision rule h such that, given a loss function $L(z, y)$, the expected loss $\mathbb{E}_Q[L(h(X), Y)]$ is small when (X, Y) is drawn from our target distribution Q. The key challenge is that we do not assume that the training distribution P and the target Q are the same. Rather, Q is unknown, and we only assume that Q generates P via Γ-biased sampling in the sense of Definition 1, for some $\Gamma \geq 1$.

For any marginal covariate distribution Q_X, let $S_\Gamma(P, Q_X)$ denote the set of all distributions Q are related to P via Γ-biased sampling and have marginal distribution over X equal to Q_X. (Recall that Definition 1 only bounds "unexplained" sampling bias allows; it allows for arbitrary sampling bias explained by X, and thus places no meaningful restrictions on Q_X). Given any choice of Q_X, we target the robust decision rule

$$h^*_{Q_X,\Gamma} \in \operatorname*{argmin}_{h \in L^2(P_X, \mathcal{X})} \sup \left\{ \mathbb{E}_Q[L(h(X), Y)] : Q \in S_\Gamma(P, Q_X) \right\}, \tag{7}$$

where $L^2(P_X, \mathcal{X})$ denotes the space of square-integrable measurable functions with respect to P_X. The formulation (7) may look challenging to use as the basis for a practical approach to learning. First, it is formulated in terms of the marginal distribution Q_X which may sometimes be known [e.g., Nie et al., 2021], but often is not known. Second, the optimization problem (7) has a min-max form that is not obviously amenable to statistical learning. The following results show how both of these challenges can be addressed.

As a preliminary to our subsequent analysis, we start by giving a more explicit characterization of the set $\mathcal{S}_\Gamma(P, Q_X)$ that can generate P under Γ-based sampling: Q can generate P via Γ-biased sampling if and only if the likelihood ratio between the conditional distributions of $Y \mid X$ of Q and P is bounded between Γ^{-1} and Γ and the density ratio between the covariate distributions of P and Q are bounded.

Lemma 1. *Let P, Q be the distributions over (X, Y). Q can generate P via Γ-biased sampling if and only if*

$$\Gamma^{-1} \leq \frac{dQ_{Y|X=x}(y)}{dP_{Y|X=x}(y)} \leq \Gamma, \quad \forall x \in \mathcal{X}, y \in \mathcal{Y} \tag{8}$$

and $\sup_{x \in \mathcal{X}} \frac{dP_X(x)}{dQ_X(x)} < C$ for some $C < \infty$. Proof in Appendix C.3.

We are now ready to spell out our first main result, i.e., a reformulation of our learning objective (7) as the minimizer of the expectation a convex function over data drawn from the training distribution P. We demonstrate that there exists a single function h^*_Γ that solves the problem (7) simultaneously for any Q_X that is absolutely continuous with respect to P_X, and furthermore this h^*_Γ can be characterized as the minimizer of a convex loss defined in terms of the observed data distribution P. We refer to the loss function in (9) the Rockafellar-Uryasev (RU) loss because the proof of Theorem 2 draws heavily from results of Rockafellar and Uryasev [2000]; we will also refer to learning via empirical minimization based on (9) as RU Regression.

Theorem 2. *Suppose that $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ are drawn i.i.d. with respect to a distribution P for some $\mathcal{X} \subseteq \mathbb{R}^d$ and $Y \subseteq \mathbb{R}$. Let $L(z, y)$ be a loss function that is convex in z for any $y \in \mathcal{Y}$, and let $\Gamma > 1$. Then the following augmented loss function,*

$$L_{RU}^{\Gamma}(z, a, y) = \Gamma^{-1} L(z, y) + (1 - \Gamma^{-1})a + (\Gamma - \Gamma^{-1})(L(z, y) - a)_+, \tag{9}$$

is convex is (z, a) for any $y \in \mathcal{Y}$. Furthermore, any solution

$$\{h_{\Gamma}^*, \alpha_{\Gamma}^*\} \in \underset{(h,\alpha) \in L^2(P_X, \mathcal{X}) \times L^2(P_X, \mathcal{X})}{\operatorname{argmin}} \mathbb{E}_P \left[L_{RU}^{\Gamma}(h(X), \alpha(X), Y) \right] \tag{10}$$

is also a solution to (7) for any Q_X that is absolutely continuous with respect to P_X, i.e., $Q_X \ll P_X$ and $\sup_{x \in \mathcal{X}} \frac{dP_X(x)}{dQ_X(x)} < \infty$. Proof in Section 2.1.

A proof of Theorem 2 is given in Section 2.1. We define notation that is used in the proof, as well as the remainder of the paper. Let $F_{x;h(x)}(z)$ be the c.d.f. of $L(h(x), Y)$, where Y is distributed according to $P_{Y|X=x}$. In other words, $F_{x;h(x)}(z)$ is the distribution over the conditional losses when $X = x$. Define the function $q_{\eta}^{L}(x; h(x))$ to be the η-th quantile of distribution over the conditional losses when $X = x$, i.e.

$$q_{\eta}^{L}(x; h(x)) = F_{x;h(x)}^{-1}(\eta). \tag{11}$$

Also, define

$$\eta(\Gamma) = \frac{\Gamma}{\Gamma + 1}. \tag{12}$$

2.1 Proof of Theorem 2

For the first claim, the convexity of L_{RU}^{Γ} follows immediately using the standard rules for composing convex functions [Boyd and Vandenberghe, 2004]. We focus on the second claim of Theorem 2. We use the following lemma to rewrite our worst-case population risk minimization problem in (7) as a worst-case conditional risk minimization.

Lemma 3. *A function $h \in L^2(P_X, \mathcal{X})$ solves (7) iff h solves*

$$\min_{h(x) \in \mathbb{R}} \sup \left\{ \mathbb{E}_{Q_{Y|X}} \left[L(h(x), Y) \mid X = x \right] : Q \in S_{\Gamma}(P, Q_X) \right\} \tag{13}$$

for every $x \in supp(P_X)$. Proof in Appendix D.1.

Using Lemma 1, we can characterize the distributions in $\mathcal{S}_{\Gamma}(P, Q_X)$ as distributions for which (8) holds and $\sup_{x \in \mathcal{X}} \frac{dP_X(x)}{dQ_X(x)} < C$ for $C < \infty$. By the Neyman-Pearson lemma, we can verify for any decision rule h,

$$\begin{aligned} &\sup\{\mathbb{E}_{Q_{Y|X}} \left[L(h(X), Y) \mid X = x \right] : Q \in S_{\Gamma}(P, Q_X)\} \\ &\quad = \mathbb{E}_{P_{Y|X}} \left[L(h(X), Y) \Big(\Gamma^{-1} + (\Gamma - \Gamma^{-1}) \mathbb{I}(L(h(X), Y) \geq q_{\eta(\Gamma)}^{L}(X; h(X))) \Big) \mid X = x \right], \end{aligned} \tag{14}$$

where $q_{\eta}^{L}(x; h(x))$ is as defined in (11) and $\eta(\Gamma)$ is as defined in (12). (13) can be rewritten as

$$\min_{h(x) \in \mathbb{R}} \mathbb{E}_{P_{Y|X}} \left[L(h(x), Y) \Big(\Gamma^{-1} + (\Gamma - \Gamma^{-1}) \mathbb{I}(L(h(x), Y) \geq q_{\eta(\Gamma)}^{L}(X; h(x))) \Big) \mid X = x \right]. \tag{15}$$

Thus, we can focus on the optimization problem in (15).

We realize that the objective in (15) is closely related to the conditional value-at-risk (CVaR) [Rockafellar and Uryasev, 2000], which is widely considered in the finance literature. For a continuous random variable W with quantile function (inverse c.d.f.) q_W and $\eta \in (0, 1)$, the η-CVaR of W is given by

$$\mathrm{CVaR}_{\eta}(W) = \mathbb{E} \left[W \mid W \geq q_W(\eta) \right].$$

Applying the CVaR definition, we realize that

$$\mathbb{E}_{P_{Y|X}}\left[L(h(X),Y)\mathbb{I}(L(h(X),Y) > q_\eta^L(X;h(X))) \mid X=x\right] = (1-\eta(\Gamma))\cdot \mathrm{CVaR}_{\eta(\Gamma)}(L(h(x),Y)). \tag{16}$$

Substituting (16) into (15) and simplifying gives the following problem

$$\min_{h(x)\in\mathbb{R}} \Gamma^{-1}\mathbb{E}_{P_{Y|X}}\left[L(h(x),Y) \mid X=x\right] + (1-\Gamma^{-1})\cdot \mathrm{CVaR}_{\eta(\Gamma)}(L(h(x),Y)). \tag{17}$$

We are now ready to use the influential result of Rockafellar and Uryasev [2000, Theorem 1], which in our setting implies that the CVaR of the loss itself can be formulated as the solution to a convex optimization problem:

$$\mathrm{CVaR}_\eta(L(h,Y)) = \min_{\alpha\in\mathbb{R}} \alpha + (1-\eta)^{-1}\mathbb{E}_Y\left[(L(h,Y)-\alpha)_+\right] \tag{18}$$

for a loss $L(h,Y)$ that depends on $h \in H \subset \mathbb{R}$ and Y, a random variable with a density. Thus, we can rewrite the term $\mathrm{CVaR}_{\eta(\Gamma)}(L(h(x),Y))$ from (17) as

$$\mathrm{CVaR}_{\eta(\Gamma)}(L(h(x),Y)) = \min_{\alpha(x)\in\mathbb{R}} \alpha(x) + \frac{1}{1-\eta(\Gamma)}\mathbb{E}_{P_{Y|X}}\left[(L(h(x),Y)-\alpha(x))_+ | X=x\right]. \tag{19}$$

Furthermore, by Theorem 2 of Rockafellar and Uryasev [2000], any minimizer of the following joint optimization also minimizes the CVaR. In particular,

$$\min_{h\in H} \mathrm{CVaR}_\eta(L(h,Y)) = \min_{(h,\alpha)\in H\times\mathbb{R}} \alpha + (1-\eta)^{-1}\mathbb{E}_Y\left[(L(h,Y)-\alpha)_+\right]. \tag{20}$$

Applying this theorem to (17), we have that

$$\begin{aligned}
&\operatorname*{argmin}_{h(x)\in\mathbb{R}} \Gamma^{-1}\cdot\mathbb{E}_{P_{Y|X}}\left[L(h(X),Y)\right) \mid X=x] + (1-\Gamma^{-1})\cdot \mathrm{CVaR}_{\eta(\Gamma)}(L(h(x),Y)) \\
&\quad= \operatorname*{argmin}_{h(x),\alpha(x)\in\mathbb{R}} \Gamma^{-1}\cdot\mathbb{E}_{P_{Y|X}}\left[L(h(x),Y))\mid X=x\right] \\
&\qquad\qquad + (1-\Gamma^{-1})\cdot\left(\alpha(x) + \frac{1}{1-\eta(\Gamma)}\mathbb{E}_{P_{Y|X}}\left[(L(h(x),Y)-\alpha(x))_+|X=x\right]\right) \\
&\quad= \operatorname*{argmin}_{h(x),\alpha(x)\in\mathbb{R}} \Gamma^{-1}\cdot\mathbb{E}_{P_{Y|X}}\left[L(h(x),Y))\mid X=x\right] + (1-\Gamma^{-1})\alpha(x) \\
&\qquad\qquad + (\Gamma-\Gamma^{-1})\mathbb{E}_{P_{Y|X}}\left[(L(h(x),Y)-\alpha(x))_+|X=x\right] \\
&\quad= \operatorname*{argmin}_{h(x),\alpha(x)\in\mathbb{R}} \mathbb{E}_{P_{Y|X=x}}\left[L_{\mathrm{RU}}^\Gamma(h(x),\alpha(x),Y) \mid X=x\right].
\end{aligned}$$

The last line follows from the definition of L_{RU}^Γ in (9). In other words, (17) can be written as the augmented conditional risk minimization

$$\min_{h(x),\alpha(x)\in\mathbb{R}} \mathbb{E}_{P_{Y|X}}\left[L_{\mathrm{RU}}^\Gamma(h(x),\alpha(x),Y) \mid X=x\right]. \tag{21}$$

Functions $h_\Gamma^*, \alpha_\Gamma^*$ that solve (21) also solve (10). In addition, any minimizer of (21) also solves (13) for any $x \in \mathrm{supp}(P_X)$. The following lemma gives that functions that minimize (13) for every $x \in \mathrm{supp}(P_X)$ also minimize (7).

Lemma 4. *Suppose $\tilde{h}, \tilde{\alpha}$ solve (13) for every $x \in \mathrm{supp}(P_X)$. Then $\tilde{h}$ solves (7) for any Q_X such that $Q_X \ll P_X$ and $\sup_{x\in\mathcal{X}} \frac{dP_X(x)}{dQ_X(x)} < \infty$. Proof in Appendix D.2.*

Thus, we can finally conclude that $h_\Gamma^*, \alpha_\Gamma^*$ also minimize (7). In other words, our robust optimization problem in (7) can be formulated as (10), a risk minimization problem under the training distribution P that involves learning an auxiliary function α along with the decision rule h.

A key aspect of RU Regression is that the optimal decision rule is agnostic to the test covariate distribution Q_X as long as it is absolutely continuous with respect to the training covariate distribution P_X. This is

because we propose to learn the minimizer of the worst-case loss *conditionally for every* $x \in \mathcal{X}$. The minimizer is a conditional quantity. We can simply study (10) to learn a decision rule that is robust to conditional shifts of the form in (8) and almost arbitrary covariate shifts.

In order for conditional risk minimization to be equivalent to the population risk minimization, we require the decision rule h and auxiliary function α to come from a flexible class, such as $L^2(P_X, \mathcal{X})$. For practical implementation, in Section 4, we propose to use joint optimization of deep neural networks to learn the solution of (10). We will use one neural network to represent h and another neural network to represent α and train the networks with the RU loss using a standard optimization algorithm, such as stochastic gradient descent or its variants.

2.2 Connections to Other DRO Frameworks

As discussed above, our approach to learning decision rules under unknown conditional shifts yields a distributionally robust optimization (DRO) problem [Ben-Tal et al., 2013]. One recent paper providing general results for DRO in a statistical learning setting is Duchi and Namkoong [2021], who consider worst-case shifts in the joint distribution over (X, Y) and robustness sets that are f-divergence balls about the training distribution P. Specifically under their learning objective, the optimal decision rule would be

$$h^* = \operatorname*{argmin}_h \sup \Big\{ \mathbb{E}_Q \left[L(h(X), Y) \right] : D_f(Q|P) \leq \rho \Big\}, \quad D_f(Q|P) = \int f\Big(\frac{dQ}{dP}\Big) dP, \tag{22}$$

where D_f is an f-divergence. Clearly, this problem is conceptually related to (7); however, there are a number of key differences that require new ideas both in terms of learning algorithms and analysis techniques.

A first difference between our DRO problem (7) and (22) is the robustness sets that are considered in each problem. To cast (8) as a constraint of the form $D_f(Q|P) \leq \rho$, we would need to consider an "improper" f-divergence, i.e. with

$$f(z) = \begin{cases} 0 & \Gamma^{-1} \leq z \leq \Gamma \\ \infty & \text{else} \end{cases}. \tag{23}$$

The fact that this function is discontinuous and unbounded means that the formal results (and proof strategies) of Duchi and Namkoong [2021] cannot be applied in our setting.

A second difference between our DRO problem (7) and (22) is that our problem can be solved by a method that jointly optimizes over the arguments of a convex risk minimization problem (10). Our result from Theorem 2 superficially resembles the dual formulation of (22):

$$\operatorname*{argmin}_h \inf_{\lambda, \eta \geq 0} \Big\{ \mathbb{E}_P \left[\lambda f^* \Big(\frac{L(h(X), Y) - \eta}{\lambda} \Big) \right] + \lambda \rho + \eta \Big\}, \tag{24}$$

where f^* is the Fenchel conjugate of f. However, comments in Namkoong and Duchi [2016] suggest that for general f-divergences, joint optimization algorithms for solving (24) would be ill-conditioned due to the dependence on λ^{-1} in the first term. In contrast, for the improper function f (23) relevant to our problem, $f^*(u) = \Gamma(u)_+ - \Gamma^{-1}(u)_-$. For this particular choice of f, λ can be removed from the optimization problem (24). Our approach exploits special structure in our distribution shift model that is not present in the problems studied in Duchi and Namkoong [2021].

A third difference between our DRO problem (7) and the problem in (22) is that (7) involves constraints on the distribution shift that hold conditionally on x, instead of a constraint on the shift in the joint distribution over (X, Y). Many previous DRO works consider robustness sets that constrain the shift in the joint distribution over (X, Y) [Duchi and Namkoong, 2021, Duchi et al., 2020, Hu et al., 2018, Michel et al., 2022, Mohajerin Esfahani and Kuhn, 2018, Oren et al., 2019, Sagawa et al., 2019] or the marginal distribution over X [Duchi et al., 2020]. However, our motivating problem yields constraints on conditional shifts that must hold simultaneously for every x, which results in a different and substantially more complicated optimization problem that requires more delicate methods and analysis. For example, Levy et al. [2020] proposes a mini-batch gradient-descent algorithm for learning the solution to (22); however, this algorithm cannot be used with conditional constraints (unless one can gather multiple observations for every x, which is impossible for continuous-valued x).

A handful of recent works consider robustness sets that place restrictions on conditional shifts [Esteban-Pérez and Morales, 2021, Oberst et al., 2021, Thams et al., 2022]. Esteban-Pérez and Morales [2021] takes statistical uncertainty to be the source of the distribution shift and considers shifts in the empirical conditional distribution for subsets of $\mathcal{X}$ with sufficiently large measure. In contrast, we consider sampling bias, which is present even in the population case with infinite samples, as the source of the distribution shift we seek to be robust against. Furthermore, our problem also requires placing constraints on the conditional shift for every x, not just subsets of $\mathcal{X}$. Oberst et al. [2021] leverages access to noisy proxies of unobserved variables for learning models that are robust to shifts in the distribution of unobservables. Thams et al. [2022] studies how to evaluate the worst-case loss under a parametric robustness set, which consists of interpretable, conditional shifts. Our work differs from Oberst et al. [2021], Thams et al. [2022] in that we do not make any fine-grained assumptions on the nature of the shift, such access to proxy variables or a parametric form. We note that the challenge of considering robustness sets that enforce conditional restrictions has also recently been considered in the literature on sensitivity analysis in causal inference [Dorn et al., 2021, Jin et al., 2022, Nie et al., 2021, Yadlowsky et al., 2018].

3 Theoretical Guarantees

In the previous section, we showed that minimax decision rule under Γ-biased sampling could be expressed as the population minimizer of a convex loss function over an augmented function space. This is helpful in understanding what the minimax decision rule looks like—and suggests that the corresponding DRO problem may be tractable. In practice, however, we of course do not have access to the full sampling distribution P, and need to choose our decision rule based on a finite (random) sample from it. Here, we investigate the properties of learning algorithms that leverage the representation result derived above, and learn decision rules via empirical minimization using the loss function L^{Γ}_{RU} given in (6).

One challenge in doing so is that $L^{\Gamma}_{\text{RU}}(z,\, a,\, y)$ is not strongly convex in (z, a); and in fact is not even strongly convex in expectation when $a < 0$. The following results show, however, that the expected RU risk has a unique minimizer—and is strongly convex and smooth in a neighborhood around the minimizer. These properties enable us to obtain nonparametric estimation guarantees by applying the method of sieves in Section 3.2. Overall, our results suggest that L^{Γ}_{RU} has sound statistical properties in finite samples, and thus that empirical minimization using this loss function is a promising approach to learning minimax decision rules under Γ-biased sampling.

3.1 Properties of Population RU Risk

First, we consider the problem of minimizing the population RU risk with respect to (h, α) over $L^2(P_X, \mathcal{X}) \times L^2(P_X, \mathcal{X})$. We consider the following norm on this product space

$$||(h, \alpha)||_{L^2(P_X,\mathcal{X})} = \sqrt{||h||^2_{L^2(P_X,\mathcal{X})} + ||\alpha||^2_{L^2(P_X,\mathcal{X})}}.$$

Under the following two assumptions, we can show that any minimizer of the population RU risk lies in a bounded subset of $L^2(P_X, \mathcal{X}) \times L^2(P_X, \mathcal{X})$.

Assumption 1. $\mathcal{X} \subset \mathbb{R}^d, \mathcal{Y} \subset \mathbb{R}$, and $\mathcal{X} \times \mathcal{Y}$ is compact.

Assumption 2. The loss function $L(\hat{y}, y) = \ell(y - \hat{y})$ for some function $\ell(z)$ that is $C_{L,l}$-strongly convex, twice-differentiable and is minimized at $\ell(0) = 0$.

Since $\mathcal{Y}$ is bounded, $\mathcal{Y} \subset [-B, B]$. We can define a bounded class of decision rules

$$\mathcal{H} = \{h \in L^2(P_X, \mathcal{X}) \mid ||h||_\infty \leq 2B\}.$$

We define a constant M_u such that

$$\sup_{h \in \mathcal{H}, x \in \mathcal{X}} q^L_{\eta(\Gamma)}(x; h(x)) < M_u, \tag{25}$$

and note that $M_u < \infty$ because $\mathcal{H}$ is bounded and $\mathcal{X} \times \mathcal{Y}$ is compact. We define the bounded class $\mathcal{A}$ for the auxiliary functions

$$\mathcal{A} = \{\alpha \in L^2(P_X, \mathcal{X}) \mid 0 \leq \alpha(x) \leq M_u \quad \forall x \in \mathcal{X}\}.$$

Let $\Theta = \mathcal{H} \times \mathcal{A}$. In the following result, we show that minimizing the population RU risk over $L^2(P_X, \mathcal{X}) \times L^2(P_X, \mathcal{X})$ is equivalent to minimizing the population RU risk over Θ.

Lemma 5. *Under Assumption 1, 2, if any minimizer of* $(h, \alpha) \mapsto \mathbb{E}_P\left[L^\Gamma_{RU}(h(X), \alpha(X), Y)\right]$ *exists over* $L^2(P_X, \mathcal{X}) \times L^2(P_X, \mathcal{X})$*, then it must lie in* Θ*. Proof in Appendix C.4.*

From now on, we will only consider minimization of the population RU risk over Θ. We can show that the population RU risk has at least one minimizer on Θ.

Lemma 6. *Under Assumption 1, 2,* $\mathbb{E}_P\left[L^\Gamma_{RU}(h(X), \alpha(X), Y)\right]$ *has at least one minimizer on* Θ*.*

To show that the population RU risk is strictly convex on Θ, we make the following assumption on the conditional distribution $P_{Y|X=x}$.

Assumption 3. For every $x \in \mathcal{X}$, we assume that $P_{Y|X=x}(y)$ is differentiable and strictly increasing in its argument and has positive density on $\mathcal{Y}$. We assume that $\sup_{x \in \mathcal{X}, y \in \mathbb{R}} p_{Y|X=x}(y) \leq C_{p,u}$, where $0 < C_{p,u} < \infty$.

Lemma 7. *Under Assumptions 1, 2, 3,* $\mathbb{E}_P\left[L^\Gamma_{RU}(h(X), \alpha(X), Y)\right]$ *is strictly convex in* (h, α) *on* Θ*. Proof in Appendix C.6.*

As a consequence of strict convexity on Θ, the population RU risk must have at most one minimizer over Θ. Meanwhile, Lemma 6 gives that it has at least one minimizer over Θ, as well. Combining these results gives that the population RU risk has a unique minimizer over Θ. Because of Lemma 5, this means that the population RU risk also has a unique minimizer over all of $L^2(P_X, \mathcal{X}) \times L^2(P_X, \mathcal{X})$.

Theorem 8. *Under Assumptions 1, 2, 3,* $\mathbb{E}_P\left[L^\Gamma_{RU}(h(X), \alpha(X), Y)\right]$ *has a unique minimizer* $(h^*_\Gamma, \alpha^*_\Gamma)$ *over* Θ*. Proof in Appendix C.7.*

In addition, we can develop an interpretation of α^*_Γ that minimizes the population RU risk.

Lemma 9. *Under Assumptions 1, 2, 3,*

$$\alpha^*_\Gamma(x) = q^L_{\eta(\Gamma)}(x; h^*_\Gamma(x)),$$

and there exists $M_l > 0$ *such that*

$$\alpha^*_\Gamma(x) > M_l \quad \forall x \in \mathcal{X}.$$

Proof in Appendix C.8.

Using Lemma 9, we can show that the population RU risk is strongly convex near the minimizer. We define constants that will be used in the proof of strong convexity. Recall that under Assumption 2, we can rewrite $L(\hat{y}, y) = \ell(y - \hat{y})$. Let ℓ_1^{-1} be the inverse of $\ell(z)$ where $z > 0$. Let ℓ_2^{-1} be the inverse of $\ell(z)$ where $z \leq 0$. Define

$$C_{a,u} := \sup_{i \in \{1,2\}} |\ell'(\ell_i^{-1}(M_u))|, \tag{26}$$

$$C_{a,l} := \inf_{i \in \{1,2\}} |\ell'(\ell_i^{-1}(M_l))|. \tag{27}$$

To define the next set of constants, we define $q^Y_c(x)$ to be the c-th quantile of Y where Y is distributed according to $P_{Y|X=x}$.

$$C_{p,l} := \inf_{c \in [1 - \frac{\eta(\Gamma)}{2}, 1 + \frac{\eta(\Gamma)}{2}], x \in \mathcal{X}} p_{Y|X=x}(q^Y_c(x)), \tag{28}$$

$$\kappa_1 := (1 - \Gamma^{-1}) \cdot \frac{C_{L,l} \cdot C_{p,l}}{2C_{p,u} \cdot (C_{a,u} \cdot C_{a,l} + 1) + C_{L,l} \cdot C_{a,l}} \cdot \frac{C_{a,l}}{C_{a,u}}. \tag{29}$$

Additionally, let

$$C_{a,l,\delta} := \inf_{i\in\{1,2\}} |\ell'(\ell_i^{-1}(M_l - \delta))|, \tag{30}$$

$$C_{p,l,\epsilon} := \inf_{c\in[1-\frac{\eta(\Gamma)}{2},1+\frac{\eta(\Gamma)}{2}],b\in[-\epsilon,\epsilon],x\in\mathcal{X}} p_{Y|X=x}(q_c^Y(x)+b). \tag{31}$$

We can show that in a $||\cdot||_\infty$-ball about the minimizer, the population RU loss is strongly convex, where the constant of strong convexity approaches κ_1 as the ball's radius shrinks.

Theorem 10. *Suppose Assumptions 1, 2, 3, hold. Let* $\mathcal{C}_\delta = \{(h,\alpha)\in\Theta \mid ||(h,\alpha)-(h^*_\Gamma,\alpha^*_\Gamma)||_\infty < \delta\}$, *and let* $0<\epsilon<\frac{1-\eta(\Gamma)}{2C_{p,u}}$. *There exists* $0<\delta(\epsilon)<M_l$ *such that* $\mathbb{E}_P\left[L^\Gamma_{RU}(h(X),\alpha(X),Y)\right]$ *is* $\kappa_{1,\epsilon}$*-strongly convex on* $\mathcal{C}_{\delta(\epsilon)}$, *where*

$$\kappa_{1,\epsilon} := (\Gamma-\Gamma^{-1})\frac{C_{L,l}\cdot(1-\eta(\Gamma)-2C_{p,u}\cdot\epsilon)\cdot C_{p,l,\epsilon}}{2C_{p,u}\cdot(C_{a,u}\cdot C_{a,l,\delta(\epsilon)}+1)+C_{L,l}\cdot C_{a,l,\delta(\epsilon)}}\cdot\frac{C_{a,l,\delta(\epsilon)}}{C_{a,u}}. \tag{32}$$

As $\epsilon\to 0$,

$$\kappa_{1,\epsilon}\to\kappa_1.$$

Proof in Appendix C.9.

To show that the population RU risk is smooth in an $||\cdot||_\infty$-ball around the minimizer, we require an additional assumption on the loss function L. Essentially, we need L to be $C_{L,u}$-smooth for some constant $0<C_{L,u}<\infty$.

Assumption 4. The second derivative of $\ell(z)$ as defined in Assumption 2 is upper bounded by $C_{L,u}$, where $0<C_{L,u}<\infty$.

The constant for smoothness depends on the constant $C_{p,u}$ from Assumption 3, $C_{a,u}$ from (26), $C_{a,l,\delta}$ from (30), and $C_{L,u}$ from Assumption 4. Let

$$\kappa_2 := (\Gamma-\Gamma^{-1})\cdot\left(2C_{p,u}\left(C_{a,u}+\frac{1}{C_{a,l}}\right)\right)+\Gamma\cdot C_{L,u}. \tag{33}$$

We can show that in an $||\cdot||_\infty$ ball about the minimizer, the population RU risk is smooth, where the constant for smoothness of the population RU risk approaches κ_2 as the radius of the ball decreases.

Theorem 11. *Suppose Assumptions 1, 2, 3, 4 hold. Let* $\mathcal{C}_\delta = \{(h,\alpha)\in\Theta \mid ||(h,\alpha)-(h^*_\Gamma,\alpha^*_\Gamma)||_\infty<\delta\}$. *For every* $0<\epsilon<\frac{1-\eta(\Gamma)}{2C_{p,u}}$, *there is* $0<\delta(\epsilon)<M_l$ *such that* $\mathbb{E}_P\left[L_{RU}(h(X),\alpha(X),Y)\right]$ *is* $\kappa_{2,\epsilon}$*-smooth in* (h,α) *on* $\mathcal{C}_{\delta(\epsilon)}$ *where*

$$\kappa_{2,\epsilon} := (\Gamma-\Gamma^{-1})\cdot\left(2C_{p,u}\left(C_{a,u}+\frac{1}{C_{a,l,\delta(\epsilon)}}\right)\right)+\Gamma\cdot C_{L,u}. \tag{34}$$

As $\epsilon\to 0$,

$$\kappa_{2,\epsilon}\to\kappa_2.$$

Proof in Appendix C.10.

3.2 Estimation Guarantees via Method of Sieves

To simplify notation, we denote $\theta := (h,\alpha)$ and rewrite the population RU risk as $\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(\theta(X),Y)\right]$. The empirical risk is accordingly

$$\widehat{\mathbb{E}}_P\left[L^\Gamma_{\mathrm{RU}}(\theta(X),Y)\right] = \frac{1}{n}\sum_{i=1}^n L^\Gamma_{\mathrm{RU}}(\theta(X_i),Y_i). \tag{35}$$

In addition, we will denote the minimizer of the population RU risk as simply $\theta^* := (h^*_\Gamma,\alpha^*_\Gamma)$, omitting the dependence on Γ.

Thus far, we have demonstrated that θ^* is the minimizer of the population RU risk over the infinite-dimensional space Θ. In practice, we aim to minimize the empirical RU loss (35). However, due to the computational difficulties of estimating infinite-dimensional models using finite samples, we do not minimize the empirical risk over Θ directly. Instead, we apply the method of sieves [Geman and Hwang, 1982]; we consider optimizing the empirical risk over an increasing sequence of sieves $\Theta_1 \subseteq \Theta_2 \subseteq \cdots \subseteq \Theta$, which are finite-dimensional parameter spaces. The sieves we consider have the property that $\inf_{\theta \in \Theta_m} ||\theta - \theta^*||_\infty \to 0$ as $m \to \infty$. To ensure consistency, we increase the complexity of the sieves with the sample size. We let

$$\hat{\theta}_n = \underset{\theta \in \Theta_n}{\operatorname{argmin}} \widehat{\mathbb{E}}_P \left[L^{\Gamma}_{\mathrm{RU}}(\theta(X), Y) \right].$$

To estimate h, it is sufficient to consider sieves that consist functions bounded between $-2B$ and $2B$. To estimate α, it is sufficient to consider sieves that consist of nonnegative, bounded functions because any minimizer of the population RU risk has $0 \leq \alpha^*(x) \leq M_u$ for all $x \in \mathcal{X}$. In order to make our sieve-based estimates $h(x), \alpha(x)$ be bounded, we use the same strategy as in Jin et al. [2022]; we truncate standard sieve space to bounded functions. The following two natural examples of truncated sieve spaces were also discussed by Jin et al. [2022]:

Example 2 (Polynomials)**.** Let $\mathrm{Pol}(J_n)$ be the space of polynomials on $[0,1]$ of degree J_n or less; that is

$$\mathrm{Pol}(J_n) = \Big\{ x \mapsto \sum_{k=0}^{J_n} a_k x^k, x \in [0,1] : a_k \in \mathbb{R} \Big\}.$$

Let $\mathrm{Pol}(J_n, a, b)$ be the space of polynomials on $[0,1]$ of degree J_n or less that are bounded between a and b; that is

$$\mathrm{Pol}(J_n, a, b) = \Big\{ x \mapsto \min(\max(f(x), a), b), x \in [0,1] : f \in \mathrm{Pol}(J_n) \Big\}.$$

Then, we define the sieve *with truncation* as $\Theta_n = \mathcal{H}_n \times \mathcal{A}_n$, where $\mathcal{H}_n = \{x \mapsto \prod_{k=1}^d f_k(x_k) : f_k \in \mathrm{Pol}(J_n, -2B, 2B), k = 1, \dots d\}$ and $\mathcal{A}_n = \{x \mapsto \prod_{k=1}^d f_k(x_k) : f_k \in \mathrm{Pol}(J_n, 0, M_u)\}$ for $J_n \to \infty$. We can also define the sieve without truncation as $\tilde{\Theta}_n = \tilde{\mathcal{H}}_n \times \tilde{\mathcal{A}}_n$, where $\tilde{\mathcal{H}}_n = \{x \mapsto \prod_{k=1}^d f_k(x_k) : f_k \in \mathrm{Pol}(J_n)\}$ for $J_n \to \infty$ and $\tilde{\mathcal{A}}_n = \{x \mapsto \prod_{k=1}^d f_k(x_k) : f_k \in \mathrm{Pol}(J_n)\}$ for $J_n \to \infty$.

Example 3 (Univariate Splines)**.** Let J_n be a positive number, and let $t_0, t_1, \dots t_{J_n}, t_{J_n+1}$ be real numbers with $0 = t_0 < t_1 < \dots t_{J_n} < t_{J_n+1} = 1$. Partition $[0,1]$ into $J_n + 1$ subintervals $I_j = [t_j, t_{j+1})$, $j = 0, \dots J_n - 1$ and $I_{J_n} = [t_{J_n}, t_{J_n+1}]$. We assume that the knots $t_1, t_2 \dots t_{J_n}$ have bounded mesh ratio:

$$\frac{\max_{0 \leq j \leq J_N} (t_{j+1} - t_j)}{\min_{0 \leq j \leq J_n} (t_{j+1} - t_j)} \leq c \text{ for some constant } c > 0.$$

Let $r \geq 1$ be an integer. A spline of order r with knots $t_1 \dots t_{J_n}$ is given by

$$\mathrm{Spl}(r, J_n) = \Big\{ \sum_{k=0}^{r-1} a_k x^k + \sum_{j=1}^{J_n} b_j [\max\{x - t_j, 0\}]^{r-1}, x \in [0,1] : a_k, b_j \in \mathbb{R} \}.$$

Let $\mathrm{Spl}(r, J_n, a, b)$ be the space of splines that are bounded between a and b; that is

$$\mathrm{Spl}(r, J_n, a, b) = \Big\{ x \mapsto \min(\max(f(x), a), b), x \in [0,1] : f \in \mathrm{Spl}(r, J_n) \Big\}.$$

Then, we define the sieve *with truncation* as $\Theta_n = \mathcal{H}_n \times \mathcal{A}_n$, where $\mathcal{H}_n = \{x \mapsto \prod_{k=1}^d f_k(x_k) : f_k \in \mathrm{Spl}(r, J_n, -2B, 2B), k = 1, \dots d\}$ and $\mathcal{A}_n = \{x \mapsto \prod_{k=1}^d f_k(x_k) : f_k \in \mathrm{Spl}(r, J_n, 0, M_u)\}$ for $J_n \to \infty$. We can also define the sieve without truncation as $\tilde{\Theta}_n = \tilde{\mathcal{H}}_n \times \tilde{\mathcal{A}}_n$, where $\tilde{\mathcal{H}}_n = \{x \mapsto \prod_{k=1}^d f_k(x_k) : f_k \in \mathrm{Spl}(r, J_n)\}$ for $J_n \to \infty$ and $\tilde{\mathcal{A}}_n = \{x \mapsto \prod_{k=1}^d f_k(x_k) : f_k \in \mathrm{Spl}(r, J_n)\}$ for $J_n \to \infty$.

We prove results that demonstrate the consistency of the sieve estimation procedure. Let

$$\theta_m^* = \underset{\theta \in \Theta_m}{\operatorname{argmin}} \, \mathbb{E}_P \left[L_{\mathrm{RU}}^{\Gamma}(\theta(X), Y) \right]. \tag{36}$$

First, we show that θ_m^* is the unique minimizer of the population RU risk over the sieve space Θ_m. Then, we prove that the sieve approximation error, the bias that results from minimizing the population RU risk over a finite-dimensional sieve space, converges to zero as the dimension of the sieve spaces goes to infinity. Then, we consider $\hat{\theta}_{m,n}$, the minimizer of the empirical risk over Θ_m, i.e.

$$\hat{\theta}_{m,n} = \underset{\theta \in \Theta_m}{\operatorname{argmin}} \, \widehat{\mathbb{E}}_P \left[L_{\mathrm{RU}}^{\Gamma}(\theta(X), Y) \right]$$

for a sufficiently large integer m. We can show that the estimation error, the error that results from estimating the minimizer of the empirical risk (in finite samples) in a fixed sieve space, converges to zero in probability.

Lemma 12. *Under Assumptions 1, 2, 3, $\mathbb{E}_P \left[L_{RU}^{\Gamma}(\theta(X), Y) \right]$ has a unique minimizer over Θ_m called θ_m^*. Proof in Appendix C.11*

Theorem 13. *Under Assumptions 1, 2, 3, as $m \to \infty$,*

$$||\theta_m^* - \theta^*||_{L^2(P_X, \mathcal{X})} \to 0.$$

Proof in Appendix C.12

Lemma 14. *Under Assumptions 1, 2, 3, $\hat{\theta}_{m,n}$ exists with probability approaching 1 and*

$$\hat{\theta}_{m,n} \xrightarrow{p} \theta_m^*$$

as $n \to \infty$ and m sufficiently large. Proof in Appendix C.13.

Combining Theorem 13 and Lemma 14 implies the consistency of the sieve estimation procedure: as $m, n \to \infty$,

$$||\hat{\theta}_{m,n} - \theta^*||_{L^2(P_X, \mathcal{X})} \leq ||\hat{\theta}_{m,n} - \theta_{m,n}^*||_{L^2(P_X, \mathcal{X})} + ||\theta_{m,n}^* - \theta^*||_{L^2(P_X, \mathcal{X})} \xrightarrow{p} 0.$$

To obtain a rate of convergence, we consider the classes of sufficiently smooth functions. Given a d-tuple $\beta = (\beta_1, \ldots \beta_d)$ of nonnegative integers, set $[\beta] = \beta_1 + \beta_2 + \cdots + \beta_d$ and let D^β denote the differential operator defined by $D^\beta = \frac{\partial^{[\beta]}}{\partial x_1^{\beta_1} \ldots \partial x_d^{\beta_d}}$. A real-valued function h on $\mathcal{X}$ is p-smooth if it is m times continuously differentiable on $\mathcal{X}$ and $D^\beta h$ satisfies a Hölder condition (Definition 4) with exponent γ for all d-tuples β of nonnegative integers with $[\beta] = m$. Denote the Hölder class, or the class of all p-smooth real-valued functions on $\mathcal{X}$, by $\Lambda^p(\mathcal{X})$, and the space of all m-times differentiable real-valued functions on $\mathcal{X}$ by $C^m(\mathcal{X})$. Define a Hölder ball with smoothness $p = m + \gamma$ as

$$\Lambda_c^p(\mathcal{X}) = \left\{ h \in C^m(\mathcal{X}) : \sup_{[\beta] \leq m} \sup_{x \in \mathcal{X}} |D^\beta h(x)| \leq c, \sup_{[\beta]=m} \sup_{\substack{x,y \in \mathcal{X}, \\ x \neq y}} \frac{|D^\beta h(x) - D^\beta h(y)|}{|x - y|_2^\gamma} \leq c \right\}.$$

To ensure that h, α are bounded, we define the truncated function class

$$\Lambda_c^p(\mathcal{X}, a, b) := \{ x \mapsto \min(\max(f(x), a), b), f \in \Lambda_c(\mathcal{X}) \}.$$

To obtain a rate of convergence for the estimators, we impose the following assumption on the true optimizer.

Assumption 5. Assume that $\theta^* \in \Lambda_c^p(\mathcal{X}, -2B, 2B) \times \Lambda_c^p(\mathcal{X}, 0, M_u)$ for some $c > 0$. We redefine $\Theta := \Lambda_c^p(\mathcal{X}, -2B, 2B) \times \Lambda_c^p(\mathcal{X}, 0, M_u)$.

We also required that the second moment of Y, where Y is distributed following $P_{Y|X=x}$, is bounded for all $x \in \mathcal{X}$.

Assumption 6. We assume that $\sup_{x \in \mathcal{X}} \mathbb{E}_{P_{Y|X}} \left[Y^2 \mid X = x \right] < \infty$.

In addition, we require the following condition on the density of P_X.

Assumption 7. P_X has a density that is bounded away from 0 and ∞, i.e. $0 < \inf_{x\in\mathcal{X}} p_X(x) < \sup_{x\in\mathcal{X}} p_X(x) < \infty$ for all $x \in \mathcal{X}$.

Under this last assumption, $||\cdot||_{L^2(P_X,\mathcal{X})} \asymp ||\cdot||_{L^2(\lambda,\mathcal{X})}$, where λ is the Lebesgue measure. Finally, with these assumptions, we can apply a result from Chen [2007] to show the following rate of convergence. The proof of the result requires balancing the sieve approximation error and estimation error. To get a handle on the sieve approximation error, we use the result from Timan [2014] that for the sieves $\tilde{\Theta}_{J_n}$ in Example 2 and 3 and $\theta^* \in \Lambda_c^p(\mathcal{X}) \times \Lambda_c^p(\mathcal{X})$ for $\mathcal{X}$ compact,

$$\inf_{\theta\in\tilde{\Theta}_{J_n}} ||\theta - \theta^*||_\infty = O(J_n^{-p}).$$

Theorem 15. *Let* $J_n = (\frac{n}{\log n})^{\frac{1}{2p+d}}$. *Under Assumptions 1, 2, 3, 4, 5, 6, 7,*

$$||\hat{\theta}_n - \theta^*||_{L^2(P_X,\mathcal{X})} = O_P\Big(\Big(\frac{\log n}{n}\Big)^{\frac{p}{2p+d}}\Big).$$

Furthermore, for some $Q_X \ll P_X$, *if the density ratio* $\sup_{x\in\mathcal{X}} \frac{dQ_X(x)}{dP_X(x)} < \infty$, *then*

$$||\hat{\theta}_n - \theta^*||_{L^2(Q_X,\mathcal{X})} = O_P\Big(\Big(\frac{\log n}{n}\Big)^{\frac{p}{2p+d}}\Big),$$

as well. Proof in Appendix C.14.

The minimax-optimal rate of convergence for nonparametric regression over the class of p-smooth functions is $O_P(n^{-\frac{p}{2p+d}})$ [Stone, 1982]; and so Theorem 15 demonstrates that up to log factors, RU regression achieves the rate of convergence we would expect for nonparametric regression. In other words, we find that minimax learning under Γ-biased sampling changes our learning objective; but doesn't meaningfully change the rate of convergence at which we can achieve good performance via empirical minimization.

4 Experiments

We evaluate the empirical performance of RU Regression when neural networks are used to learn h, α. First, we demonstrate that RU Regression enables us to learn models that are robust to Γ-biased sampling in simulation experiments with synthetic data. Second, we apply RU Regression in a semi-synthetic experiment with patient length-of-stay data from the MIMIC-III dataset Johnson et al. [2016a]. The code for our experiments is available in `https://github.com/roshni714/ru_regression`.

4.1 Deep Learning Implementation

Following best practices in applied machine learning, we implement our baselines and proposed method using neural networks [Goodfellow et al., 2016]. From a statistical perspective, neural networks can be seen as a practical alternative to sieve methods that automate the selection of relevant basis functions [Chen and White, 1999, Farrell et al., 2021, Schmidt-Hieber, 2020]. The benefits of neural networks include that they can be used as a black-box primitive for flexible function classes, they are straightforward to train using standard deep learning libraries, and they require less manual hyperparameter tuning than classical sieve-based approaches.

A neural network can be thought of as a function $f_\theta : \mathcal{X} \to \mathbb{R}$, where θ denotes the parameters of the network. The output space of the neural network is often the space of outcomes $\mathcal{Y}$ but can also take other values. We use Pytorch, a standard deep learning library, to instantiate, train, validate, and test the neural networks [Paszke et al., 2019]. Using the Pytorch library, it is straightforward to compute the training loss, update the parameters of the network during training using a variant of stochastic gradient descent called Adam optimization [Kingma and Ba, 2014], measure the validation loss the network incurs during training, and save the network parameters that yield the lowest validation loss for later evaluation at test-time.

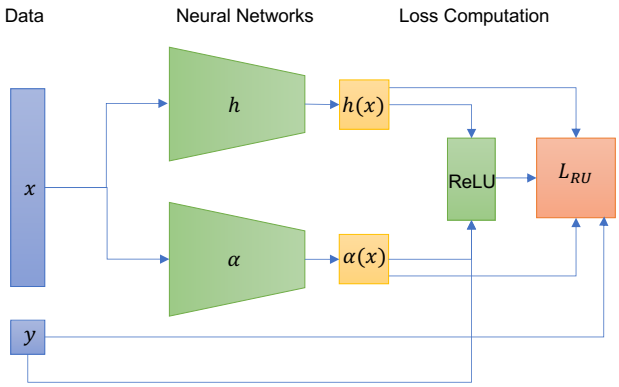

Figure 1: Model architecture for RU Regression.

Our proposed method, Rockafellar-Uryasev regression, is implemented using two neural networks. One of the networks represents the decision rule $h : \mathcal{X} \to \mathcal{Y}$, while the other network represents the auxiliary function $\alpha : \mathcal{X} \to \mathbb{R}$. A visualization for the model architecture is provided in Figure 1. Recall from Theorem 2 that RU Regression is a joint optimization over both (h, α). Mirroring this in our implementation, we propose to learn the parameters of the networks h and α simultaneously. To do so, the covariates X from a training sample (X, Y) are passed to both networks h and α, and the outputs of both networks $h(X), \alpha(X)$ are obtained. Next, we compute $L^{\Gamma}_{\text{RU}}(h(X), \alpha(X), Y)$ by summing the three terms of the RU loss (9). Recall that the third term of the RU loss depends on $(L(h(X), Y) - \alpha(X))_+$. This term can be represented as $\text{ReLU}(L(h(X), Y) - \alpha(X))$, where the ReLU (rectified linear unit) function is a commonly used neural network "activation" or transform available in Pytorch. After computing the loss, we can compute the gradient of the RU loss with respect to the parameters of network h and the parameters of network α and update the parameters of both networks using the Adam optimizer.

4.2 Simulations with Synthetic Data

We perform two simulations with synthetic data. We first consider a one-dimensional toy example because it permits visualization of the data distributions and the learned models. Next, we show that similar trends hold in a high-dimensional simulation. Implementation details for these experiments can be found in Appendix A.

4.2.1 Methods

We compare the performance of two baselines and our proposed method.

1. Standard ERM - We fit a neural network model with the squared loss function

$$L(z, y) = (y - z)^2 \tag{37}$$

   on the training data.

2. Oracle ERM - We fit a neural network model with the squared loss function (37) on data sampled from the test distribution.

3. Rockafellar-Uryasev Regression (RU Regression) - We fit two neural networks with the RU loss on the training data.

The two baselines, Standard ERM and Oracle ERM, are each implemented using a single neural network, which represents the decision rule $h : \mathcal{X} \to \mathcal{Y}$. Both methods are trained using the squared loss function. Standard ERM and RU Regression are trained on samples from the training distribution (which may differ from the test distribution), while Oracle ERM is given access to data sampled from the test distribution at train-time. Additional implementation details are in Appendix A.

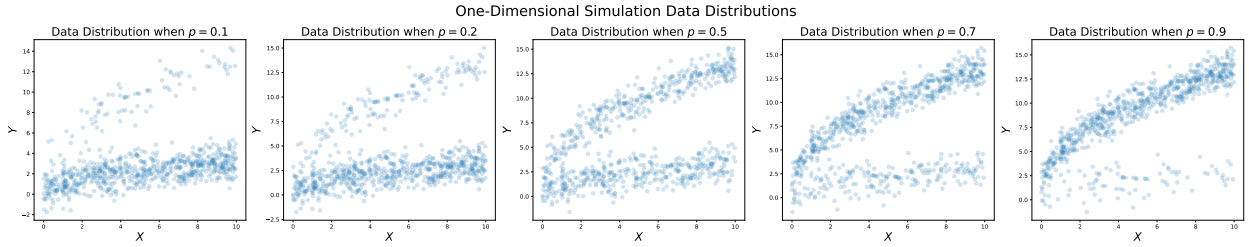

Figure 2: From left to right, we visualize the distribution over (X_i, Y_i) as p varies in $\{0.1, 0.2, 0.5, 0.7, 0.9\}$. We note that as p increases the proportion of samples where $U_i = 1$ increases.

4.2.2 One-Dimensional Toy Example

Data Generation. We generate a synthetic dataset of samples of the form (X_i, Y_i, U_i), where $X_i \in \mathbb{R}$ represents observed covariates, $Y_i \in \mathbb{R}$ represents the outcome, and $U_i \in \{0, 1\}$ represents an unobserved variable that influences the outcome Y_i. We suppose the data is distributed as follows

$$X_i \sim \text{Uniform}[0, 10], \quad U_i \sim \text{Bernoulli}(p), \quad Y_i | X_i \sim N(\sqrt{X}_i + U_i(3\sqrt{X}_i + 1), 1). \tag{38}$$

The outcomes Y_i can be clustered into two bands corresponding to $U_i = 1$ and $U_i = 0$, respectively. In this simulation, we consider a biased training distribution where $p = 0.2$, so we are less likely to observe examples with $U_i = 1$. Meanwhile, the possible test distributions are generated by varying p, e.g. $p \in \{0.1, 0.2, 0.5, 0.7, 0.9\}$. These data distributions are visualized in Figure 2. For all methods, the train, validation, and test sets consists of 7000, 1400, and 10000 samples, respectively.

Results. When the test distributions have $p \in \{0.1, 0.2\}$, the training distribution, which has $p = 0.2$, is generated with a low amount of sampling bias. In Table 1, we observe that Standard ERM achieves low test MSE in these cases.

However, for test distributions with $p \in \{0.5, 0.7, 0.9\}$, the training distribution is a more biased sample of the test distribution. In these cases, we observe that Standard ERM yields high test MSE. The RU Regression methods achieve higher test MSE on the original training distribution than Standard ERM but are more robust than Standard ERM in the presence of sampling bias. Note that Oracle ERM outperforms both the Standard ERM and RU Regression methods; this is expected because the Oracle ERM model is trained on data from the same distribution as the test distribution.

In addition, we visualize the regression functions learned from each of the methods. From the left plot of Figure 3, it is clear that the regression model learned via Standard ERM incurs high error on samples with $U_i = 1$ and low error on samples with $U_i = 0$, which explains why the method performs poorly on distributions with higher p (higher proportion of samples with $U_i = 1$). Furthermore, we observe that increasing Γ yields regression functions that incur lower error on samples with $U_i = 1$, relative to the Standard ERM model. The Oracle ERM model visualized in Figure 3 is the model that is trained on data generated when $p = 0.5$. We see that this model makes similar predictions as the RU Regression models, which explains why the RU Regression models perform similarly to the Oracle ERM model on the $p = 0.5$ test distribution.

Furthermore, we verify that the solution learned by the neural network is consistent with Theorem 9, which states that

$$\alpha^*_\Gamma(x) = q^L_{\eta(\Gamma)}(x; h^*_\Gamma(x)) \quad \forall x \in \mathcal{X}.$$

For each RU Regression method, we plot the function $\hat{\alpha}_\Gamma(x)$ learned by the neural network. In addition, with access to the data generating process, we can explictly compute the function $q^L_{\eta(\Gamma)}(X; \hat{h}_\Gamma(X))$. In the right plot of Figure 3, we observe that $\hat{\alpha}_\Gamma$ closely matches $q^L_{\eta(\Gamma)}(X; \hat{h}_\Gamma(X))$ across the possible values of X.

4.2.3 High-Dimensional Experiment

Data Generation. We generate a synthetic dataset of samples of the form (X_i, Y_i, U_i), where $X_i \in \mathbb{R}^d$ represents observed covariates, $Y_i \in \mathbb{R}$ represents the outcome, and $U_i \in \{0, 1\}$ represents an unobserved

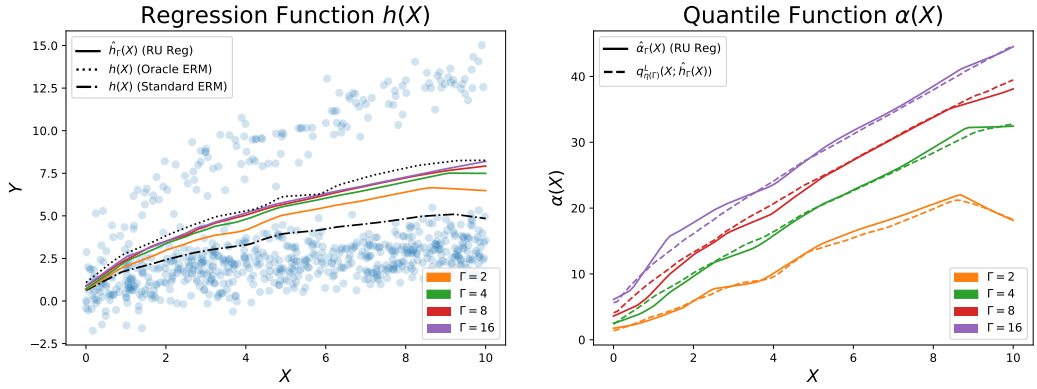

Figure 3: **Left**: We visualize the decision rules $\hat{h}$ that are learned in our one-dimensional toy example. Standard ERM incurs low error on samples with $U_i = 0$ but high error on samples with $U_i = 1$. Models learned via RU Regression incur lower error on samples with $U_i = 1$. The Oracle ERM model visualized here is the model that is trained on the data distribution when $p = 0.5$. **Right**: We visualize the auxiliary function $\hat{\alpha}_\Gamma$ that is learned in the RU Regression methods for our one-dimensional toy example. We realize that the learned $\hat{\alpha}_\Gamma$ closely tracks $q^L_{\eta(\Gamma)}(x; \hat{h}_\Gamma(x))$ as expected.

Method	Test MSE				
	$p = 0.1$	$p = 0.2$	$p = 0.5$	$p = 0.7$	$p = 0.9$
Standard ERM	6.939 ± 0.174	10.480 ± 0.126	20.866 ± 0.304	27.880 ± 0.484	34.913 ± 0.668
RU Regression ($\Gamma = 2$)	10.074 ± 0.247	11.846 ± 0.138	17.029 ± 0.236	20.522 ± 0.308	24.046 ± 0.540
RU Regression ($\Gamma = 4$)	12.456 ± 0.431	13.388 ± 0.309	16.093 ± 0.179	17.898 ± 0.300	19.750 ± 0.584
RU Regression ($\Gamma = 8$)	13.419 ± 0.306	14.057 ± 0.255	15.895 ± 0.133	17.119 ± 0.143	18.388 ± 0.304
RU Regression ($\Gamma = 16$)	13.613 ± 0.400	14.197 ± 0.308	15.873 ± 0.140	16.983 ± 0.262	18.142 ± 0.480
Oracle ERM	6.306 ± 0.187	10.480 ± 0.126	15.743 ± 0.152	13.341 ± 0.123	6.274 ± 0.176

Table 1: Results from the one-dimensional simulation experiment. We report the mean and standard deviation of the test MSE from 6 random trials, where the randomness is over the dataset generation. Standard ERM incurs high test MSE for high values of p. RU Regression is more robust to sampling bias than Standard ERM. RU Regression matches the performance of Oracle ERM at $p = 0.5$.

variable that influences the outcome Y_i. Since we aim to consider a high-dimensional example, we set $d = 16$. We suppose the data is distributed as follows

$$X_i \sim \text{Uniform}[0,1]^d, \quad U_i \sim \text{Bernoulli}(p), \quad Y_i | X_i \sim N(\mathbf{a}^T X_i + 0.5 \cdot U_i, 0.1), \tag{39}$$

where $\mathbf{a} \in \mathbb{R}^d$ is a constant vector. Similar to the one-dimensional example, the outcomes Y_i can be clustered into two hyperplanes $Y_i = \mathbf{a}^T X_i + 0.5$ for samples with $U_i = 1$ and $Y = \mathbf{a}^T X_i$ for samples with $U_i = 0$. As in the one-dimensional example, we consider distribution shifts which result from varying p, the probability that $U_i = 1$. We consider a biased training distribution where $p = 0.2$, so examples with $U_i = 1$ occur with lower frequency than examples with $U_i = 0$. At test-time, we evaluate the learned models on data distributions where $p \in \{0.1, 0.2, 0.5, 0.7, 0.9\}$. For all methods, the train, validation, and test sets consists of 100000, 20000, and 20000 samples, respectively.

Results. The results from the high-dimensional simulation are consistent with those from the one-dimensional simulation. From Table 2, Standard ERM achieves low test MSE when the amount of sampling bias is low, meaning that the test distribution has $p \in \{0.1, 0.2\}$. However, the test MSE of Standard ERM increases when the amount of sampling bias is high, when $p \in \{0.5, 0.7, 0.9\}$. The RU Regression methods achieve higher test MSE on the original training distribution than Standard ERM but are more robust than Standard ERM under sampling bias. We note that RU Regression matches the performance of the Oracle ERM model when $p = 0.5$.

Method	Test MSE				
	$p = 0.1$	$p = 0.2$	$p = 0.5$	$p = 0.7$	$p = 0.9$
Standard ERM	0.028 ± 0.000	0.043 ± 0.000	0.088 ± 0.002	0.118 ± 0.002	0.148 ± 0.003
RU Regression ($\Gamma = 2$)	0.041 ± 0.002	0.049 ± 0.001	0.071 ± 0.001	0.086 ± 0.003	0.100 ± 0.004
RU Regression ($\Gamma = 4$)	0.054 ± 0.008	0.057 ± 0.006	0.067 ± 0.002	0.073 ± 0.006	0.080 ± 0.011
RU Regression ($\Gamma = 8$)	0.056 ± 0.003	0.058 ± 0.002	0.066 ± 0.001	0.071 ± 0.002	0.076 ± 0.004
RU Regression ($\Gamma = 16$)	0.057 ± 0.003	0.059 ± 0.002	0.066 ± 0.000	0.070 ± 0.002	0.074 ± 0.003
Oracle ERM	0.025 ± 0.000	0.043 ± 0.000	0.066 ± 0.000	0.056 ± 0.000	0.025 ± 0.000

Table 2: Results from the high-dimensional ($d = 16$) simulation experiment. We report the mean and standard deviation of the test MSE from 6 random trials, where the randomness is over the dataset generation. Standard ERM incurs high test MSE for high values of p, where the amount of sampling bias is high. RU Regression is more robust to sampling bias than Standard ERM. RU Regression matches the performance of Oracle ERM at $p = 0.5$.

4.3 MIMIC-III Data

Accurate patient length-of-stay predictions are useful for scheduling and hospital resource management [Harutyunyan et al., 2019]. Many recent works study the problem of predicting patient length-of-stay from patient covariates [Daghistani et al., 2019, Morton et al., 2014, Sotoodeh and Ho, 2019]. In this experiment, we evaluate our approach on electronic health record data drawn from the publicly available MIMIC-III dataset [Johnson et al., 2016a]. We study the robustness of regression models when the distribution of patients observed at test time differs from the distribution of patients observed at train-time.

Data. In this experiment, the observed covariates X_i consist of 17 different medical measurements of a patient recorded within the first 24 hours of hospital stay (see Appendix A.3 for details on the particular covariates). The outcome Y_i is the patient length-of-stay in the ICU in days. We split the original dataset into train, validation, and test sets consisting of 7045, 4697, and 7829 samples, respectively.

To simulate a setting where the data we observed is a biased draw from our true target distribution, we imagine that datapoints from the true underlying distribution are observed with probability $\pi_i = 1/Y_i$ (i.e., using notation from Definition 1, we have $\mathbb{P}\left[S_i \,\middle|\, Y_i = y\right] = 1/y$). Under this assumption, we can use the test set to get unbiased estimates for the loss of a rule learned on the training set given new draws from either the training set, or the true target distribution:

$$\begin{aligned} \text{Training Environment MSE} &= \sum_{i=1}^{n_{\text{test}}} L(h(X_i), Y_i) \,/\, n_{\text{test}} \\ \text{Target Environment MSE} &= \sum_{i=1}^{n_{\text{test}}} \pi_i^{-1} L(h(X_i), Y_i) \,/\, \sum_{i=1}^{n_{\text{test}}} \pi_i^{-1}. \end{aligned} \tag{40}$$

We note that these π_i remain unobserved, and are not used for any algorithm in learning. They are simply used to define a hypothetical target environment under which we seek to perform well despite biased sampling (with unknown sampling bias).

We compare the following methods.

1. Standard ERM - We fit a neural network model with the squared loss function (Equation 37) on the training data.

2. Rockafellar-Uryasev Regression (RU Regression) - We fit two neural networks with the L_{RU}^{Γ} loss function, where one network learns h and the other network learns α, on the training data. The model architecture is visualized in Figure 1.

Results. As seen in Table 3, RU Regression trades performance on the training distribution for robustness to sampling bias. RU Regression performs worse than standard ERM in the training environment but is more accurate than the Standard ERM in the the target environment, where patients with high length-of-stay occur with higher frequency than in the training environement. Thus, at a modest cost in shift-free

Method	Weighted Test MSE	
	Training Environment	Target Environment
Standard ERM	3.230 ± 0.079	6.926 ± 0.265
RU Regression ($\Gamma = 1.50$)	3.227 ± 0.074	6.663 ± 0.253
RU Regression ($\Gamma = 2.00$)	3.274 ± 0.072	6.607 ± 0.247
RU Regression ($\Gamma = 2.50$)	3.349 ± 0.070	6.313 ± 0.237
RU Regression ($\Gamma = 3.00$)	3.441 ± 0.068	6.060 ± 0.224

Table 3: Results from MIMIC-III Experiment. We report the weighted test MSE and bootstrap standard error with 5000 bootstrap samples.

accuracy, our method achieved considerable improvements in the presence of sampling bias. We emphasize that RU Regression was not given any information on how the test set might differ from the training set; we simply posited that the shift is some re-weighting of the type (8) and asked RU Regression to be robust to any such shift (up to a factor $\Gamma = 3$). We report the bootstrap standard error obtained with 5000 bootstrap samples.

5 Discussion

In this paper, we considered a model for sampling bias, Γ-biased sampling, and proposed an approach to learning minimax decision rules under Γ-biased sampling. Under our model, selection bias may depend on unobservables—and the analyst may not be able to model sampling bias. As such, the optimal decision rule under the target distribution is not identified; and the best the analyst can do is to seek a decision rule with minimax guarantees under all target distributions that may have generated the observed data under Γ-biased sampling. One of our key results is that, although our learning problem may at first appear intractable, we can in fact turn it into a convex problem over an augmented function space by leveraging a result of Rockafellar and Uryasev [2000].

One question we have not focused on in this paper is how to choose Γ in practice, i.e., how to set the maximal bias parameter in Definition 1. We emphasize that Γ is not something that's identified from the data; rather, it's a parameter that the decision maker must choose when designing their learning algorithm. Setting $\Gamma = 1$ corresponds to the usual empirical risk minimization algorithm, with no robustness guarantees under potential sampling bias. Using a larger value $\Gamma > 1$ enables the analyst to gain robustness to sampling bias at the cost of potentially worsening performance in the training environment.

One practical way to navigate the choice of Γ is, following Imbens [2003], to consider values of Γ that help make decision rules robust across different available samples. For example, if one seeks to design a generally applicable risk prediction model using data only from two hospitals A and B whose patients come from different populations, one could examine which values of Γ enable one to use data from hospital A that work well in hospital B, and vice-versa. While such an exercise does not tell us which value would be best for accuracy on the (unknown) target distribution, it can at least shed light on the order of magnitude of values for Γ that are likely to be helpful in practice.

Finally, we note that it is interesting to consider how our results relate to the broader literature on "robust" learning. There is a broad literature on methods for learning that are robust to data contamination. For example, there has been interest in models where a fraction ε of the data comes from a different distribution [Chen et al., 2016, Huber, 1964], or was chosen by an adversary [Charikar et al., 2017, Diakonikolas et al., 2019, Lugosi and Mendelson, 2021]. Interestingly, however, methods that seek robustness to data corruption effectively down-weight the influence of outliers, because otherwise a small fraction of corrupted examples could affect results arbitrarily much. In contrast, in our setting, we tend to give larger weight to samples with large loss—because under biased sampling a small number of samples with large loss in the training distribution could reflect a much larger fraction of the true target. In other words, approaches that seek robustness to data corruption end up to a large extent doing the opposite of what we do here in order to achieve robustness to sampling bias. This tension suggests that a learning algorithm cannot simply be "robust". One can make choices that make an algorithm robust to some possible problems with the

training distribution (e.g., sampling bias, or data corruption), but these choices will involve trade-offs that may reduce robustness across other dimensions.

References

Isaiah Andrews and Emily Oster. A simple approximation for evaluating external validity bias. *Economics Letters*, 178:58–62, 2019.

Peter M Aronow and Donald KK Lee. Interval estimation of population means under unknown but bounded probabilities of sample selection. *Biometrika*, 100(1):235–240, 2013.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

Orazio Attanasio, Adriana Kugler, and Costas Meghir. Subsidizing vocational training for disadvantaged youth in colombia: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, 3(3):188–220, 2011.

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

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66 (3):1025–1044, 2020.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60, 2017.

Mengjie Chen, Chao Gao, and Zhao Ren. A general decision theory for Huber's *epsilon*-contamination model. *Electronic Journal of Statistics*, 10(2):3752–3774, 2016.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632, 2007.

Xiaohong Chen and Xiaotong Shen. Sieve extremum estimates for weakly dependent data. *Econometrica*, pages 289–314, 1998.

Xiaohong Chen and Halbert White. Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691, 1999.

Tahani A Daghistani, Radwa Elshawi, Sherif Sakr, Amjad M Ahmed, Abdullah Al-Thwayee, and Mouaz H Al-Mallah. Predictors of in-hospital length of stay among cardiac patients: a machine learning approach. *International journal of cardiology*, 288:140–147, 2019.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48 (2):742–864, 2019.

Jacob Dorn, Kevin Guo, and Nathan Kallus. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *arXiv preprint arXiv:2112.11449*, 2021.

John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses for latent covariate mixtures. *arXiv preprint arXiv:2007.13982*, 2020.

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

Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022.

Adrián Esteban-Pérez and Juan M Morales. Distributionally robust stochastic programs with side information based on trimmings. *Mathematical Programming*, pages 1–37, 2021.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

Stuart Geman and Chii-Ruey Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *The annals of Statistics*, pages 401–414, 1982.

David C Goff, Jr, Donald M Lloyd-Jones, Glen Bennett, Sean Coady, Ralph B D'agostino, Raymond Gibbons, Philip Greenland, Daniel T Lackland, Daniel Levy, Christopher J O'donnell, et al. 2013 acc/aha guideline on the assessment of cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines. *Circulation*, 129(25_suppl_2):S49–S73, 2014.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101 (23):e215–e220, 2000.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Muhammet Gul and Erkan Celik. An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments. *Health Systems*, 9(4):263–284, 2020.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1): 73–101, 1964.

Guido W Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, 2003.

Ying Jin, Zhimei Ren, and Zhengyuan Zhou. Sensitivity analysis under the f-sensitivity models: Definition, estimation and inference. *arXiv preprint arXiv:2203.04373*, 2022.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016a.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016b.

Nathan Kallus and Angela Zhou. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

Gabor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410, 2021.

Charles F Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246, 2004.

Paul Michel, Tatsunori Hashimoto, and Graham Neubig. Distributionally robust models with parametric likelihood ratios. *arXiv preprint arXiv:2204.06340*, 2022.

Luke W Miratrix, Stefan Wager, and Jose R Zubizarreta. Shape-constrained partial identification of a population mean under unknown probabilities of sample selection. *Biometrika*, 105(1):103–114, 2018.

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

April Morton, Eman Marzban, Georgios Giannoulis, Ayush Patel, Rajender Aparasu, and Ioannis A Kakadiaris. A comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients. In *2014 13th International Conference on Machine Learning and Applications*, pages 428–431. IEEE, 2014.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108 (2):299–319, 2021.

Xinkun Nie, Guido Imbens, and Stefan Wager. Covariate balancing sensitivity analysis for extrapolating randomized trials across locations. *arXiv preprint arXiv:2112.04723*, 2021.

Michael Oberst, Nikolaj Thams, Jonas Peters, and David Sontag. Regularizing towards causal invariance: Linear models with proxies. In *International Conference on Machine Learning*, pages 8260–8270. PMLR, 2021.

Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

R Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2: 21–42, 2000.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Mani Sotoodeh and Joyce C Ho. Improving length of stay prediction using a hidden markov model. *AMIA Summits on Translational Science Proceedings*, 2019:425, 2019.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

Jörg Stoye. Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1):70–81, 2009.

Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):369–386, 2011.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755, 2015.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

Nikolaj Thams, Michael Oberst, and David Sontag. Evaluating robustness to dataset shift via parametric robustness sets. *arXiv preprint arXiv:2205.15947*, 2022.

Aleksandr Filippovich Timan. *Theory of approximation of functions of a real variable*. Elsevier, 2014.

Elizabeth Tipton. Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, 38(3):239–266, 2013.

Elizabeth Tipton. How generalizable is your experiment? an index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, 39(6):478–501, 2014.

Sara A Van de Geer and Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Vladimir N Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8.

Sheng-Min Wang, Changsu Han, Soo-Jung Lee, Tae-Youn Jun, Ashwin A Patkar, Prakash S Masand, and Chi-Un Pae. Efficacy of antidepressants: bias in randomized clinical trials and related issues. *Expert Review of Clinical Pharmacology*, 11(1):15–25, 2018.

Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM conference on health, inference, and learning*, pages 222–235, 2020.

Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*, 2018.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, (forthcoming), 2022.

A Experiment Details

A.1 One-Dimensional Toy Example

A.1.1 Models

For the Standard ERM and Oracle ERM models, we train a neural network with 2 hidden layers and 128 units per layer and ReLU activation to learn the regression function h. For the RU Regression model, we jointly train two neural networks to learn the regression function h and the quantile function α, respectively. A visualization of the model architecture for RU Regression is provided in Figure 1. Each of the neural networks has 2 hidden layers and 64 units per layer and ReLU activation. We note that overall the Standard ERM and Oracle ERM models have 18.8K trainable parameters, and the RU Regression model has 10.6K trainable parameters.

A.1.2 Dataset Splits

For all methods, the train, validation, and test sets consists of 7000, 1400, and 10000 samples, respectively. For Standard ERM and RU Regression, the train and validation sets are generated via the data model specified in Equation 38 with $p = 0.2$. For Oracle ERM, the train and validation set is generated with the same data model with the parameter p matching that of the test distribution. All methods are evaluated on the same test sets, which are generated via the data model in Equation 38 with parameter p taking value in $[0.1, 0.2, 0.5, 0.7, 0.9]$. For each of 6 random seeds [0, 1, 2, 3, 4, 5], a new dataset (Standard ERM/RU Regression train and validation sets, Oracle ERM train and validation sets, and test sets) is generated.

A.1.3 Training Procedure

The models are trained for a maximum of 100 epochs with batch size equal to 1750 and we use the Adam optimizer with learning rate 1e-2. Each epoch we check the loss obtained on the validation set and select the model that minimizes the loss on the validation set.

A.2 High-Dimensional Experiment

A.2.1 Models

We use the same models as in the one-dimensional experiment. See Section A.1.1 for details.

A.2.2 Dataset Splits

For all methods, the train, validation, and test sets consists of 100000, 20000, and 20000 samples, respectively. In the data model in Equation 39, we set

$$\begin{aligned}\mathbf{a} = [&0.098, 0.430, 0.206, 0.090, -0.153, 0.292, -0.125, 0.784,\\ &0.927, -0.233, 0.583, 0.0578, 0.136, 0.851, -0.858, -0.826]\end{aligned}$$

in all experiments. For Standard ERM and RU Regression, the train and validation sets are generated via Equation 39 with $p = 0.2$. For Oracle ERM, the train and validation set is generated with the same data model with the parameter p matching that of the test distribution. All methods are evaluated on the same test sets, which are generated via the data model in Equation 39 with parameter p taking value in $[0.1, 0.2, 0.5, 0.7, 0.9]$. For each of 6 random seeds [0, 1, 2, 3, 4, 5], a new dataset (Standard ERM/RU Regression train and validation sets, Oracle ERM train and validation sets, and test sets) is generated.

A.2.3 Training Procedure

The models are trained for a maximum of 50 epochs with batch size equal to 25000 and we use the Adam optimizer with learning rate 1e-2. Each step we check the loss obtained on the validation set and select the model that minimizes the loss on the validation set.

A.3 MIMIC-III Experiment

A.3.1 Dataset

Medical Information Mart for Intensive Care III (MIMIC-III) is a freely accessible medical database of critically ill patients admitted to the intensive care unit (ICU) at Beth Israel Deaconess Medical Center (BIDMC) from 2001 to 2012 [Johnson et al., 2016b, Goldberger et al., 2000]. During that time, BIDMC switched clinical information systems from Carevue (2001-2008) to Metavision (2008-2012). To ensure data consistency, only data archived via the Metavision system was used in the dataset.

A.3.2 Feature Selection and Data Preprocessing

We select the same patient features and imputed values as in Harutyunyan et al. [2019]. A total of 17 variables were extracted from the chartevents table to include in the dataset - capillary refill rate, blood pressure (systolic, diastolic, and mean), fraction of inspired oxygen, Glasgow Coma Score (eye opening response, motor response, verbal response, and total score), serum glucose, heart rate, respiratory rate, oxygen saturation, respiratory rate, temperature, weight, and arterial pH. For each unique ICU stay, values were extracted for the first 24 hours upon admission to the ICU and averaged. Normal values were imputed for missing variables as shown in Table 4.

Variable	MIMIC-III item ids from chartevents table	Imputed value
Capillary refill rate	(223951, 224308)	0
Diastolic blood pressure	(220051, 227242, 224643, 220180, 225310)	59.0
Systolic blood pressure	(220050, 224167, 227243, 220179, 225309)	118.0
Mean blood pressure	(220052, 220181, 225312)	77.0
Fraction inspired oxygen	(223835)	0.21
GCS eye opening	(220739)	4
GCS motor response	(223901)	6
GCS verbal response	(223900)	5
GCS total	(220739 + 223901 + 223900)	15
Glucose	(228388, 225664, 220621, 226537)	128.0
Heart Rate	(220045)	86
Height	(226707, 226730)	170.0
Oxygen saturation	(220227, 220277, 228232)	98.0
Respiratory rate	(220210, 224688, 224689, 224690)	19
Temperature	(223761, 223762)	97.88
Weight	(224639, 226512, 226531)	178.6
pH	(223830)	7.4

Table 4: Variables included in dataset

Following the cohort selection procedure in Wang et al. [2020], we further restrict to patients with covariates within physiologically valid range of measurements and length-of-stay less than or equal to 10 days.

A.3.3 Training Details

Models. For the Standard ERM model, we train a neural network with 2 hidden layers and 128 units per layer and ReLU activation to learn the regression function h. For the RU Regression model, we jointly train two neural networks to learn the regression function h and the quantile function α, respectively. Each of the neural networks has 2 hidden layers and 64 units per layer and ReLU activation. A visualization of the model architecture for RU Regression is provided in Figure 1. We note that overall the Standard ERM model has 18.8K trainable parameters, and the RU Regression model has 10.6K trainable parameters.

Dataset Splits. For all methods, the train, validation, and test sets consists of 7045, 4697, and 7829 samples, respectively.

B Standard Results

Definition 4. A function h on $\mathcal{X}$ is said to satisfy a Holder condition with exponent β if there is a positive number γ such that $|h(x) - h(x_0)| \leq \gamma |x - x_0|^\beta$ for $x_0, x \in \mathcal{X}$.

Lemma 16. *If there is a function $Q_0(\theta)$ such that (i) $Q_0(\theta)$ is uniquely maximized at θ_0; (ii) θ_0 is an element of the interior of a convex set Θ and $\hat{Q}_n(\theta)$ is concave; and (iii) $\hat{Q}_n(\theta) \to Q_0(\theta)$for all $\theta \in \Theta$, then $\hat{\theta}_n$ exists with probability approaching one and $\hat{\theta}_n \xrightarrow{p} \theta_0$ (Theorem 2.7, Newey and McFadden [1994]).*

Lemma 17. *If a functional $J : V \to \mathbb{R}$ is Gâteaux differentiable J' at $u_0 \in V$ and has a relative extremum at u_0, then $J'(u_0; v) = 0$ for all $v \in V$.*

Lemma 18. *If $\{e_i\}$ is an orthonormal basis (a maximal orthonormal sequence) in a Hilbert space H then for any element $u \in H$ the 'Fourier-Bessel series' converges to u:*

$$u = \sum_{i=1}^{\infty} \langle u, e_i \rangle e_i.$$

Lemma 19. *Let X be a Hilbert space, and suppose $f : X \to [-\infty, \infty]$ is lower semicontinuous and convex. If C is a closed, bounded, and convex subset of X, then f achieves its minimum on C; i.e., there is some $x_0 \in C$ with $f(x_0) = \inf_{x \in C} f(x)$.*

Lemma 20. *Let A be a 2×2 symmetric matrix with $\operatorname{tr}(A) > 0$ and $\det(A) \geq 0$. Then*

$$\lambda_{\min}(A) \geq \frac{\det A}{\operatorname{tr} A}, \quad \lambda_{\max}(A) \leq \operatorname{tr} A.$$

Proof in Appendix D.3.

Lemma 21. *Let $H(h, \alpha) = G(h) + F(h, \alpha)$, where G is strongly convex and Gâteaux differentiable in h and F is jointly convex in (h, α), strictly convex in α, and Gâteaux differentiable in (h, α). Then H is strictly convex in (h, α). Proof in Appendix D.4.*

C Proofs of Main Results

C.1 Notation

We introduce notation that is used in the proofs and technical lemmas.

$$L^{\Gamma}_{\mathrm{RU},1}(z, y) := \Gamma^{-1} L(z, y) \tag{41}$$
$$L^{\Gamma}_{\mathrm{RU},2}(a) := (1 - \Gamma^{-1}) a \tag{42}$$
$$L^{\Gamma}_{\mathrm{RU},3}(z, y, a) := (\Gamma - \Gamma^{-1}) \cdot (L(z, y) - a)_+. \tag{43}$$

Define

$$R_{f,c} := \{x \in \mathcal{X} \mid f(x) < c\} \tag{44}$$
$$S_{f,c} := \{x \in \mathcal{X} \mid f(x) > c\}. \tag{45}$$

When we consider loss functions L that satisfy Assumption 2, we define

$$\ell_1(y) := \begin{cases} \ell(y) & y > 0 \\ 0 & y \leq 0 \end{cases}, \quad \ell_2(y) := \begin{cases} 0 & y > 0 \\ \ell(y) & y \leq 0 \end{cases}, \tag{46}$$
$$T_{1,x}(c) := \mathbb{E}_{P_{Y|X=x}} \left[\ell(Y - c) \mid X = x \right], \tag{47}$$
$$T_{3,x}(c, d) := \begin{cases} \mathbb{E}_{P_{Y|X=x}} \left[(\ell(Y - c) - d) \mathbb{I}(\ell(Y - c) > d) \mid X = x \right] & d > 0 \\ \mathbb{E}_{P_{Y|X=x}} \left[\ell(Y - c) - d \mid X = x \right] & d \leq 0 \end{cases}. \tag{48}$$

C.2 Technical Lemmas

We prove lemmas about the transforms $T_{1,x}(c), T_{3,x}(c,d)$. These enable us to establish more general properties of the RU loss.

Lemma 22. *Under Assumption 2, $T_{1,x}(c)$ is twice-differentiable in c and*

$$\mathbb{E}_P\left[L^{\Gamma}_{RU,1}(h(X),Y)\right] = \Gamma^{-1}\mathbb{E}_{P_X}\left[T_{1,X}(h(X))\right].$$

Proof in Appendix D.5.

Lemma 23. *Under Assumption 2, 3, $T_{3,x}(c,d)$ is differentiable in c,d. In particular,*

$$T^d_{3,x}(c,d) = \begin{cases} -\Pr(\ell(Y-c) > d \mid X = x) & d > 0 \\ -1 & d \le 0 \end{cases}.$$

Equivalently,

$$T^d_{3,x}(c,d) = \begin{cases} -1 + P_{Y|X=x}(c+\ell_1^{-1}(d)) - P_{Y|X=x}(c+\ell_2^{-1}(d)) & d > 0 \\ -1 & d \le 0 \end{cases}.$$

In addition, $T_{3,x}(c,d)$ is twice-differentiable in c,d when $d > 0$. The second derivatives are

$$T^{cc}_{3,x}(c,d) = \sum_{i\in\{1,2\}} |\ell'(\ell_i^{-1}(d))| \cdot p_{Y|X=x}(c+\ell_i^{-1}(d)) + \mathbb{E}_{P_{Y|X}}\left[\ell''(Y-c)\mathbb{I}(\ell(Y-c) > d)\right],$$

$$T^{dd}_{3,x}(c,d) = \sum_{i\in\{1,2\}} \frac{p_{Y|X=x}(c+\ell_i^{-1}(d))}{|\ell'(\ell_i^{-1}(d))|},$$

$$T^{cd}_{3,x}(c,d) = p_{Y|X=x}(c+\ell_1^{-1}(d)) - p_{Y|X=x}(c+\ell_2^{-1}(d)),$$

where ℓ_1^{-1} is the inverse of $\ell(z)$ when $z > 0$ and ℓ_2^{-1} is the inverse of $\ell(z)$ when $z < 0$.

Also,

$$\mathbb{E}_P\left[L^{\Gamma}_{RU,3}(h(X),\alpha(X),Y)\right] = (\Gamma - \Gamma^{-1})\mathbb{E}_{P_X}\left[T_{3,X}(h(X),\alpha(X))\right].$$

Proof in Appendix D.6.

Lemma 24. *Under Assumption 2, 3, there are symmetric matrices $A_x(c,d), B_x(c,d)$ such that*

$$A_x(c,d) \preceq \nabla^2 T_{3,x}(c,d) \preceq B_x(c,d)$$

when $d > 0$. The entries of $A_x(c,d)$ are given by

$$A_{x,11}(c,d) = \sum_{i\in\{1,2\}} |\ell'(\ell_i^{-1}(d))| \cdot p_{Y|X=x}(c+\ell_i^{-1}(d)) + C_{L,l} \cdot \Pr(\ell(Y-c) > d \mid X = x)$$

$$A_{x,22}(c,d) = \sum_{i\in\{1,2,\}} \frac{p_{Y|X=x}(c+\ell_i^{-1}(d))}{|\ell'(\ell_i^{-1}(d))|},$$

$$A_{x,12}(c,d) = p_{Y|X=x}(c+\ell_1^{-1}(d)) - p_{Y|X=x}(c+\ell_2^{-1}(d)).$$

The entries of $B_x(c,d)$ are given by

$$B_{x,11}(c,d) = \sum_{i\in\{1,2\}} |\ell'(\ell_i^{-1}(d))| \cdot p_{Y|X=x}(c+\ell_i^{-1}(d)) + \mathbb{E}_{P_{Y|X=x}}\left[\ell''(Y-c) \mid X = x\right],$$

$$B_{x,22}(c,d) = \sum_{i\in\{1,2\}} \frac{p_{Y|X=x}(c+\ell_i^{-1}(d))}{|\ell'(\ell_i^{-1}(d))|},$$

$$B_{x,12}(c,d) = p_{Y|X=x}(c+\ell_1^{-1}(d)) - p_{Y|X=x}(c+\ell_2^{-1}(d)).$$

Proof in Appendix D.7.

We give a few additional lemmas related to the RU loss. These lemmas are used in proofs of many of the results from Section 3.1.

Lemma 25. *Under Assumption 1, 2 3,* $\mathbb{E}_P\left[L^\Gamma_{RU}(h(X),\alpha(X),Y)\right]$ *is Gâteaux differentiable in* (h,α) *on* $L^2(P_X,\mathcal{X})\times L^2(P_X,\mathcal{X})$ *and twice-Gâteaux differentiable in* (h,α) *on* $\mathcal{C}$*, where*

$$\mathcal{C}=\{(h,\alpha)\in\Theta \mid \alpha(x)>0 \quad \forall x\in\mathcal{X}\}.$$

Proof in Appendix D.8.

Lemma 26. *Under Assumption 2,* $\mathbb{E}_P\left[L^\Gamma_{RU,1}(h(X),Y)\right]$ *is strongly convex in* h*. Proof in Appendix D.9.*

Lemma 27. *Under Assumptions 1, 2, 3,* $\mathbb{E}_P\left[L^\Gamma_{RU,3}(h(X),\alpha(X),Y)\right]$ *is strictly convex in* α *on* $\mathcal{A}$*. Proof in Appendix D.10.*

C.3 Proof of Lemma 1

First, suppose that Q generates P via Γ-biased sampling. We show that (8) holds and that $\sup_{x\in\mathcal{X}}\frac{dP_X(x)}{dQ_X(x)}<C$ for some $C<\infty$.

$$\begin{aligned}
\frac{dQ_{Y|X=x}(y)}{dP_{Y|X=x}(y)} &= \frac{d\tilde{Q}_{Y|X}(y)}{d\tilde{Q}_{Y|X,S=1}(y)} && (49)\\
&= \frac{d\tilde{Q}_{X,Y}(x,y)}{d\tilde{Q}_X(x)}\cdot\frac{d\tilde{Q}_{X|S=1}(x)}{d\tilde{Q}_{X,Y|S=1}(x,y)} && (50)\\
&= \frac{d\tilde{Q}_{X,Y}(x,y)}{d\tilde{Q}_{X,Y|S=1}(x,y)}\cdot\frac{d\tilde{Q}_{X|S=1}(x)}{d\tilde{Q}_X(x)} && (51)\\
&= \frac{d\tilde{Q}_{X,Y}(x,y)}{d\tilde{Q}_{X,Y|S=1}(x,y)}\cdot\frac{d\tilde{Q}_{X|S=1}(x)}{d\tilde{Q}_X(x)}\cdot\frac{\mathbb{P}_{\tilde{Q}}\left[S=1\right]}{\mathbb{P}_{\tilde{Q}}\left[S=1\right]} && (52)\\
&= \frac{\mathbb{P}_{\tilde{Q}}\left[S=1\mid X=x\right]}{\mathbb{P}_{\tilde{Q}}\left[S=1\mid X=x,Y=y\right]} && (53)\\
&\in [\Gamma^{-1},\Gamma]. && (54)
\end{aligned}$$

(53) follows from Bayes' Rule. (54) follows from (3). So, (8) holds.

We also show that the covariate density ratio between P and Q is bounded. We note that

$$\begin{aligned}
dP_X(x) &= dQ_X(x)\cdot\frac{\mathbb{P}_{\tilde{Q}}\left[S=1\mid X=x\right]}{\mathbb{P}_{\tilde{Q}}\left[S=1\right]}\\
&\leq \frac{1}{\mathbb{P}_{\tilde{Q}}\left[S=1\right]}.
\end{aligned}$$

Thus, we have that $\sup_{x\in\mathcal{X}}\frac{dP_X(x)}{dQ_X(x)}$ is upper bounded by a constant.

Second, we show the converse. Let Q be a distribution over (X,Y) that satisfies (8). We define $\tilde{Q}$ to be a distribution over (X,Y,S), where $X\in\mathcal{X},Y\in\mathcal{Y},S\in\{0,1\}$.We set $\tilde{Q}_{X,Y}=Q$ and define that

$$\mathbb{P}_{\tilde{Q}}\left[S=1\mid X=x,Y=y\right]=\frac{1}{N}\frac{dP(x,y)}{dQ(x,y)}, \tag{55}$$

where $N\leq 1/C\Gamma$. Note that $\mathbb{P}_{\tilde{Q}}\left[S=1\mid X=x,Y=y\right]\in[0,1]$ because (8) holds and $\sup_{x\in\mathcal{X}}\frac{dP_X(x)}{dQ_X(x)}<C$.

To show that converse holds, we must verify (3) holds for $\tilde{Q}$ and that $\tilde{Q}_{X,Y|S=1}=P$. First, we verify (3). We compute $\mathbb{P}_{\tilde{Q}}\left[S=1\mid X=x\right]$.

$$\mathbb{P}_{\tilde{Q}}\left[S=1\mid X=x\right]=\mathbb{E}_{\tilde{Q}}\left[\frac{1}{N}\cdot\frac{dP(X,Y)}{dQ(X,Y)}\mid X=x\right]=\frac{1}{N}\cdot\frac{dP_X(x)}{dQ_X(x)}. \tag{56}$$

Note that

$$\mathbb{P}_{\tilde{Q}}\left[S=1 \mid X=x, Y=y\right]=\frac{1}{N}\cdot\frac{dP(x,y)}{dQ(x,y)}=\mathbb{P}_{\tilde{Q}}\left[S=1 \mid X=x\right]\cdot\frac{dP_{Y|X=x}(y)}{dQ_{Y|X=x}(y)}.$$

From (8), we have that

$$\frac{dP_{Y|X=x}(y)}{dQ_{Y|X=x}(y)}\in[\Gamma^{-1},\Gamma].$$

So, we have that (3) holds for $\tilde{Q}$.

Now, we can verify that $\tilde{Q}_{X,Y|S=1}=P$. We aim to verify that

$$d\tilde{Q}_{X,Y|S=1}(x,y)=dP(x,y). \tag{57}$$

We have that

$$\begin{aligned}
d\tilde{Q}_{X,Y,S=1}(x,y)&=\mathbb{P}_{\tilde{Q}}\left[S=1,X=x,Y=y\right]\\
&=\mathbb{P}_{\tilde{Q}}\left[S=1 \mid X=x,Y=y\right]\cdot\mathbb{P}_{\tilde{Q}}\left[X=x,Y=y\right]\\
&=\frac{1}{N}\cdot\frac{dP_{X,Y}(x,y)}{dQ_{X,Y}(x,y)}\cdot dQ_{X,Y}(x,y)\\
&=\frac{1}{N}\cdot dP(x,y).
\end{aligned}$$

In addition, from (56), we have that

$$\begin{aligned}
\mathbb{P}_{\tilde{Q}}\left[S=1\right]&=\mathbb{E}_{\tilde{Q}_X}\left[\mathbb{P}_{\tilde{Q}}\left[S=1 \mid X=x\right]\right]\\
&=\mathbb{E}_{\tilde{Q}_X}\left[\frac{1}{N}\cdot\frac{dP_X(X)}{dQ_X(X)}\right]\\
&=\frac{1}{N}.
\end{aligned}$$

Thus, we have that

$$d\tilde{Q}_{X,Y|S=1}(x,y)=\frac{d\tilde{Q}_{X,Y,S=1}(x,y)}{\mathbb{P}_{\tilde{Q}}\left[S=1\right]}=dP_{X,Y}(x,y).$$

Therefore, we have that Q can generate P under Γ-biased sampling.

C.4 Proof of Lemma 5

Suppose for the sake of contradiction (h,α) is a minimizer of the population RU risk and $(h,\alpha)\notin\Theta$. There are three cases

1. $(h,\alpha)\in\mathcal{H}^c\times\mathcal{A}$,
2. $(h,\alpha)\in\mathcal{H}\times\mathcal{A}^c$,
3. $(h,\alpha)\in\mathcal{H}^c\times\mathcal{A}^c$.

First, we focus on the case where $(h,\alpha)\in\mathcal{H}^c\times\mathcal{A}$. We consider $\bar{h}$,

$$\bar{h}(x)=\begin{cases}h(x) & h(x)\in[-2B,2B]\\ 2B & h(x)>2B\\ -2B & h(x)<-2B\end{cases}.$$

We note that $(\bar{h},\alpha)\in\Theta$. We define $R_{h,-2B}$ and $S_{h,2B}$ following (44) and (45).

$$\begin{aligned}
&\mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y)\right] - \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU}}(\bar{h}(X),\alpha(X),Y)\right] \\
&= \mathbb{E}_P\left[(L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y) - L^{\Gamma}_{\mathrm{RU}}(\bar{h}(X),\alpha(X),Y)\mathbb{I}(R_{h,-2B})\right] \\
&\quad + \mathbb{E}_P\left[(L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y) - L^{\Gamma}_{\mathrm{RU}}(\bar{h}(X),\alpha(X),Y)\mathbb{I}(S_{h,2B})\right]
\end{aligned}$$

because they only differ on $R_{h,-2B}$ and $S_{h,2B}$. Analyzing the second term on the right side above, we see that

$$\begin{aligned}
&\mathbb{E}_P\left[(L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y) - L^{\Gamma}_{\mathrm{RU}}(\bar{h}(X),\alpha(X),Y))\cdot\mathbb{I}(S_{h,2B})\right] \\
&= \mathbb{E}_{P_X}\left[\Big(\Gamma^{-1}T_{1,X}(h(X),\alpha(X))) + (\Gamma-\Gamma^{-1})\cdot T_{3,X}(h(X),\alpha(X))\Big)\mathbb{I}(S_{h,2B})\right],
\end{aligned}$$

where $T_{1,X}, T_{3,X}$ are defined in Lemma 22 and Lemma 23, respectively. For $x \in S_{h,2B}$,

$$\begin{aligned}
&\Gamma^{-1}T_{1,x}(h(x),\alpha(x)) + (\Gamma-\Gamma^{-1})\cdot T_{3,x}(h(x),\alpha(x)) - \Gamma^{-1}T_{1,x}(\bar{h}(x),\alpha(x)) - (\Gamma-\Gamma^{-1})\cdot T_{3,x}(\bar{h}(x),\alpha(x)) \\
&= (h(x)-\bar{h}(x))\cdot\Big(\Gamma^{-1}T^{c}_{1,x}(\tilde{h}(x),\alpha(x)) + (\Gamma-\Gamma^{-1})\cdot T^{c}_{3,X}(\tilde{h}(x),\alpha(x))\Big) \quad \tilde{h}(x)\in[\bar{h}(x),h(x)] && \text{(58a)} \\
&= (h(x)-\bar{h}(x))\cdot\mathbb{E}_{P_{Y|X=x}}\left[\Gamma^{-1}\cdot(-\ell'(Y-\tilde{h}(x))) + (\Gamma-\Gamma^{-1})\cdot(-\ell'(Y-\tilde{h}(x)))\cdot\mathbb{I}(\ell(Y-\tilde{h}(x))>\alpha(x))\right] && \text{(58b)} \\
&\geq (h(x)-\bar{h}(x))\cdot\mathbb{E}_{P_{Y|X=x}}\left[\Gamma^{-1}\cdot(-\ell'(Y-\tilde{h}(x)))\right] && \text{(58c)} \\
&> 0. && \text{(58d)}
\end{aligned}$$

(58a) follows from the Mean Value Theorem, the differentiability of $T_{1,x}$ (Lemma 22), and the differentiability of $T_{3,x}$ (Lemma 23). (58b) follows from Lemma 22 and Lemma 23. The inequality in (58c) comes from the observation that for $x \in S_{h,2B}$, we have that $Y - \tilde{h}(x) \leq -B$ because $Y \in [-B, B]$ and $\tilde{h}(x) \in [2B, h(x)]$. So, $-\ell'(Y-\tilde{h}(x)) > 0$. Meanwhile, $h(x) - \bar{h}(x) > 0$. So, the product of $-\ell'(Y-\tilde{h}(x))\cdot(h(x)-\bar{h}(x)) > 0$. Since $\Pr(\ell(Y-\tilde{h}(x)) > \alpha(x)|X=x) \geq 0$, (58c) holds. For the same reason, (58d) holds as well. Thus, if $S_{h,2B}$ has positive measure, then

$$\mathbb{E}_P\left[(L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y) - L^{\Gamma}_{\mathrm{RU}}(\bar{h}(X),\alpha(X),Y))\mathbb{I}(X\in S_{h,2B})\right] > 0.$$

An analogous argument can be used to show that for $R_{h,-2B}$ with positive measure,

$$\mathbb{E}_P\left[(L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y) - L^{\Gamma}_{\mathrm{RU}}(\bar{h}(X),\alpha(X),Y))\mathbb{I}(X\in R_{h,-2B})\right] > 0.$$

Thus, as long as $R_{h,-2B}\cup S_{h,2B}$ has positive measure, which must be the case under our assumption that the minimizer $(h,\alpha)\in\mathcal{H}^c\times\mathcal{A}$, then there is $(\bar{h},\alpha)\in\Theta$ that achieves lower population RU risk. This is a contradiction, so the minimizer cannot be in $\mathcal{H}^c\times\mathcal{A}$.

Now, we consider the next case that the minimizer $(h,\alpha)\in\mathcal{H}\times\mathcal{A}^c$. Consider $\bar{\alpha}\in\mathcal{A}$,

$$\bar{\alpha}(x) = \begin{cases} 0 & \alpha(x) < 0 \\ \alpha(x) & 0 \leq \alpha(x) \leq M_u \\ M_u & \alpha(x) > M_u \end{cases}$$

Note that $(h,\bar{\alpha})\in\Theta$. We define $R_{\alpha,0}$ and S_{α,M_u} according to (44) and (45), respectively. We have that

$$\begin{aligned}
&\mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y)\right] - \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU}}(h(X),\bar{\alpha}(X),Y)\right] \\
&= \mathbb{E}_P\left[(L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y) - L^{\Gamma}_{\mathrm{RU}}(h(X),\bar{\alpha}(X),Y))\mathbb{I}(R_{\alpha,0})\right] \\
&\quad + \mathbb{E}_P\left[(L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y) - L^{\Gamma}_{\mathrm{RU}}(h(X),\bar{\alpha}(X),Y))\mathbb{I}(S_{\alpha,M_u})\right].
\end{aligned}$$

because they only differ on $R_{\alpha,0}$ and S_{α,M_u}. We find that

$$\begin{aligned}
&\mathbb{E}_P\left[(L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y) - L^\Gamma_{\mathrm{RU}}(h(X),\bar{\alpha}(X),Y))\cdot\mathbb{I}(R_{\alpha,0})\right] \\
&= (1-\Gamma^{-1})\mathbb{E}_P\left[(\alpha(X)-\bar{\alpha}(X))\mathbb{I}(R_{\alpha,0})\right] + (\Gamma-\Gamma^{-1})\mathbb{E}_P\left[(L(h(X),Y)-\alpha(X))_+\cdot\mathbb{I}(R_{\alpha,0})\right] \\
&\quad - (\Gamma-\Gamma^{-1})\mathbb{E}_P\left[(L(h(X),Y)-\bar{\alpha}(X))_+\cdot\mathbb{I}(R_{\alpha,0})\right] \\
&= (1-\Gamma^{-1})\mathbb{E}_X\left[\alpha(X)\mathbb{I}(R_{\alpha,0})\right] + (\Gamma-\Gamma^{-1})\mathbb{E}_P\left[(L(h(X),Y)-\alpha(X))\mathbb{I}(R_{\alpha,0})\right] \\
&\quad - (\Gamma-\Gamma^{-1})\mathbb{E}_P\left[L(h(X),Y)\mathbb{I}(R_{\alpha,0})\right] \\
&= (1-\Gamma)\mathbb{E}_P\left[\alpha(X)\cdot\mathbb{I}(R_{\alpha,0})\right].
\end{aligned}$$

If $R_{\alpha,0}$ has positive measure, then

$$\mathbb{E}_P\left[(L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y) - L^\Gamma_{\mathrm{RU}}(h(X),\bar{\alpha}(X),Y))\cdot\mathbb{I}(R_{\alpha,0})\right] > 0$$

because on $R_{\alpha,0}$, we have that $\alpha(X) < 0$ and also $(1-\Gamma) < 0$. In addition,

$$\begin{aligned}
&\mathbb{E}_P\left[(L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y) - L^\Gamma_{\mathrm{RU}}(h(X),\bar{\alpha}(X),Y))\cdot\mathbb{I}(S_{\alpha,M_u})\right] \\
&= \mathbb{E}_{P_X}\left[\mathbb{E}_{P_{Y|X}}\left[L^\Gamma_{\mathrm{RU},2}(\alpha(X)) - L^\Gamma_{\mathrm{RU},2}(\bar{\alpha}(X)) + L^\Gamma_{\mathrm{RU},3}(h(X),\alpha(X),Y) - L^\Gamma_{\mathrm{RU},3}(h(X),\bar{\alpha}(X),Y)\mid X\right]\mathbb{I}(S_{\alpha,M_u})\right].
\end{aligned}$$

For $x \in S_{\alpha,M_u}$, we compute

$$\begin{aligned}
&\mathbb{E}_{P_{Y|X}}\left[L^\Gamma_{\mathrm{RU},2}(\alpha(X)) - L^\Gamma_{\mathrm{RU},2}(\bar{\alpha}(X)) + L^\Gamma_{\mathrm{RU},3}(h(X),\alpha(X),Y) - L^\Gamma_{\mathrm{RU},3}(h(X),\bar{\alpha}(X),Y)\mid X=x\right] && \text{(59a)}\\
&= \mathbb{E}_{P_{Y|X=x}}\left[(1-\Gamma^{-1})(\alpha(X)-\bar{\alpha}(X))\mid X=x\right] && \text{(59b)}\\
&\quad + \mathbb{E}_{P_{Y|X=x}}\left[(\Gamma-\Gamma^{-1})\Big(T_{3,X}(h(X),\alpha(X)) - T_{3,X}(h(X),\bar{\alpha}(X))\Big)\mid X=x\right] && \text{(59c)}\\
&= (1-\Gamma^{-1})(\alpha(x)-\bar{\alpha}(x)) + (\Gamma-\Gamma^{-1})\Big(T_{3,x}(h(x),\alpha(x)) - T_{3,x}(h(x),\bar{\alpha}(x))\Big) && \text{(59d)}\\
&= (1-\Gamma^{-1})(\alpha(x)-\bar{\alpha}(x)) + (\Gamma-\Gamma^{-1})\cdot(\alpha(x)-\bar{\alpha}(x))\cdot T^d_{3,x}(h(x),\tilde{\alpha}(x)) \quad \tilde{\alpha}(x)\in[\bar{\alpha}(x),\alpha(x)] && \text{(59e)}\\
&= (1-\Gamma^{-1})(\alpha(x)-\bar{\alpha}(x)) + (\Gamma-\Gamma^{-1})\cdot(\alpha(x)-\bar{\alpha}(x))\cdot(-1+F_{x;h(x)}(\tilde{\alpha}(x))) && \text{(59f)}\\
&> (1-\Gamma^{-1})(\alpha(x)-\bar{\alpha}(x)) + (\Gamma-\Gamma^{-1})\cdot(\alpha(x)-\bar{\alpha}(x))\cdot(-1+\eta(\Gamma)) && \text{(59g)}\\
&= 0. && \text{(59h)}
\end{aligned}$$

In the above derivation, we have that (59d) follows from Lemma 23 and Assumption 2. Next, we apply the Mean Value Theorem to $T_{3,x}(c,d)$ to arrive at (59e). After that, we use the definition of $T^d_{3,x}(c,d)$ for $d>0$ from Lemma 23, where $\tilde{\alpha}(x) > 0$. Finally, we recall that $F_{x;h(x)}$ is the distribution over $L(h(x),Y) = \ell(Y - h(x))$ when Y is distributed according to $P_{Y|X=x}$. We can show (59g) as follows. Since $\tilde{\alpha}(x)\in[\bar{\alpha}(x),\alpha(x)]$ and $x\in S_{\alpha,M_u}$, we have that

$$F_{x;h(x)}(\tilde{\alpha}(x)) \geq F_{x;h(x)}(\bar{\alpha}(x)) = F_{x;h(x)}(M_u),$$

and we have that

$$q^L_{\eta(\Gamma)}(x;h(x)) = F^{-1}_{x;h(x)}(\eta(\Gamma)) < M_u$$

by the definition of M_u (25). So, we see that $\eta(\Gamma) < F_{x;h(x)}(M_u)$. In addition, we note that $\alpha(x)-\bar{\alpha}(x) > 0$ for $x\in S_{\alpha,M_u}$ and $\Gamma-\Gamma^{-1} > 0$. We conclude that if S_{α,M_u} has positive measure, then

$$\mathbb{E}_P\left[(L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y) - L^\Gamma_{\mathrm{RU}}(h(X),\bar{\alpha}(X),Y))\cdot\mathbb{I}(S_{\alpha,M_u})\right] > 0.$$

Thus, as long as $R_{\alpha,0}\cup S_{\alpha,M_u}$ has positive measure, which must be the case because we assumed that $(h,\alpha)\in\mathcal{H}\times\mathcal{A}^c$, there is $(h,\bar{\alpha})\in\Theta$ that achieves lower population RU risk than the minimizer (h,α). This is a contradiction, so any minimizer cannot be in $\mathcal{H}\times\mathcal{A}^c$.

Combining the two previous arguments, we can show that any minimizer also cannot be in $\mathcal{H}^c\times\mathcal{A}^c$. Thus, any minimizer of the population RU risk must lie in Θ.

C.5 Proof of Lemma 6

The main goal of this proof is to apply Lemma 19 to the function $\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\right]$ and set Θ. Clearly, the population RU risk is continuous. We have the RU loss is convex from the first part of Theorem 2, so the population RU risk is also convex in (h,α). In addition, $\Theta \subset L^2(P_X,\mathcal{X}) \times L^2(P_X,\mathcal{X})$, which is a Hilbert space. In addition, since L^∞ balls are closed in $L^2(P_X,\mathcal{X})$, and Θ consists of a product of L^∞ balls (one of which is not centered at 0), so Θ is closed in $L^2(P_X,\mathcal{X})$. Also, Θ is convex and bounded. Thus Lemma 19 holds, so $\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\right]$ must achieve a minimum on Θ.

C.6 Proof of Lemma 7

Let

$$\begin{aligned} F(h,\alpha) &= \mathbb{E}_P\left[L^\Gamma_{\mathrm{RU},3}(h(X),\alpha(X),Y)\right] \\ G(h) &= \mathbb{E}_P\left[L^\Gamma_{\mathrm{RU},1}(h(X),Y)\right] \\ H(h,\alpha) &= \mathbb{E}_P\left[L^\Gamma_{\mathrm{RU},1}(h(X),Y)\right] + \mathbb{E}_P\left[L^\Gamma_{\mathrm{RU},3}(h(X),\alpha(X),Y)\right]. \end{aligned}$$

Note that

$$\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\right] = H(h,\alpha) + \mathbb{E}_P\left[L^\Gamma_{\mathrm{RU},2}(\alpha(X))\right]. \tag{60}$$

Since the population RU risk is the sum of H and a function that is convex in (h,α), then it suffices to show that H is strictly convex. The main goal of this proof is to show that the conditions of Lemma 21 hold so that we can conclude that H, as defined above, is strictly convex in (h,α).

First, we note that F, G, H are all Gâteaux differentiable by Lemma 25.

Second, we show that G satisfies the conditions of Lemma 21. By Lemma 26, G is strongly convex with constant $\Gamma^{-1}C_{L,l}$.

Third, we show that F satisfies the conditions of Lemma 21. It follows from the first part of Theorem 2 that F is jointly convex in (h,α). Also, F is strictly convex in α on $\mathcal{A}$ by Lemma 27.

As a result, F, G satisfy the conditions of Lemma 21. So, we have that $H(h,\alpha)$ is strictly convex in (h,α). Furthermore, because $\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU},2}(\alpha)\right]$ is convex in α and does not depend on h, it is also jointly convex in (h,α). Due to the decomposition in (60), $\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\right]$ is the sum of a strictly convex function and a convex function in (h,α), and is thus strictly convex.

C.7 Proof of Theorem 8

First, by Lemma 5 we have that

$$\operatorname*{argmin}_{(h,\alpha)\in L^2(P_X,\mathcal{X})\times L^2(P_X,\mathcal{X})} \mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\right] = \operatorname*{argmin}_{(h,\alpha)\in\Theta} \mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\right].$$

Second, we can show that $\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\right]$ has a unique minimizer on Θ. By Lemma 7, we have that $\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\right]$ is strictly convex on Θ, so it has at most one minimizer on the convex set Θ. From Lemma 6, we have that $\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\right]$ has at least one minimizer on Θ. Thus, there is a unique minimizer $(h^*_\Gamma,\alpha^*_\Gamma)$ on Θ. Finally, $(h^*_\Gamma,\alpha^*_\Gamma)$ is also the unique minimizer over $L^2(P_X,\mathcal{X})\times L^2(P_X,\mathcal{X})$.

C.8 Proof of Lemma 9

Let $L(h,\alpha) = \mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(h(X),\alpha(X),Y)\right]$ as the population RU risk. Since $L(h,\alpha)$ is Gâteaux differentiable (Lemma 25) and has a unique minimizer at $(h^*_\Gamma,\alpha^*_\Gamma)$ (Theorem 8), we can use Lemma 17 to realize that the Gâteaux derivative in the direction ϕ is equal to 0 for all $\phi \in L^2(P_X,\mathcal{X})$, i.e.

$$L'_\alpha(h^*_\Gamma,\alpha^*_\Gamma;\phi) = 0, \quad \forall \phi \in L^2(P_X,\mathcal{X}).$$

Recall that from Lemma 25, we have that

$$L'_\alpha(h,\alpha;\phi) = (1-\Gamma^{-1})\mathbb{E}_{P_X}\left[\phi(X)\right] + (\Gamma-\Gamma^{-1})\cdot\mathbb{E}_{P_X}\left[T^d_{3,X}(h^*_\Gamma(X),\alpha^*_\Gamma(X))\phi(X)\right].$$

So, at $(h^*_\Gamma, \alpha^*_\Gamma)$, we have that

$$\mathbb{E}_{P_X}\left[\phi(X)\cdot\left(\frac{1-\Gamma^{-1}}{\Gamma-\Gamma^{-1}}+T^d_{3,X}(h^*_\Gamma(X),\alpha^*_\Gamma(X))\right)\right]=0,\quad \forall\phi\in L^2(P_X,\mathcal{X}).$$

We note that by Lemma 23,

$$T^d_{3,x}(h(x),\alpha(x))=-1+F_{x;h(x)}(\alpha(x)),$$

where $F_{x;h(x)}$ is the distribution over $L(h(x),Y)$ where Y is distributed according to $P_{Y|X=x}$. So, we have that

$$\mathbb{E}_{P_X}\left[\phi(X)\cdot(-\eta(\Gamma)+F_{X,h^*_\Gamma(X)}(\alpha^*_\Gamma(X))\right]=0,\quad \forall\phi\in L^2(P_X,\mathcal{X}).$$

So, $-\eta(\Gamma)+F_{x,h^*_\Gamma(x)}(\alpha^*_\Gamma(x))$ must be equal to 0 almost everywhere for the above equation to hold for all ϕ. Therefore, we conclude that

$$\alpha^*_\Gamma(x)=F^{-1}_{x;h^*_\Gamma(x)}(\eta(\Gamma))=q^L_{\eta(\Gamma)}(x;h^*_\Gamma(x)).$$

Now, with this definition of α^*_Γ, we can show that there exists $M_l>0$ such that $\alpha^*_\Gamma(x)>M_l$ for all $x\in\mathcal{X}$. We aim to show that $\inf_{x\in\mathcal{X},h\in\mathcal{H}} q^L_{\eta(\Gamma)}(x;h(x))>0$. For convenience, we define

$$m(x):=q^L_{\frac{1}{2}}(x;h(x)).$$

We note that $\eta(\Gamma)>\frac{1}{2}$. So,

$$q^L_{\eta(\Gamma)}(x;h(x))\geq m(x).$$

We have that for any $x\in\mathcal{X},h\in\mathcal{H}$,

$$\Pr(L(h(X),Y)\leq m(X)\mid X=x)=\frac{1}{2}.$$

Recall that under Assumption 2, $L(h(x),y)=\ell(y-h(x))$. We can apply Assumption 2 to see that

$$\Pr(Y\in[h(x)+\ell_2^{-1}(m(x)),h(x)+\ell_1^{-1}(m(x))]\mid X=x)=\frac{1}{2}.$$

Now, we can use the upper bound on the density of $p_{Y|X=x}$ from Assumption 3 to see that

$$C_{p,u}\cdot(\ell_1^{-1}(m(x))-\ell_2^{-1}(m(x)))\geq\frac{1}{2}.$$

Rearranging, we have that

$$\ell_1^{-1}(m(x))-\ell_2^{-1}(m(x))\geq\frac{1}{2C_{p,u}}.$$

So,

$$\max\{\ell_1^{-1}(m(x)),-\ell_2^{-1}(m(x))\}\geq\frac{1}{4C_{p,u}}.$$

Applying ℓ to both sides, we conclude that

$$m(x)\geq\ell\left(\frac{1}{4C_{p,u}}\right).$$

Since $\frac{1}{4C_{p,u}}>0$, we have that $m(x)$ is lower bounded by a positive constant for any choice of $h\in\mathcal{H},x\in\mathcal{X}$. Thus, we have that

$$\alpha^*_\Gamma(x)=q^L_{\eta(\Gamma)}(x;h(x))\geq\inf_{x\in\mathcal{X},h\in\mathcal{H}}q^L_{\frac{1}{2}}(x;h(x))\geq\ell\left(\frac{1}{4C_{p,u}}\right).$$

So, let $M_l=\ell(\frac{1}{4C_{p,u}})/2$. Then $\alpha^*(x)>M_l$ for all $x\in\mathcal{X}$.

C.9 Proof of Theorem 10

The constant for strong convexity depends on lower bounds on the conditional density $p_{Y|X=x}(\cdot)$ and on $|\ell'(\ell^{-1}(\cdot))|$ when these functions are evaluated over a particular region. To ensure that they can be lower bounded, we pick the radius of the $||\cdot||_\infty$ ball about the minimizer.

Define

$$g_i(x; h, \alpha) = h(x) + \ell_i^{-1}(\alpha(x)). \tag{61}$$

For $0 < \epsilon < \frac{1-\eta(\Gamma)}{2C_{p,u}}$, we pick $0 < \delta(\epsilon) < M_l$ to ensure that for $(h, \alpha) \in \mathcal{C}_{\delta(\epsilon)}$, we have that

$$\sup_{x\in\mathcal{X}, i\in\{1,2\}} |g_i(x; h, \alpha) - g_i(x; h^*_\Gamma, \alpha^*_\Gamma)| < \epsilon.$$

By Lemma 9, we have that $\alpha^*_\Gamma(x) > M_l$ for all $x \in \mathcal{X}$. We consider $(h, \alpha) \in \mathcal{C}_{\delta(\epsilon)}$. For such α, we have that $||\alpha - \alpha^*_\Gamma||_\infty \le \delta(\epsilon)$, and so $\alpha(x) \ge M_l - \delta(\epsilon)$ for all $x \in \mathcal{X}$. Since $\delta(\epsilon) < M_l$, for $(h, \alpha) \in \mathcal{C}_{\delta(\epsilon)}$, we have that $\alpha(x) > 0$. Since the RU loss is twice-differentiable when $\alpha(x) > 0$ (Lemma 25), we have that it is twice-differentiable on $\mathcal{C}_{\delta(\epsilon)}$.

Let $L(h, \alpha), L_1(h, \alpha), L_3(h, \alpha)$ be shorthand for the population RU risk, the first term of the population RU risk, and the third term of the population RU risk, respectively.

$$\begin{aligned}
L(h, \alpha) &= \mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(h(X), \alpha(X), Y)\right], \\
L_1(h, \alpha) &= \mathbb{E}_P\left[L^\Gamma_{\mathrm{RU},1}(h(X), Y)\right], \\
L_3(h, \alpha) &= \mathbb{E}_P\left[L^\Gamma_{\mathrm{RU},3}(h(X), \alpha(X), Y)\right].
\end{aligned}$$

We compute the second Gâteaux derivative of the population RU risk.

$$\begin{aligned}
&\langle L''(h, \alpha; \psi, \phi), (\psi, \phi)\rangle && \text{(62a)}\\
&= \langle L''_1(h, \alpha; \psi, \phi) + L''_3(h, \alpha; \psi, \phi), (\psi, \phi)\rangle && \text{(62b)}\\
&\ge \langle L''_3(h, \alpha; \psi, \phi), (\psi, \phi)\rangle && \text{(62c)}\\
&= (\Gamma - \Gamma^{-1}) \cdot \mathbb{E}_{P_X}\left[\begin{bmatrix}\psi(X) & \phi(X)\end{bmatrix} \nabla^2 T_{3,X}(h(X), \alpha(X)) \begin{bmatrix}\psi(X)\\ \phi(X)\end{bmatrix}\right] && \text{(62d)}\\
&\ge (\Gamma - \Gamma^{-1})\mathbb{E}_{P_X}\left[\begin{bmatrix}\psi(X) & \phi(X)\end{bmatrix} A_X(h(X), \alpha(X)) \begin{bmatrix}\psi(X)\\ \phi(X)\end{bmatrix}\right] && \text{(62e)}
\end{aligned}$$

(62b) follows from Lemma 25. (62c) holds because $\langle L''_1(h, \alpha; (\psi, \phi), (\psi, \phi))\rangle \ge 0$ because $L_1(h, \alpha)$ is strongly convex in h (Lemma 26) and does not depend on α. Next, (62d) follows from Lemma 23. Finally, $A_x(c, d)$ is the lower bound on the Hessisan matrix of $T_{3,x}(c, d)$ defined in Lemma 24.

To develop a lower bound for $\langle L''(h, \alpha; \psi, \phi), (\psi, \phi)\rangle$, we aim to apply Lemma 20 to $A_x(h(x), \alpha(x))$. Before we verify the conditions of Lemma 20, we introduce the following notation

$$\begin{aligned}
a_{i,x} &:= |\ell'(\ell_i^{-1}(\alpha(x)))| \quad i = 1, 2,\\
f_{i,x} &:= p_{Y|X=x}(h(x) + \ell_i^{-1}(\alpha(x))) \quad i = 1, 2,
\end{aligned}$$

and we develop upper and lower bounds on $a_{i,x}$ for $i \in \{1, 2\}, \sum_{i\in\{1,2\}} f_{i,x}$, and $1 - F_{x;h(x)}(\alpha(x))$.

First, we focus on $a_{i,x}$. By the definition of Θ, we have that $\alpha(x) \le M_u$. Since $|\ell'(\ell_i^{-1}(y))|$ is strictly increasing in y and on $C_{\delta(\epsilon)}$, $\alpha(x) \ge M_l - \delta(\epsilon)$ for all $x \in \mathcal{X}$, we can recall the definition of $C_{a,l,\delta}, C_{a,u}$ from (30), (26) to see that

$$0 < C_{a,l,\delta(\epsilon)} \le a_{i,x} \le C_{a,u} < \infty \quad i = 1, 2, x \in \mathcal{X}.$$

Second, we aim to show that $\sum_{i=\{1,2\}} f_{i,x}$ is similarly upper and lower bounded. The upper bound is straightforward from Assumption 3. To obtain the lower bound, we first analyze $\sum_{i\in\{1,2\}} p_{Y|X=x}(h^*_\Gamma(x) + \ell_i^{-1}(\alpha^*_\Gamma(x))$, which can be written as $\sum_{i\in\{1,2\}} p_{Y|X=x}(g_i(x; h^*_\Gamma, \alpha^*_\Gamma))$ using the definition of g in (61).

Let $\ell_i^{-1}(q^L_{\eta(\Gamma)}(x; h^*_\Gamma))$ corresponds to the $c_{i,x}$-th quantile of Y, where Y is distributed following $P_{Y|X=x}$. We realize that

$$\begin{aligned}\sum_{i\in\{1,2\}} p_{Y|X=x}(g_i(x;h^*_\Gamma,\alpha^*_\Gamma)) &= \sum_{i\in\{1,2\}} p_{Y|X=x}(h^*_\Gamma(x)+\ell_i^{-1}(\alpha^*_\Gamma(x))\\ &= \sum_{i\in\{1,2\}} p_{Y|X=x}(h^*_\Gamma(x)+\ell_i^{-1}(q^L_{\eta(\Gamma)}(x;h^*_\Gamma(x))))\\ &= \sum_{i\in\{1,2\}} p_{Y|X=x}(h^*_\Gamma(x)+q^Y_{c_{i,x}}(x)-h^*_\Gamma(x))\\ &= \sum_{i\in\{1,2\}} p_{Y|X=x}(q^Y_{c_{i,x}}(x)).\end{aligned}$$

Furthermore, we realize that either $c_{1,x}$ or $c_{2,x}$ lies in $[1-\frac{\eta(\Gamma)}{2}, 1+\frac{\eta(\Gamma)}{2}]$. First, because $q^L_{\eta(\Gamma)}(x;h^*_\Gamma(x))$ corresponds to the $\eta(\Gamma)$-th quantile of the conditional losses, we must have that

$$c_{1,x}-c_{2,x}=\eta(\Gamma). \tag{63}$$

In addition, $c_{1,x}\le 1$, so $c_{2,x}\le 1-\eta(\Gamma)$. So, $c_{2,x}\in[0,1-\eta(\Gamma)]$. Suppose that $c_{2,x}\in[1-\frac{\eta(\Gamma)}{2},1-\eta(\Gamma)]$, then clearly the desired claim holds. If $c_{2,x}\notin[1-\frac{\eta(\Gamma)}{2},1-\eta(\Gamma)]$, this means that $c_{2,x}\in[0,1-\frac{\eta(\Gamma)}{2})$. So, we must have that $c_{1,x}\in[\eta(\Gamma),1+\frac{\eta(\Gamma)}{2})$. Thus, we have that at least one of $c_{1,x},c_{2,x}$ lies in the interval $[1-\frac{\eta(\Gamma)}{2},1+\frac{\eta(\Gamma)}{2}]$.

Now, we have that

$$\sum_{i\in\{1,2\}} f_{i,x}=\sum_{i\in\{1,2\}} p_{Y|X=x}(g_i(x;h,\alpha)),$$

and $\delta(\epsilon)$ was chosen so that for $(h,\alpha)\in\mathcal{C}_{\delta(\epsilon)}$

$$\sup_{x\in\mathcal{X},i\in\{1,2\}} |g_i(x;h,\alpha)-g_i(x;h^*_\Gamma,\alpha^*_\Gamma)| = \sup_{x\in\mathcal{X},i\in\{1,2\}} |g_i(x;h,\alpha)-p_{Y|X=x}(q^Y_{c_{i,x}}(x))|<\epsilon.$$

Thus, we realize that for $(h,\alpha)\in\mathcal{C}_{\delta(\epsilon)}$,

$$g_i(x;h,\alpha)=q^Y_{c_{i,x}}(x)+b_i(x),\quad b_i(x)\in(-\epsilon,\epsilon), i\in\{1,2\}, x\in\mathcal{X}. \tag{64}$$

So, for $(h,\alpha)\in\mathcal{C}_{\delta(\epsilon)}$, we realize that a lower bound on $\sum_{i\in\{1,2\}} f_{i,x}=\sum_{i\in\{1,2\}} p_{Y|X}(g_i(x;h,\alpha))$ is given by $C_{p,l,\epsilon}$ from (31). Thus, we have that

$$0<C_{p,l,\epsilon}\le\sum_{i\in\{1,2\}} f_{i,x}\le 2C_{p,u}<\infty\quad i=1,2,x\in\mathcal{X},$$

and clearly each $f_{i,x}$ must be nonnegative.

Third, we aim to show that $1-F_{x;h(x)}(\alpha(x))$ is similarly upper and lower bounded on $\mathcal{C}_{\delta(\epsilon)}$. Clearly, an upper bound on this quantity is 1. To compute the lower bound, we see that for $(h,\alpha)\in\mathcal{C}_{\delta(\epsilon)}$,

$$\begin{aligned}1-F_{x;h(x)}(\alpha(x)) &= 1-P_{Y|X=x}(g_1(x;h,\alpha))+P_{Y|X=x}(g_2(x;h,\alpha))\\ &= 1-P_{Y|X=x}(q^Y_{c_{1,x}}(x)+b_1(x))+P_{Y|X=x}(q^Y_{c_{2,x}}(x)+b_2(x))\quad b_1(x),b_2(x)\in(-\epsilon,\epsilon)\\ &\ge 1-c_{1,x}-C_{p,u}\cdot\epsilon+c_{2,x}-C_{p,u}\cdot\epsilon\\ &= 1-\eta(\Gamma)-2C_{p,u}\epsilon\\ &> 0.\end{aligned}$$

The first line follows from the definition of F and g_i from (61). In the second line, we apply (64). In the third line, we note that the c.d.f. of $P_{Y|X=x}$ at $q^Y_{c_{i,x}}(x)+b_i(x)$ can be closely approximated by the value of the c.d.f. at $q^Y_{c_{i,x}}(x)$. Next, we apply (63). The last line follows because $\epsilon<\frac{1-\eta(\Gamma)}{2C_{p,u}}$. Thus, we have that

$$1-F_{x;h(x)}(\alpha(x))\ge 1-\eta(\Gamma)-2C_{p,u}\epsilon>0. \tag{65}$$

Now, we finally verify the conditions of Lemma 20. We note that $A_x(h(x), \alpha(x))$ is a symmetric matrix by definition. We realize that $\operatorname{tr} A_x(h(x), \alpha(x)) \geq 0$ because

$$\begin{aligned}
\operatorname{tr} A_x(h(x), \alpha(x)) &= A_{x,11}(h(x), \alpha(x)) + A_{x,22}(h(x), \alpha(x)) && (66)\\
&= \sum_{i\in\{1,2\}} a_{i,x} \cdot f_{i,x} + \sum_{i\in\{1,2\}} \frac{f_{i,x}}{a_{i,x}} + C_{L,\ell}(1 - F_{x;h(x)}(\alpha(x))) && (67)\\
&\geq C_{L,\ell}(1 - F_{x;h(x)}(\alpha(x))) && (68)\\
&> 0. && (69)
\end{aligned}$$

(68) follows from the observation that $f_{i,x}, a_{i,x} \geq 0$. (69) follows from (65). In addition, we see that $\det A_x(h(x), \alpha(x)) \geq 0$ because

$$\begin{aligned}
&\det A_x(h(x), \alpha(x)) && (70a)\\
&= A_{x,11}(h(x), \alpha(x)) \cdot A_{x,22}(h(x), \alpha(x)) - (A_{x,12}(h(x), \alpha(x)))^2 && (70b)\\
&= \Big(\sum_{i\in\{1,2\}} a_{i,x} \cdot f_{i,x} + C_{L,l}(1 - F_{x;h(x)}(\alpha(x))) \Big) \cdot \Big(\sum_{i\in\{1,2\}} \frac{f_{i,x}}{a_{i,x}} \Big) - (f_{1,x} - f_{2,x})^2 && (70c)\\
&= \Big(\frac{a_{1,x}}{a_{2,x}} + \frac{a_{2,x}}{a_{1,x}} + 2 \Big) \cdot f_{1,x} \cdot f_{2,x} + C_{L,l}(1 - F_{x;h(x)}(\alpha(x))) \cdot \Big(\sum_{i\in\{1,2\}} \frac{f_{i,x}}{a_{i,x}} \Big) && (70d)\\
&\geq C_{L,l} \cdot \Big(\sum_{i\in\{1,2\}} \frac{f_{i,x}}{a_{i,x}} \Big) \cdot (1 - F_{x;h(x)}(\alpha(x))) && (70e)\\
&\geq C_{L,l} \cdot \frac{1}{C_{a,u}} \cdot \Big(\sum_{i\in\{1,2\}} f_{i,x} \Big) \cdot (1 - \eta(\Gamma) - 2C_{p,u}\epsilon) && (70f)\\
&\geq C_{L,l} \cdot \frac{1}{C_{a,u}} \cdot C_{p,l,\epsilon} \cdot (1 - \eta(\Gamma) - 2C_{p,u}\epsilon) && (70g)\\
&> 0. && (70h)
\end{aligned}$$

Thus, we can apply Lemma 20 to $A_x(h(x), \alpha(x))$ to see that

$$\lambda_{\min}(A_x(h(x), \alpha(x)) \geq \frac{\det A_x(h(x), \alpha(x))}{\operatorname{tr} A_x(h(x), \alpha(x))}.$$

We can combine the lower bound on $\det A$ from (70g) with the following upper bound on $\operatorname{tr} A$ to find a lower bound on $\lambda_{\min}(A_x(h(x), \alpha(x))$ that does not depend on the choice of $x \in \mathcal{X}$ and $(h, \alpha) \in \mathcal{C}_{\delta(\epsilon)}$.

$$\begin{aligned}
\operatorname{tr} A_x(h(x), \alpha(x)) &= \sum_{i\in\{1,2\}} a_{i,x} \cdot f_{i,x} + \sum_{i\in\{1,2\}} \frac{f_{i,x}}{a_{i,x}} + C_{L,l}(1 - F_{x;h(x)}(\alpha(x))) && (71a)\\
&\leq 2C_{p,u}(C_{a,u} + \frac{1}{C_{a,l,\delta(\epsilon)}}) + C_{L,l} && (71b)\\
&= \frac{2C_{p,u}(C_{a,u} \cdot C_{a,l,\delta(\epsilon)} + 1) + C_{L,l} \cdot C_{a,l,\delta(\epsilon)}}{C_{a,l,\delta(\epsilon)}} && (71c)
\end{aligned}$$

Therefore, applying(71c) and (70g), we find that

$$\begin{aligned}
\lambda_{\min}(A_x(h(x), \alpha(x)) &\geq \frac{\det A_x(h(x), \alpha(x))}{\operatorname{tr} A_x(h(x), \alpha(x))}\\
&\geq \Big(C_{L,l} \cdot (1 - \eta(\Gamma) - 2C_{p,u}\epsilon) \cdot \frac{1}{C_{a,u}} \cdot C_{p,l,\epsilon} \Big) \cdot \Big(\frac{C_{a,l,\delta(\epsilon)}}{2C_{p,u}(C_{a,u} \cdot C_{a,l,\delta(\epsilon)} + 1) + C_{L,l} \cdot C_{a,l,\delta(\epsilon)}} \Big)\\
&\geq \frac{C_{L,l} \cdot (1 - \eta(\Gamma) - 2C_{p,u} \cdot \epsilon) \cdot C_{p,l,\epsilon}}{2C_{p,u} \cdot (C_{a,u} \cdot C_{a,l,\delta(\epsilon)} + 1) + C_{L,l} \cdot C_{a,l,\delta(\epsilon)}} \cdot \frac{C_{a,l,\delta(\epsilon)}}{C_{a,u}}.
\end{aligned}$$

Recall the definition of $\kappa_{1,\epsilon}$ from (32). We realize that for all $x \in \mathcal{X}, (h,\alpha) \in \mathcal{C}_{\delta(\epsilon)}$, we have that

$$(\Gamma - \Gamma^{-1}) \cdot A_x(h(x),\alpha(x)) \succeq \kappa_{1,\epsilon} \cdot I_2.$$

Revisiting (62e), we have that

$$\begin{aligned}
\langle L''(h,\alpha;(\psi,\phi)),(\psi,\phi)\rangle &\geq (\Gamma - \Gamma^{-1}) \cdot \mathbb{E}_{P_X}\left[\begin{bmatrix}\psi(X) & \phi(X)\end{bmatrix} A_X(h(X),\alpha(X)) \begin{bmatrix}\psi(X)\\ \phi(X)\end{bmatrix}\right] && (72)\\
&\geq \mathbb{E}_{P_X}\left[\begin{bmatrix}\psi(X) & \phi(X)\end{bmatrix} \kappa_{1,\epsilon} I_2 \begin{bmatrix}\psi(X)\\ \phi(X)\end{bmatrix}\right] && (73)\\
&= \kappa_{1,\epsilon}\mathbb{E}_{P_X}\left[\psi(X)^2 + \phi(X)^2\right] && (74)\\
&= \kappa_{1,\epsilon}||(\psi,\phi)||^2. && (75)
\end{aligned}$$

Thus, $\mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y)\right]$ is $\kappa_{1,\epsilon}$ strongly convex in (h,α) on $\mathcal{C}_{\delta(\epsilon)}$. We note that as $\epsilon \to 0$, then $\delta(\epsilon) \to 0$, as well. So, $C_{a,l,\delta(\epsilon)} \to C_{a,l}$, where $C_{a,l}$ is defined in (27) and $C_{p,l,\epsilon} \to C_{p,l}$, where $C_{p,l}$ is defined in (28), and $\epsilon \cdot C_{p,u} \to 0$. So, we have that

$$\begin{aligned}
&\lim_{\epsilon\to 0}(\Gamma - \Gamma^{-1})\frac{C_{L,l}\cdot(1-\eta(\Gamma)-2C_{p,u}\cdot\epsilon)\cdot C_{p,l,\epsilon}}{2C_{p,u}\cdot(C_{a,u}\cdot C_{a,l,\delta(\epsilon)}+1)+C_{L,l}\cdot C_{a,l,\delta(\epsilon)}}\cdot\frac{C_{a,l,\delta(\epsilon)}}{C_{a,u}}\\
&= (\Gamma - \Gamma^{-1})\cdot\frac{C_{L,l}\cdot(1-\eta(\Gamma))\cdot C_{p,l}}{2C_{p,u}\cdot(C_{a,u}\cdot C_{a,l}+1)+C_{L,l}\cdot C_{a,l}}\cdot\frac{C_{a,l}}{C_{a,u}}\\
&= (1 - \Gamma^{-1})\cdot\frac{C_{L,l}\cdot C_{p,l}}{2C_{p,u}\cdot(C_{a,u}\cdot C_{a,l}+1)+C_{L,l}\cdot C_{a,l}}\cdot\frac{C_{a,l}}{C_{a,u}}
\end{aligned}$$

Thus, as $\epsilon \to 0$, then $\kappa_{1,\epsilon} \to \kappa_1$, where κ_1 is defined in (29).

C.10 Proof of Theorem 11

Let $L(h,\alpha), L_1(h,\alpha), L_3(h,\alpha), a_{i,x}, f_{i,x}$ be defined as in the proof of Theorem 10. To show that the population RU risk is κ_2-smooth on $\mathcal{C}_{\delta(\epsilon)}$, we show that

$$\langle L''_{h\alpha}(h,\alpha;\psi,\phi),(\psi,\phi)\rangle \leq \kappa_2||(\psi,\phi)||^2_{L^2(P_X,\mathcal{X})}.$$

We have that

$$\begin{aligned}
&\langle L''(h,\alpha;\psi,\phi),(\psi,\phi)\rangle\\
&= \langle L''_1(h,\alpha;\psi,\phi) + L''_3(h,\alpha;\psi,\phi),(\psi,\phi)\rangle\\
&\leq \mathbb{E}_{P_X}\left[\begin{bmatrix}\psi(X) & \phi(X)\end{bmatrix}\cdot\left(\Gamma^{-1}\nabla^2 T_{1,X}(h(X),\alpha(X)) + (\Gamma-\Gamma^{-1})\nabla^2 T_{3,X}(h(X),\alpha(X))\right)\begin{bmatrix}\psi(X)\\ \phi(X)\end{bmatrix}\right]\\
&\leq \mathbb{E}_{P_X}\left[\begin{bmatrix}\psi(X) & \phi(X)\end{bmatrix}\cdot\left(\Gamma^{-1}\nabla^2 T_{1,X}(h(X),\alpha(X)) + (\Gamma-\Gamma^{-1})\nabla^2 B_X(h(X),\alpha(X))\right)\begin{bmatrix}\psi(X)\\ \phi(X)\end{bmatrix}\right],
\end{aligned}$$

The second line follows from Lemma 25. The matrix $B_x(h(x),\alpha(x))$ is as defined in Lemma 24. It suffices to show that there is $\kappa_{2,\epsilon}$ such that

$$\Gamma^{-1}\nabla^2 T_{1,x}(h(x),\alpha(x)) + (\Gamma-\Gamma^{-1})B_x(h(x),\alpha(x)) \preceq \kappa_{2,\epsilon} I_2 \quad \forall x \in \mathcal{X}.$$

Applying Lemma 22 and Assumption 4, we have that

$$\nabla^2 T_{1,x}(h(x),\alpha(x)) = \begin{bmatrix}\mathbb{E}_{P_{Y|X=x}}\left[\ell''(Y-h(x))\right] & 0\\ 0 & 0\end{bmatrix} \preceq C_{L,u} I_2. \tag{76}$$

From the proof of Theorem 10, for $0 < \epsilon < \frac{1-\eta(\Gamma)}{2C_{p,u}}$, there exists $0 < \delta(\epsilon) < M_l$ so that for $(h, \alpha) \in C_{\delta(\epsilon)}$, $(\Gamma - \Gamma^{-1})\nabla^2 T_{3,x}(h(x), \alpha(x))$ is positive definite. So, on this set $\mathcal{C}_{\delta(\epsilon)}, B_x(h(x), \alpha(x))$ is also certainly positive definite. So, $B_x(h(x), \alpha(x))$ satisfies the conditions of Lemma 20, so we can conclude that $\lambda_{\max}(B_x(h(x), \alpha(x))) \leq \operatorname{tr} B_x(h(x), \alpha(x))$. We can compute an upper bound on $\operatorname{tr} B_x(h(x), \alpha(x))$. We note that for $(h, \alpha) \in \mathcal{C}_{\delta(\epsilon)}$, $\alpha(x) \geq M_l - \delta(\epsilon)$ for all $x \in \mathcal{X}$ because $\alpha^*(x) > M_l$ by Lemma 9 and $||\alpha - \alpha^*||_\infty < \delta(\epsilon)$. So, we have that

$$\begin{aligned}\operatorname{tr} B_x(h(x), \alpha(x)) &= \sum_{i \in \{1,2\}} a_{i,x} \cdot f_{i,x} + \sum_{i \in \{1,2\}} \frac{f_{i,x}}{a_{i,x}} + \mathbb{E}_{P_{Y|X=x}}\left[\ell''(Y - h(x))\right] \\ &\leq 2C_{p,u}(C_{a,u} + \frac{1}{C_{a,l,\delta(\epsilon)}}) + C_{L,u}.\end{aligned}$$

We arrive at the second inequality by recalling the definition of $C_{p,u}$ from Assumption 3, $C_{a,u}$ from (26), $C_{a,l,\delta}$ from (30), and $C_{L,u}$ from Assumption 4. So, we have that

$$B_x(h(x), \alpha(x)) \preceq \left(2C_{p,u}(C_{a,u} + \frac{1}{C_{a,l}}) + C_{L,u}\right) I_2. \tag{77}$$

Combining the constants from (76) and (77), we have that for $(h, \alpha) \in \mathcal{C}_{\delta(\epsilon)}$

$$\Gamma^{-1}\nabla^2 T_{1,x}(h(x), \alpha(x)) + (\Gamma - \Gamma^{-1}) B_x(h(x), \alpha(x)) \preceq \kappa_{2,\epsilon} I_2 \quad \forall x \in \mathcal{X},$$

where $\kappa_{2,\epsilon}$ is defined as in (34). Thus, we conclude that $\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(h(X), \alpha(X), Y)\right]$ is $\kappa_{2,\epsilon}$-smooth in (h, α) on $\mathcal{C}_{\delta(\epsilon)}$. As, $\epsilon \to 0$, $\delta(\epsilon) \to 0$. So, $C_{a,l,\delta(\epsilon)} \to C_{a,l}$. This implies that $\kappa_{2,\epsilon} \to \kappa_2$ as the radius of the $||\cdot||_\infty$-ball shrinks.

C.11 Proof of Lemma 12

We note that Θ_m is a convex subset of Θ. By Lemma 7, the population RU risk is strictly convex on Θ. So, it is strictly convex on Θ_m, which means that it has at most one minimizer on Θ_m. In addition, by an analogous argument as the proof of Lemma 6, the population RU risk has at least one minimizer on Θ_m. Combining these two facts, it has a unique minimizer on Θ_m called θ^*_m.

C.12 Proof of Theorem 13

In this proof, we use the following lemma.

Lemma 28. *Define $\pi_m : \Theta \to \Theta_m$ to be the projection of θ^* onto Θ_m. Under Assumptions 1, 2,3,*

$$||\pi_m(\theta^*) - \theta^*||_{L^2(P_X, \mathcal{X})} \to 0.$$

Proof in Appendix D.11.

To simplify notation, let $L(\theta) = \mathbb{E}_P\left[L^\Gamma_{\mathrm{RU}}(\theta(X), Y)\right]$. For the sake of contradiction, assume that θ^*_m does not limit to θ^*. This means that there exists $\delta_1 > 0$ such that for every $m \in \mathbb{N}$, there is $A_m \geq m$ such that

$$||\theta^*_{A_m} - \theta^*||_{L^2(P_X, \mathcal{X})} > \delta_1.$$

We have that $\theta^* \in \Theta_m \subset \Theta$. In addition, $\theta^* \in \Theta$ by Lemma 5. By the strict convexity of the population RU risk on Θ (Lemma 7), for some $\epsilon > 0$, we have that

$$L(\theta_{A_m}) > L(\theta^*) + \epsilon$$

because by strict convexity, $||\theta - \theta^*||_{L^2(P_X, \mathcal{X})} > \delta_1$ implies that $L(\theta) > L(\theta^*) + \epsilon$ for some $\epsilon > 0$.

Note that $L(\theta)$ is continuous at θ^*, so there exists $\delta_2 > 0$ such that $||\theta - \theta^*||_{L^2(P_X, \mathcal{X})} < \delta_2$ implies that $|L(\theta) - L(\theta^*)| < \epsilon$. Since θ^* is the unique minimizer of the population RU risk, we have that $L(\theta) < L(\theta^*) + \epsilon$

in particular. Since $\pi_m(\theta^*) \to \theta^*$, there exists $M \in \mathbb{N}$ such that $||\pi_m(\theta^*) - \theta^*||_{L^2(P_X,\mathcal{X})} < \delta_2$ for $m \geq M$. By continuity of the population RU risk, we have that

$$L(\pi_m(\theta^*)) < L(\theta^*) + \epsilon \quad \text{for } m \geq M. \tag{78}$$

In addition, there exists $A_M \geq M$ so that $||\theta^*_{A_M} - \theta^*||_{L^2(P_X,\mathcal{X})} > \delta$, implying that

$$L(\theta^*_{A_M}) > L(\theta^*) + \epsilon.$$

However, this is a contradiction because $\theta^*_{A_M}$ is by definition the unique minimizer of the population RU risk over Θ_{A_M}, but we find that $\pi_{A_M}(\theta^*) \in \Theta_{A_M}$ satisfies

$$L(\pi_{A_M}(\theta^*)) < L(\theta^*_{A_M}).$$

Thus, we must have that $||\theta^*_m - \theta^*||_{L^2(P_X,\mathcal{X})} \to 0$ as $m \to \infty$.

C.13 Proof of Lemma 14

The goal of the proof is to verify that the conditions of Lemma 16 hold so that we can conclude that $\hat{\theta}_{m,n}$ exists with probability approaching 1 and $\hat{\theta}_{m,n} \xrightarrow{p} \theta^*_m$. First, we note that over the sieve space Θ_m, the population RU risk is uniquely minimized at θ^*_m by Theorem 12. To check the second condition, we observe that for m sufficiently large, $\theta^*_m \in \text{Int}(\Theta_m)$ because $\theta^*_m \to \theta^*$ by Theorem 13 and $\theta^* = (h^*, \alpha^*)$ where $0 < M_l \leq \alpha^*(x) < M_u$ for all $x \in \mathcal{X}$. Furthermore, it follows from the first part of Theorem 2 that $\theta \mapsto L^{\Gamma}_{\text{RU}}(\theta(x), y)$ is convex, which implies that the empirical risk $\widehat{\mathbb{E}}_P\left[L_{\text{RU}}(\theta(X), Y)\right]$ is also convex. Third, by the Weak Law of Large Numbers, we have the following pointwise convergence

$$\widehat{\mathbb{E}}_P\left[L^{\Gamma}_{\text{RU}}(\theta(X), Y)\right] \xrightarrow{p} \mathbb{E}_P\left[L^{\Gamma}_{\text{RU}}(\theta(X), Y)\right].$$

Thus, $\widehat{\mathbb{E}}_P\left[L_{\text{RU}}(\theta(X), Y)\right]$ and $\mathbb{E}_P\left[L^{\Gamma}_{\text{RU}}(\theta(X), Y)\right]$ satisfy the conditions of Lemma 16. So, we have that $\hat{\theta}_{m,n}$ exists with probability approaching 1 and $\hat{\theta}_{m,n} \xrightarrow{p} \theta^*_m$.

C.14 Proof of Theorem 15

The main goal of this proof is to show that the following theorem applies to our setting.

Theorem 29 (Chen [2007], Theorem 3.2)**.** *Let Z_i be distributed i.i.d. following a distribution P. Let $\theta^* \in \Theta$ be the population risk minimizer*

$$\theta^* = \underset{\theta \in \Theta}{\text{argmin}}\, \mathbb{E}_P\left[l(\theta, Z_i)\right].$$

Let $\hat{\theta}_n$ be the empirical risk minimizer given by

$$\frac{1}{n}\sum_{i=1}^{n} l(\hat{\theta}_n, Z_i) \leq \inf_{\theta \in \Theta_n} \frac{1}{n}\sum_{i=1}^{n} l(\theta, Z_i) + O_P(\epsilon_n^2).$$

Let $||\cdot||$ be a norm on Θ such that $||\hat{\theta}_n - \theta^*|| = o_P(1)$. Let $\mathcal{F}_n = \{l(\theta, Z_i) - l(\theta^*, Z_i) : ||\theta - \theta^*|| \leq \delta, \theta \in \Theta_n\}$. For some constant $b > 0$, let

$$\delta_n = \inf\left\{\delta \in (0,1) : \frac{1}{\sqrt{n}\delta^2}\int_{b\delta^2}^{\delta} \sqrt{H_{[]}(w^{1+\frac{d}{2p}}, \mathcal{F}_n, ||\cdot||)}dw \leq 1\right\},$$

where $H_{[]}(w, \mathcal{F}_n, ||\cdot||_r)$ is the $L^r(P)$ metric entropy with bracketing of the class $\mathcal{F}_n$.

Assume that the following conditions hold.

1. *In a neighborhood of θ^*, $\mathbb{E}\left[l(\theta, Z_i) - l(\theta^*, Z_i)\right] \asymp ||\theta - \theta^*||^2$.*
2. *There is $C_1 > 0$ s.t. for all small $\epsilon > 0$*

$$\sup_{\theta \in \Theta_n : ||\theta - \theta^*|| \leq \epsilon} \text{Var}\left[l(\theta, Z_i) - l(\theta^*, Z_i)\right] \leq C_1\epsilon^2.$$

3. For any $\delta > 0$, there exists a constant $s \in (0, 2)$ such that

$$\sup_{\theta \in \Theta_n : ||\theta - \theta^*|| \leq \delta} |l(\theta, Z_i) - l(\theta^*, Z_i)| \leq \delta^s U(Z_i)$$

with $\mathbb{E}\left[U(Z_i)^\gamma\right] \leq C_2$ for some $\gamma \geq 2$.

Then $||\hat{\theta}_n - \theta^*|| = O_P(\epsilon_n)$, where

$$\epsilon_n = \max\{\delta_n, \inf_{\theta \in \Theta_n} ||\theta^* - \theta||\}.$$

We will use the following lemmas to show that the conditions of the above theorem are satisfied for our setting.

Lemma 30 (Chen and Shen [1998], Lemma 2)**.** *For $\theta \in \Lambda_c^p(\mathcal{X})$, we have that $||\theta||_\infty \leq 2c^{1 - \frac{2p}{2p+d}} ||\theta||_{L^2(\lambda, \mathcal{X})}^{\frac{2p}{2p+d}}$, where λ is the Lebesgue measure.*

Lemma 31. *Under Assumptions 2, 4 5, 6, for any $h \in \Lambda_c^p(\mathcal{X})$, there exists $\bar{L}(X, Y)$ such that*

$$|L(h(x), y) - L(h_\Gamma^*(x), y))| \leq \bar{L}(x, y) \cdot |h(x) - h_\Gamma^*(x)|,$$

where $\sup_{x \in \mathcal{X}} \mathbb{E}_{P_{Y|X}}\left[\bar{L}(x, Y)^2 \mid X = x\right] \leq M < \infty$. Proof in Appendix D.12.

For the metric, we will use $||\cdot||_{L^2(P_X, \mathcal{X})}$. Since any function $\theta \in \Theta$ only depends on X, $||\cdot||_{L^2(P_X, \mathcal{X})} = ||\cdot||_{L^2(P, \mathcal{X} \times \mathcal{Y})}$. From Theorem 13, Lemma 14, we have that $\hat{\theta}_n \xrightarrow{p} \theta^*$ with respect to the $L^2(P_X, \mathcal{X})$ norm. So, $||\theta^* - \hat{\theta}_n||_{L^2(P_X \mathcal{X})} = o_P(1)$.

First, we note that our observed data (X_i, Y_i) is i.i.d.

We aim to verify the second condition. We note that by Theorems 10 and 11, the population RU risk is strongly convex and smooth in a $||\cdot||_\infty$-ball about the minimizer θ^*. We note that all θ in this $||\cdot||_\infty$-ball about θ^* also must lie in a $||\cdot||_{L^2(P_X, \mathcal{X})}$-ball about θ^*. So, in a $L^2(P_X, \mathcal{X})$-neighborhood of θ^*, we have that

$$\mathbb{E}_P\left[L_{\text{RU}}^\Gamma(\theta(X), Y)\right] - \mathbb{E}_P\left[L_{\text{RU}}^\Gamma(\theta^*(X), Y)\right] \asymp ||\theta - \theta^*||_{L^2(P_X, \mathcal{X})}^2.$$

We aim to verify the third condition. First, we show the following three intermediate results.

$$\mathbb{E}_P\left[(L(h(X), Y) - L(h_\Gamma^*(X), Y))^2\right] \lesssim ||h - h_\Gamma^*||_{L^2(P_X, \mathcal{X})}^2. \tag{79}$$

$$\mathbb{E}_P\left[(\alpha(X) - \alpha_\Gamma^*(X))^2\right] \asymp ||\alpha - \alpha_\Gamma^*||_{L^2(P_X, \mathcal{X})}^2. \tag{80}$$

$$\mathbb{E}_P\left[((L(h(X), Y) - \alpha(X))_+ - (L(h_\Gamma^*(X), Y) - \alpha_\Gamma^*(X))_+)^2\right] \lesssim ||\theta - \theta^*||_{L^2(P_X, \mathcal{X})}^2. \tag{81}$$

(79) can be shown by apply Lemma 31.

$$\begin{aligned}
\mathbb{E}_P\left[(L(h(X), Y) - L(h_\Gamma^*(X), Y))^2\right] &= \mathbb{E}_P\left[\bar{L}(X, Y)^2 \cdot (h(X) - h_\Gamma^*(X))^2\right] \\
&= \mathbb{E}_{P_X}\left[\mathbb{E}_{P_{Y|X}}\left[\bar{L}(X, Y)^2 \cdot (h(X) - h_\Gamma^*(X))^2 \mid X = x\right]\right] \\
&\leq \sup_{x \in \mathcal{X}} \mathbb{E}_{P_{Y|X}}\left[\bar{L}(x, Y)^2 \mid X = x\right] \cdot ||h - h_\Gamma^*||_{L^2(P_X, \mathcal{X})}^2 \\
&\asymp ||h - h_\Gamma^*||_{L^2(P_X, \mathcal{X})}^2.
\end{aligned}$$

(80) is true by definition. So, we proceed to show (81). We use (79), (80).

$$\begin{aligned}
&\mathbb{E}_P\left[((L(h(X), Y) - \alpha(X))_+ - (L(h_\Gamma^*(X), Y) - \alpha_\Gamma^*(X))_+)^2\right] \\
&\leq \mathbb{E}_P\left[((L(h(X), Y) - \alpha(X)) - (L(h_\Gamma^*(X), Y) - \alpha_\Gamma^*(X)))^2\right] \\
&= \mathbb{E}_P\left[((L(h(X), Y) - L(h_\Gamma^*(X), Y)) - (\alpha(X) - \alpha_\Gamma^*(X)))^2\right] \\
&\leq 2\mathbb{E}_P\left[(L(h(X), Y) - L(h_\Gamma^*(X), Y))^2\right] + 2\mathbb{E}_P\left[(\alpha(X) - \alpha_\Gamma^*(X))^2\right] \\
&\lesssim ||h - h_\Gamma^*||_{L^2(P_X, \mathcal{X})}^2 + ||\alpha - \alpha_\Gamma^*||_{L^2(P_X, \mathcal{X})}^2 \\
&= ||\theta - \theta^*||_{L^2(P_X, \mathcal{X})}^2.
\end{aligned}$$

Now, we consider $\theta \in \mathcal{B}_\epsilon$ where

$$\mathcal{B}_\epsilon = \{\theta \in \Theta_n \mid ||\theta - \theta^*||_{L^2(P_X,\mathcal{X})} \leq \epsilon\}.$$

We aim to show that $\mathrm{Var}_P\left[L^\Gamma_{\mathrm{RU}}(\theta(X), Y) - L^\Gamma_{\mathrm{RU}}(\theta^*(X), Y)\right] \lesssim \epsilon^2$ when $||\theta - \theta^*||_{L^2(P_X,\mathcal{X})} \leq \epsilon$.

$$\begin{aligned}
&\mathrm{Var}_P\left[L^\Gamma_{\mathrm{RU}}(\theta(X), Y) - L^\Gamma_{\mathrm{RU}}(\theta^*(X), Y)\right] \\
&\leq \mathbb{E}_P\left[(L^\Gamma_{\mathrm{RU}}(\theta(X), Y) - L^\Gamma_{\mathrm{RU}}(\theta^*(X), Y))^2\right] \\
&\leq 3\mathbb{E}_P\left[(L(h(X), Y) - L(h^*_\Gamma(X), Y))^2\right] + 3\mathbb{E}_P\left[(\alpha(X) - \alpha^*_\Gamma(X))^2\right] \\
&\quad + 3\mathbb{E}_P\left[((L(h(X), Y) - \alpha(X))_+ - (L(h^*_\Gamma(X), Y) - \alpha^*_\Gamma(X))_+)^2\right] \\
&\lesssim \|h - h^*_\Gamma\|^2_{L^2(P_X,\mathcal{X})} + \|\alpha - \alpha^*_\Gamma\|^2_{L^2(P_X,\mathcal{X})} + \|\theta - \theta^*\|^2_{L^2(P_X,\mathcal{X})} \\
&\lesssim \|\theta - \theta^*\|^2_{L^2(P_X,\mathcal{X})}.
\end{aligned}$$

The second line comes from the Cauchy-Schwarz inequality and the second last line comes from (79), (80), and (81). This prove the third condition.

Finally, we verify the fourth condition. We consider $\theta \in \mathcal{B}_\delta$, where

$$\mathcal{B}_\delta = \{\theta \in \Theta_n \mid ||\theta - \theta^*||^2_{L^2(P_X,\mathcal{X})} \leq \delta\}.$$

Using a similar argument as in the previous condition, we apply Lemma 31.

$$\begin{aligned}
|L^\Gamma_{\mathrm{RU}}(\theta(x), y) - L^\Gamma_{\mathrm{RU}}(\theta^*(x), y)| &\lesssim |\bar{L}(x, y) \cdot (h(x) - h^*(x))| + |\alpha(x) - \alpha^*(x)| && (82)\\
&\lesssim |\bar{L}(x, y)| \cdot ||\theta - \theta^*||_\infty && (83)\\
&\lesssim |\bar{L}(x, y)| \cdot ||\theta - \theta^*||^{\frac{2p}{2p+d}}_{L^2(\lambda)} && (84)\\
&\lesssim |\bar{L}(x, y)| \cdot ||\theta - \theta^*||^{\frac{2p}{2p+d}}_{L^2(P_X,\mathcal{X})}. && (85)
\end{aligned}$$

Since Assumption 5 holds, we can apply Lemma 30 to see that for $\theta \in \Theta$, $||\theta||_\infty \lesssim ||\theta||^{\frac{2p}{2p+d}}_{L^2(\lambda)}$, where λ is the Lebesgue measure. This gives (84). Under Assumption 7, $||\theta - \theta'||_{L^2(P_X,\mathcal{X})} \asymp ||\theta - \theta'||_{L^2(\lambda)}$, which gives (85).

Therefore, the fourth condition holds with $s = \frac{2p}{2p+d}$ and $U(X_i, Y_i) = |\bar{L}(X_i, Y_i)|$. So, by Theorem 29, we have that $||\hat{\theta}_n - \theta^*||_{L^2(P_X,\mathcal{X})} = O_P(\max\{\delta_n, \inf_{\theta\in\Theta_n} ||\theta - \theta^*||_{L^2(P_X,\mathcal{X})}\})$.

Let $\mathcal{F}_n = \{L^\Gamma_{\mathrm{RU}}(\theta(X_i), Y_i) - L^\Gamma_{\mathrm{RU}}(\theta^*(X_i), Y_i) : ||\theta - \theta^*||_{L^2(P_X,\mathcal{X})} \leq \delta, \theta \in \Theta_n\}$. Let $H_{[]}(w, \mathcal{F}_n, ||\cdot||_{L^2(P_X,\mathcal{X})})$ be the $L^2(P_X, \mathcal{X})$-metric entropy with bracketing of the class $\mathcal{F}_n$.

Since in our setting, we satisfy the fourth condition of Theorem 29 with $s = \frac{2p}{2p+d}$,

$$H_{[]}(w, \mathcal{F}_n, ||\cdot||_2) \leq \log N(w^{1+\frac{d}{2p}}, \Theta_n, ||\cdot||_{L^2(P_X,\mathcal{X})}).$$

Recall that $\tilde{\Theta}_n$ is the sieve space *without truncation*. We note that the covering number of Θ_n is upper bounded by the covering number of $\tilde{\Theta}_n$, so we have that

$$H_{[]}(w, \mathcal{F}_n, ||\cdot||_2) \leq \log N(w^{1+\frac{d}{2p}}, \tilde{\Theta}_n, ||\cdot||_{L^2(P_X,\mathcal{X})}).$$

For the finite-dimensional linear sieves, such as those in Example 2 and 3 without truncation, we have that

$$\log N(w^{1+\frac{d}{2p}}, \tilde{\Theta}_n, ||\cdot||_{L^2(P_X,\mathcal{X})}) \lesssim \dim(\tilde{\Theta}_n) \log(\frac{1}{w})$$

from Van de Geer and van de Geer [2000]. Then, we have that

$$\frac{1}{\sqrt{n}\delta^2}\int_{b\delta^2}^{\delta} \sqrt{\log N(w^{1+\frac{d}{2p}}, \tilde{\Theta}_n, ||\cdot||_{L^2(P_X,\mathcal{X})})}dw \lesssim \frac{1}{\delta}\sqrt{\frac{\dim(\tilde{\Theta}_n)}{n}\log\frac{1}{\delta}}.$$

We realize that

$$\delta_n \asymp \sqrt{\frac{\dim(\tilde{\Theta}_n)\log n}{n}}.$$

We note that $\tilde{\Theta}_n = \tilde{\mathcal{H}}_n \times \tilde{\mathcal{A}}_n$. We have that $\dim(\tilde{\Theta}_n) = 2J_n^d = O(J_n^d)$. Plugging this in, we have that

$$\delta_n \asymp \sqrt{\frac{J_n^d \log n}{n}}.$$

Now, we can bound the approximation error $\inf_{\theta \in \Theta_n} ||\theta^* - \theta||_{L^2(P_X,\mathcal{X})}$. Since the truncation of the sieve space is a contraction map to the true minimizer, we have that

$$\inf_{\theta \in \Theta_n} ||\theta^* - \theta||_{L^2(P_X,\mathcal{X})} \leq \inf_{\theta \in \tilde{\Theta}_n} ||\theta^* - \theta||_\infty \leq O(J_n^{-p}),$$

where the last inequality follows from Timan [2014]. So, we can set $J_n = (\frac{n}{\log n})^{\frac{1}{2p+d}}$. Thus, we have that $||\hat{\theta} - \theta^*||_{L^2(P_X,\mathcal{X})} = O_P\Big(\Big(\frac{\log n}{n}\Big)^{\frac{p}{2p+d}}\Big)$.

Finally, we can also show that for $Q_X \ll P_X$, if $\sup_{x \in \mathcal{X}} \frac{dQ_X(x)}{dP_X(x)} < C$ for some $C < \infty$, then the same rate of convergence holds. We have that

$$\begin{aligned}
||\hat{\theta}_n - \theta^*||_{L^2(Q_X,\mathcal{X})} &= \Big(\mathbb{E}_{Q_X}\left[(\hat{\theta}_n(x) - \theta^*(x))^2\right]\Big)^{\frac{1}{2}} \\
&= \Big(\mathbb{E}_{P_X}\left[(\hat{\theta}_n(x) - \theta^*(x))^2 \cdot \frac{dP_X(x)}{dQ_X(x)}\right]\Big)^{\frac{1}{2}} \\
&\leq \Big(\mathbb{E}_{P_X}\left[(\hat{\theta}_n(x) - \theta^*(x))^2\right]\Big)^{\frac{1}{2}} \cdot \Big(\sup_{x \in \mathcal{X}} \Big|\frac{dP_X(x)}{dQ_X(x)}\Big|\Big)^{\frac{1}{2}} \\
&= ||\hat{\theta}_n - \theta^*||_{L^2(Q_X,\mathcal{X})} \cdot \sqrt{C} \\
&= O_P\Big(\Big(\frac{\log n}{n}\Big)^{\frac{p}{2p+d}}\Big).
\end{aligned}$$

D Proofs of Technical Lemmas

D.1 Proof of Lemma 3

Let $h^* \in L^2(Q_X, \mathcal{X})$ be the solution to (7). Let the function $\tilde{h}$ be minimizer of (13) at every x. Since $\tilde{h}$ solves (13) for every $x \in \text{supp}(Q_X)$,

$$\sup_{Q \in S_\Gamma(P,Q_X)} \mathbb{E}_{Q_{Y|X}}\left[L(\tilde{h}(X), Y) \mid X = x\right] \leq \sup_{Q \in S_\Gamma(P,Q_X)} \mathbb{E}_{Q_{Y|X}}\left[L(h^*(X), Y) \mid X = x\right].$$

Given any marginal distribution Q_X, we can marginalize over X to see that

$$\mathbb{E}_{Q_X}\left[\sup_{Q_{Y|X}:Q \in S_\Gamma(P)} \mathbb{E}_{Q_{Y|X}}\left[L(\tilde{h}(X), Y) \mid X\right]\right] \leq \mathbb{E}_{Q_X}\left[\sup_{Q_{Y|X}:Q \in S_\Gamma(P)} \mathbb{E}_{Q_{Y|X}}\left[L(h^*(X), Y) \mid X = x\right]\right].$$

Based on our definition of $S_\Gamma(P, Q_X)$, we note that for any $h \in L^2(Q_X, \mathcal{X})$

$$\sup_{Q \in S_\Gamma(P,Q_X)} \mathbb{E}_Q\left[L(h(X), Y)\right] = \mathbb{E}_{Q_X}\left[\sup_{Q_{Y|X}:Q \in S_\Gamma(P,Q_X)} \mathbb{E}_{Q_{Y|X}}\left[L(h(X), Y) \mid X\right]\right].$$

Thus, we have that

$$\sup_{Q \in S_\Gamma(P,Q_X)} \mathbb{E}_Q\left[L(\tilde{h}(X), Y)\right] \leq \sup_{Q \in S_\Gamma(P,Q_X)} \mathbb{E}_Q\left[L(h^*(X), Y)\right].$$

Finally, by definition of h^* we must also have that

$$\sup_{Q \in S_\Gamma(P,Q_X)} \mathbb{E}_Q\left[L(h^*(X), Y)\right] \leq \sup_{Q \in S_\Gamma(P,Q_X)} \mathbb{E}_Q\left[L(\tilde{h}(X), Y)\right].$$

These last two inequalities yield the desired equivalence.

D.2 Proof of Lemma 4

Let $\mathcal{T}$ be any set with nonzero measure with respect to Q_X. Then $\mathcal{T}$ must also have nonzero measure with respect to P_X because $Q_X \ll P_X$. Since $\tilde{h}$ solves (13) for every $x \in \text{supp}(P_X)$, we have that

$$\sup_{Q_{Y|X}:Q\in S_\Gamma(P,Q_X)} \mathbb{E}_{Q_{Y|X}}\left[L(\tilde{h}(X),Y) \mid X \in \mathcal{T}\right] \leq \sup_{Q_{Y|X}:Q\in S_\Gamma(P,Q_X)} \mathbb{E}_{Q_{Y|X}}\left[L(h(X),Y) \mid X \in \mathcal{T}\right]$$

for any $h \in L^2(Q_X, \mathcal{X})$ and any set $\mathcal{T}$ with nonzero measure with respect to Q_X. This is sufficient to show that h^*_Γ is a solution to (7) for any $Q_X \ll P_X$.

D.3 Proof of Lemma 20

We note that the eigenvalues of a 2x2 matrix must satisfy

$$\lambda^2 - (\operatorname{tr} A)\lambda + \det A = 0.$$

Since $\operatorname{tr} A \geq 0$ and $\det A \geq 0$, the minimum eigenvalue is given by So,

$$\lambda_{\min}(A) = \frac{\operatorname{tr} A - \sqrt{(\operatorname{tr} A)^2 - 4\det A}}{2}.$$

Let $x = \operatorname{tr} A$ and $y = \sqrt{(\operatorname{tr} A)^2 - 4\det A}$. Note that $y \leq x$ because $\det A \geq 0$. Then we have that

$$\lambda_{\min}(A) = \frac{x-y}{2} = \frac{x^2-y^2}{2(x+y)} \geq \frac{x^2-y^2}{2(x+x)} = \frac{x^2-y^2}{4x} = \frac{\det A}{\operatorname{tr} A}.$$

In addition, we have that

$$\lambda_{\max}(A) = \frac{x+y}{2} \leq \frac{x+x}{2} = x = \operatorname{tr} A.$$

D.4 Proof of Lemma 21

Since H is Gâteaux differentiable with derivative equal to $H'_{h,\alpha}$, we aim to show that, for any $h, \tilde{h}, \alpha, \tilde{\alpha}$,

$$\langle H'_{h,\alpha}(h,\alpha) - H'_{\tilde{h},\tilde{\alpha}}(0,0), (h-\tilde{h}, \alpha-\tilde{\alpha})\rangle > 0$$

to establish strict convexity. Without loss of generality, we assume that $(\tilde{h},\tilde{\alpha}) = (0,0)$. Define the Gâteaux derivative of F and G in (h,α) to be $F'_{h,\alpha}, G'_{h,\alpha}$, respectively. Let the Gâteaux derivative of G with respect to h be G'_h. We have that

$$\begin{aligned}\langle H'_{h,\alpha}(h,\alpha) - H'_{h,\alpha}(0,0), (h,\alpha)\rangle &= \langle F'_{h,\alpha}(h,\alpha) + G'_{h,\alpha}(h,\alpha) - F'_{h,\alpha}(0,0) - G'_{h,\alpha}(0,0), (h,\alpha)\rangle \\ &= \langle F'_{h,\alpha}(h,\alpha) - F'_{h,\alpha}(0,0), (h,\alpha)\rangle + \langle G'_h(h,\alpha) - G'_h(0,0), h\rangle.\end{aligned}$$

Note that G does not depend on α, so the Gâteaux derivative is $G'_\alpha(h,\alpha) = G'_\alpha(0,0) = 0$. Since F is jointly convex in (h,α) and G is strongly convex in h and does not depend on α, both terms above are nonnegative.

If $h \neq 0$, then we have that

$$\begin{aligned}\langle H'_{h,\alpha}(h,\alpha) - H'_{h,\alpha}(0,0), (h,\alpha)\rangle &\geq \langle G'_h(h,\alpha) - G'_h(0,0), h) \\ &\geq \mu_1 ||h||^2 > 0,\end{aligned}$$

where the last line follows follows from G's strong convexity in h. If $h = 0$ and $\alpha \neq 0$, then we have that

$$\begin{aligned}\langle H'_{h,\alpha}(h,\alpha) - H'_{h,\alpha}(0,0), (h,\alpha)\rangle &= \langle H'_{h,\alpha}(0,\alpha) - H'_{h,\alpha}(0,0), (0,\alpha)\rangle \\ &= \langle F'_{h,\alpha}(0,\alpha) - F'_{h,\alpha}(0,0), (0,\alpha)\rangle \\ &> 0,\end{aligned}$$

where the last inequality follows due to the strict convexity of F in α. Thus, H is strictly convex in (h,α).

D.5 Proof of Lemma 22

We have that $T^c_{1,x}(c) = -\mathbb{E}_{P_{Y|X=x}}\left[\ell'(Y-c)\right]$. In addition, $T^{cc}_{1,x}(c) = \mathbb{E}_{P_{Y|X=x}}\left[\ell''(Y-c)\right]$. So, $T_{1,x}$ is twice differentiable in c. In addition, we realize that

$$\begin{aligned}\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU},1}(h(X),Y)\right] &= \mathbb{E}_{P_X}\left[\mathbb{E}_{P_{Y|X=x}}\left[L^\Gamma_{\mathrm{RU},1}(h(X),Y)\right]\right] \\ &= \Gamma^{-1}\mathbb{E}_{P_X}\left[T_{X,1}(h(X))\right].\end{aligned}$$

D.6 Proof of Lemma 23

First, we compute the first derivatives of $T_{3,x}(c,d)$. Second, we compute the second derivatives of $T_{3,x}(c,d)$ when $d>0$. Finally, we show that $T_{3,x}$ can be used to express $\mathbb{E}_P\left[L^\Gamma_{\mathrm{RU},3}(h(X),\alpha(X),Y)\right]$.

Computing derivatives for the $d\le 0$ case is straightforward. When $d\le 0$, we have that

$$\begin{aligned}T^c_{3,x}(c,d) &= -\mathbb{E}_{P_{Y|X}}\left[\ell'(Y-c)\mid X=x\right] \\ T^d_{3,x}(c,d) &= -1.\end{aligned}$$

Now, we consider the $d>0$ case. To compute the derivatives of

$$T_{3,x}(c,d) = \mathbb{E}_{P_{Y|X}}\left[(\ell(Y-c)-d)\mathbb{I}(\ell(Y-c)>d)\mid X=x\right],$$

we first identify when the condition $\ell(Y-c)>d$ is satisfied. The strong convexity of ℓ given by Assumption 2 implies that ℓ is strictly increasing on $y>0$ and ℓ is strictly decreasing on $y<0$. We define ℓ_1^{-1} to be the inverse of $\ell(y)$ on $y>0$. We define ℓ_2^{-1} to be the inverse of $\ell(y)$ on $y<0$. By the Inverse Function Theorem, we have that

$$(\ell_i^{-1})'(z) = \frac{1}{\ell'(\ell_i^{-1}(z))} \quad i=1,2. \tag{86}$$

We note that $\ell_1^{-1}(z)>0$, and $\ell(y)$ strictly increasing on $y>0$, so $\ell'(\ell_1^{-1}(z))>0$. By (86), we have that $(\ell_1^{-1})'(z)>0$. This means that ℓ_1^{-1} is strictly increasing on its domain. By an analogous argument, we have that $(\ell_2^{-1})'(z)<0$ and ℓ_2^{-1} is strictly decreasing on its domain.

Based on the results above, we realize that for $d>0$,

$$\{y\in\mathbb{R}\mid \ell(y-c)>d\} = \{y\in\mathbb{R}\mid y-c>\ell_1^{-1}(d)\}\cup\{y\in\mathbb{R}\mid y-c<\ell_2^{-1}(d)\}.$$

Thus, we can rewrite $T_{3,x}(c,d)$ for $d>0$ as follows

$$T_{3,x}(c,d) = \mathbb{E}_{Y|X=x}\left[(\ell(Y-c)-d)\mathbb{I}(Y-c>\ell_1^{-1}(d))\right] + \mathbb{E}_{Y|X=x}\left[(\ell(Y-c)-d)\mathbb{I}(Y-c<\ell_2^{-1}(d))\right].$$

Now, we can compute the derivatives of $T_{3,x}(c,d)$ on $d>0$ as follows.

$$\begin{aligned}T^d_{3,x}(c,d) &= \mathbb{E}_{P_{Y|X}}\left[-1\cdot\mathbb{I}(\ell(Y-c)>d)\mid X=x\right] \\ &= \mathbb{E}_{P_{Y|X=x}}\left[-1\cdot\mathbb{I}(Y-c>\ell_1^{-1}(d))\right] + \mathbb{E}_{P_{Y|X}}\left[-1\cdot\mathbb{I}(Y-c<\ell_2^{-1}(d))\mid X=x\right] \\ &= -\Pr(Y>c+\ell_1^{-1}(d)\mid X=x) - \Pr(Y<c+\ell_2^{-1}(d)\mid X=x) \\ &= -1 + P_{Y|X=x}(c+\ell_1^{-1}(d))) - P_{Y|X=x}(c+\ell_2^{-1}(d)).\end{aligned}$$

Another way to express $T^d_{3,x} = -\Pr(\ell(Y-c)>d|X=x)$.

We realize that

$$\lim_{d\to 0^+} T^d_{3,x}(c,d) = -1 + P_{Y|X=x}(-c) - P_{Y|X=x}(-c) = -1 = \lim_{d\to 0^-} T^d_{3,x}(c,d),$$

so $T_{3,x}(c,d)$ is differentiable at $d=0$. Also,

$$\begin{aligned}T^c_{3,x}(c,d) &= -\mathbb{E}_{P_{Y|X}}\left[\ell'(Y-c)\mathbb{I}(\ell(Y-c)>d)\mid X=x\right] \\ &= -\mathbb{E}_{P_{Y|X}}\left[\ell'(Y-c)\cdot\mathbb{I}(Y-c>\ell_1^{-1}(d))\right] - \mathbb{E}_{Y|X=x}\left[\ell'(Y-c)\cdot\mathbb{I}(Y-c<\ell_2^{-1}(d))\mid X=x\right]\end{aligned}$$

We realize that

$$\lim_{d\to 0^+} T^c_{3,x}(c,d) = -\mathbb{E}_{P_{Y|X}}\left[\ell'(Y-c) \mid X = x\right] = \lim_{d\to 0^-} T^c_{3,x}(c,d),$$

so $T_{3,x}(c,d)$ is differentiable with respect to c.

Second, we compute the second derivatives of $T_{3,x}(c,d)$ when $d > 0$. It is straightforward to see that

$$T^{dc}_{3,x}(c,d) = p_{Y|X=x}(c+\ell_1^{-1}(d)) - p_{Y|X=x}(c+\ell_2^{-1}(d)).$$

In addition, we have that

$$\begin{aligned} T^{dd}_{3,x}(c,d) &= p_{Y|X=x}(c+\ell_1^{-1}(d)) \cdot \frac{1}{\ell'(\ell_1^{-1}(d))} - p_{Y|X=x}(c+\ell_2^{-1}(d)) \cdot \frac{1}{\ell'(\ell_2^{-1}(d))} \\ &= \sum_{i\in\{1,2\}} p_{Y|X=x}(c+\ell_i^{-1}(d)) \cdot \frac{1}{|\ell'(\ell_i^{-1}(d))|}. \end{aligned}$$

The second line follows because $\ell'(\ell_2^{-1}(y)) < 0$. Finally, we compute $T^{cc}_{3,x}(c,d)$. First, we recall $T^c_{3,x}(c,d)$ from Lemma 23 and simplify it as follows.

$$\begin{aligned} T^c_{3,x}(c,d) &= -\mathbb{E}_{P_{Y|X}}\left[\ell'(Y-c)\mathbb{I}(\ell(Y-c) > d) \mid X = x\right] \\ &= -\mathbb{E}_{P_{Y|X}}\left[\ell'(Y-c)\mathbb{I}(Y > \ell_1^{-1}(d)+c) \mid X = x\right] - \mathbb{E}_{P_{Y|X}}\left[\ell'(Y-c)\mathbb{I}(Y < \ell_2^{-1}(d)+c) \mid X = x\right] \\ &= -\int_{\ell_1^{-1}(d)+c}^{\infty} \ell'(y-c)p_{Y|X=x}(y)dy - \int_{-\infty}^{\ell_2^{-1}(d)+c} \ell'(y-c)p_{Y|X=x}(y)dy \\ &= -\int_{\ell_1^{-1}(d)}^{\infty} \ell'(y)p_{Y|X=x}(y+c)dy - \int_{-\infty}^{\ell_2^{-1}(d)} \ell'(y)p_{Y|X=x}(y+c)dy. \end{aligned}$$

Now, we compute $T^{cc}_{3,x}(c,d)$ by differentiating with respect to c and applying integration by parts.

$$\begin{aligned} T^{cc}_{3,x}(c,d) &= -\int_{\ell_1^{-1}(d)}^{\infty} \ell'(y)p'_{Y|X=x}(y+c)dy - \int_{-\infty}^{\ell_2^{-1}(d)} \ell'(y)p'_{Y|X=x}(y+c)dy \\ &= -\Big(\ell'(y)p_{Y|X=x}(y+c)\Big|_{\ell_1^{-1}(d)}^{\infty} - \int_{\ell_1^{-1}(d)}^{\infty} p_{Y|X=x}(y+c)\ell''(y)dy\Big) \\ &\quad - \Big(\ell'(y)p_{Y|X=x}(y+c)\Big|_{-\infty}^{\ell_2^{-1}(d)} - \int_{-\infty}^{\ell_2^{-1}(d)} p_{Y|X=x}(y+c)\ell''(y)dy\Big) \\ &= \ell'(\ell_1^{-1}(d))p_{Y|X=x}(\ell_1^{-1}(d)+c) + \int_{\ell_1^{-1}(d)}^{\infty} p_{Y|X=x}(y+c)\ell''(y)dy \\ &\quad - \ell'(\ell_2^{-1}(d))p_{Y|X=x}(\ell_2^{-1}(d)+c) + \int_{-\infty}^{\ell_2^{-1}(d)} p_{Y|X=x}(y+c)\ell''(y)dy \\ &= \ell'(\ell_1^{-1}(d))p_{Y|X=x}(\ell_1^{-1}(d)+c) + \int_{c+\ell_1^{-1}(d)}^{\infty} p_{Y|X=x}(y)\ell''(y-c)dy \\ &\quad - \ell'(\ell_2^{-1}(d))p_{Y|X=x}(\ell_2^{-1}(d)+c) + \int_{-\infty}^{c+\ell_2^{-1}(d)} p_{Y|X=x}(y)\ell''(y-c)dy \\ &= \sum_{i\in\{1,2\}} |\ell'(\ell_i^{-1}(d))| \cdot p_{Y|X=x}(\ell_i^{-1}(d)+c) + \mathbb{E}_{P_{Y|X}}\left[\ell''(Y-c)\mathbb{I}(\ell(Y-c) > d) \mid X = x\right]. \end{aligned}$$

Thus, when $d > 0$, $T_{3,x}(c,d)$ is twice differentiable in (c,d).

Lastly, we find that

$$\mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},3}(h(X),\alpha(X),Y)\right]$$

$$
\begin{aligned}
&= (\Gamma - \Gamma^{-1}) \cdot \mathbb{E}_P \left[(\ell(Y - h(X)) - \alpha(X))_+ \right] \\
&= (\Gamma - \Gamma^{-1}) \cdot \mathbb{E}_{P_X} \left[\mathbb{E}_{P_{Y|X}} \left[\ell(Y - h(X)) - \alpha(X) \right)_+ \mid X \right] \\
&= (\Gamma - \Gamma^{-1}) \cdot \mathbb{E}_{P_X} \begin{bmatrix} \begin{cases} \mathbb{E}_{P_{Y|X}} \left[(\ell(Y - h(X)) - \alpha(X)) \mathbb{I}(\ell(Y - h(X)) > \alpha(X)) \right] & \alpha(X) > 0 \\ \mathbb{E}_{P_{Y|X}} \left[(\ell(Y - h(X)) - \alpha(X)) \right] & \alpha(X) \leq 0 \end{cases} \end{bmatrix} \\
&= (\Gamma - \Gamma^{-1}) \mathbb{E}_{P_X} \left[T_{3,X}(h(X), \alpha(X)) \right].
\end{aligned}
$$

D.7 Proof of Lemma 24

Now, define a symmetric 2×2 matrix $A_x(c, d)$ where

$$
\begin{aligned}
A_{x,11}(c, d) &= T^{cc}_{3,x}(c, d) - \mathbb{E}_{Y|X=x} \left[\ell''(Y - c) \mathbb{I}(\ell(Y - c) > d) \right] + C_{L,l} \cdot \Pr(\ell(Y - c) > d \mid X = x) \\
A_{x,22}(c, d) &= T^{dd}_{3,x}(c, d) \\
A_{x,12}(c, d) &= T^{dc}_{3,x}(c, d),
\end{aligned}
$$

where F is the distribution over $\ell(Y - c)$ where Y follows $P_{Y|X=x}$. Under Assumption 2, we have that ℓ is $C_{\ell,l}$-strongly convex, so

$$
\mathbb{E}_{Y|X=x} \left[\ell''(Y - c) \mathbb{I}(\ell(Y - c) > d) \right] - C_{L,l} \cdot \Pr(\ell(Y - c) > d | X = x) \geq 0.
$$

Thus, we have that

$$
\begin{aligned}
\nabla^2 T_{3,x}(c, d) - A_x(c, d) &= \begin{bmatrix} \mathbb{E}_{Y|X=x} \left[(\ell''(Y - c) - C_{L,l}) \mathbb{I}(\ell(Y - c) > d) \right] & 0 \\ 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} \mathbb{E}_{Y|X=x} \left[(\ell''(Y - c) \mathbb{I}(\ell(Y - c) > d) \right] - C_{L,l} \Pr(\ell(Y - c) > d \mid X = x) & 0 \\ 0 & 0 \end{bmatrix} \\
&\succeq 0.
\end{aligned}
$$

So,

$$
\nabla^2 T_{3,x}(c, d) \succeq A_x(c, d). \tag{87}
$$

We can also define a symmetric 2×2 matrix $B_x(c, d)$ where

$$
\begin{aligned}
B_{x,11}(c, d) &= T^{cc}_{3,x}(c, d) + \mathbb{E}_{Y|X=x} \left[\ell''(Y - c) \mathbb{I}(\ell(Y - c) \leq d) \right] \\
B_{x,22}(c, d) &= T^{dd}_{3,x}(c, d) \\
B_{x,12}(c, d) &= T^{dc}_{3,x}(c, d).
\end{aligned}
$$

Under Assumption 2, we have that ℓ is strongly convex, so

$$
\mathbb{E}_{Y|X=x} \left[\ell''(Y - c) \mathbb{I}(\ell(Y - c) \leq d) \right] \geq C_{\ell,l} \cdot \Pr(\ell(Y - c) \leq d | X = x) \geq 0.
$$

Thus, we have that

$$
B_x(c, d) - \nabla^2 T_{3,x}(c, d) = \begin{bmatrix} \mathbb{E}_{Y|X=x} \left[\ell''(Y - c) \mathbb{I}(\ell(Y - c) \leq d) \right] & 0 \\ 0 & 0 \end{bmatrix} \succeq 0.
$$

So,

$$
\nabla^2 T_{3,x}(c, d) \preceq B_x(c, d). \tag{88}
$$

Combining (87) and (88) yields the desired result.

D.8 Proof of Lemma 25

Let $L(h,\alpha) = \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU}}(h(X),\alpha(X),Y)\right]$. First, we verify Gâteaux differentiability with respect to α. We show that the directional derivative of $L(h,\alpha)$ with respect to α in the direction ϕ exists for all $\phi \in L^2(P_X,\mathcal{X})$. We note that the directional derivative with respect to α in the direction ϕ is given by

$$L'_\alpha(h,\alpha;\phi) = \lim_{\theta\to 0^+} \frac{L(h,\alpha+\theta\phi) - L(h,\alpha)}{\theta}.$$

We simplify the numerator as follows

$$\begin{aligned}
&L(h,\alpha+\theta\phi) - L(h,\alpha)\\
&= \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},2}((\alpha+\theta\phi)(X))\right] + \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},3}(h(X),(\alpha+\theta\phi)(X),Y)\right]\\
&\quad - \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},2}(\alpha(X))\right] - \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},3}(h(X),\alpha(X),Y)\right]\\
&= \theta(1-\Gamma^{-1})\cdot\mathbb{E}_{P_X}\left[\phi(X)\right] + (\Gamma-\Gamma^{-1})\cdot\mathbb{E}_{P_X}\left[T_{3,X}(h(X),(\alpha+\theta\phi)(X)) - T_{3,X}(h(X),\alpha(X))\right].
\end{aligned}$$

The first line follows because only the second and third term of the RU loss depend on α. The second line follows by Lemma 23. We analyze the second term on the right side of the above equation. We note that by Lemma 23, the map $T_{3,x}(c,d)$ is differentiable with respect to c,d. So,

$$\lim_{\theta\to 0^+} \frac{T_{3,x}(h(x),\alpha(x)+\theta\phi(x)) - T_{3,x}(h(x),\alpha(x))}{\theta} = T^d_{3,x}(h(x),\alpha(x))\phi(x).$$

Therefore, we have that

$$\begin{aligned}
L'_\alpha(h,\alpha;\phi) &= \lim_{\theta\to 0^+}(1-\Gamma^{-1})\cdot\frac{\theta\mathbb{E}_{P_X}\left[\phi(X)\right]}{\theta}\\
&\quad + \lim_{\theta\to 0^+}(\Gamma-\Gamma^{-1})\cdot\frac{\mathbb{E}_{P_X}\left[T_{3,X}(h(X),\alpha(x)+\theta\phi(X)) - T_{3,X}(h(X),\alpha(X))\right]}{\theta}\\
&= (1-\Gamma^{-1})\cdot\mathbb{E}_{P_X}\left[\phi(X)\right] + (\Gamma-\Gamma^{-1})\cdot\mathbb{E}_{P_X}\left[T^d_{3,X}(h(X),\alpha(X))\phi(X)\right]\\
&= \mathbb{E}_{P_X}\left[((1-\Gamma^{-1}) + (\Gamma-\Gamma^{-1})\cdot T^d_{3,X}(h(X),\alpha(X))\cdot\phi(X)\right].
\end{aligned}$$

Since the directional derivative of $L(h,\alpha)$ with respect to α and in the direction ϕ exists for all $\phi\in\mathcal{A}$, then $L(h,\alpha)$ is Gâteaux differentiable in α.

We use a similar technique to verify Gâteaux differentiability with respect to h. We show that the directional derivative of $L(h,\alpha)$ with respect to h in the direction ψ exists for $\psi\in\mathcal{H}$. We recall that the directional derivative of $L(h,\alpha)$ with respect to h in the direction ψ is given by

$$L'_h(h,\alpha;\psi) = \lim_{\theta\to 0^+}\frac{L(h+\theta\psi,\alpha) - L(h,\alpha)}{\theta}. \tag{89}$$

We simplify the directional derivative in (89) as follows.

$$\begin{aligned}
L'_h(h,\alpha;\psi) &= \lim_{\theta\to 0^+}\frac{L(h+\theta\psi,\alpha) - L(h,\alpha)}{\theta}\\
&= \frac{\mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},1}((h+\theta\psi)(X),Y) - L^{\Gamma}_{\mathrm{RU},1}(h(X),Y)\right]}{\theta}\\
&\quad + \lim_{\theta\to 0^+}\frac{\mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},3}((h+\theta\psi)(X),\alpha(X),Y) - L^{\Gamma}_{\mathrm{RU},3}(h(X),\alpha(X),Y)\right]}{\theta}\\
&= \lim_{\theta\to 0^+}\Gamma^{-1}\cdot\frac{\mathbb{E}_{P_X}\left[T_{1,X}(h(X)+\theta\psi(X)) - T_{1,X}(h(X))\right]}{\theta}\\
&\quad + \lim_{\theta\to 0^+}(\Gamma-\Gamma^{-1})\cdot\frac{\mathbb{E}_{P_X}\left[T_{3,X}(h(X)+\theta\psi(X),\alpha(X)) - T_{3,X}(h(X),\alpha(X))\right]}{\theta}\\
&= \mathbb{E}_{P_X}\left[(\Gamma^{-1}\cdot T^c_{1,X}(h(X)) + (\Gamma-\Gamma^{-1})\cdot T^c_{3,X}(h(X),\alpha(X))\cdot\psi(X)\right].
\end{aligned}$$

The first line follows because only the first and third terms of the RU loss depend on h. The second line follows because of Lemma 22 and Lemma 23. The third line follows from the differentiability of $T_{1,x}, T_{3,x}$, which is given by Lemmas 22 and 23. Since the directional derivative of $L(h, \alpha)$ with respect to h and in the direction ψ exists for all $\psi \in L^2(P_X, \mathcal{X})$, and the directional derivative can be expressed as a continuous linear function (given the inner product on $L^2(P_X, \mathcal{X})$), then $L(h, \alpha)$ is Gâteaux differentiable in h.

We can compute second derivatives of $L(h, \alpha)$ on $\mathcal{C}$ by applying Lemma 22 and Lemma 23. Note that $T_{3,x}$ is twice-differentiable when $d > 0$. For $(h, \alpha) \in \mathcal{C}$, we have that $\alpha(x) \geq 0$. We note that the restriction of $\mathcal{C}$ to the coordinate that corresponds to h is $L^2(P_X, \mathcal{X})$. Let $\mathcal{A}'$ be the resitrction of $\mathcal{C}$ to the coordinate that corresponds to α. In the following result, we consider $\psi_1, \psi_2 \in L^2(P_X, \mathcal{X})$ and $\phi_1, \phi_2 \in \mathcal{A}'$. We find that

$$\begin{aligned}
L''_{hh}(h, \alpha; \psi_1, \psi_2) &= L''_{1,hh}(h, \alpha; \psi_1, \psi_2) + L''_{3,hh}(h, \alpha; \psi_1, \psi_2) \\
&= \Gamma^{-1}\mathbb{E}_{P_X}\left[T^{cc}_{1,X}(h(X))\psi_1(X)\psi_2(X)\right] + (\Gamma - \Gamma^{-1})\mathbb{E}_{P_X}\left[T^{cc}_{3,X}(h(X), \alpha(X))\psi_1(X)\psi_2(X)\right]. \\
L''_{h\alpha}(h, \alpha; \psi_1, \phi_1) &= L''_{3,h\alpha}(h, \alpha; \psi_1, \phi_1) \\
&= (\Gamma - \Gamma^{-1})\mathbb{E}_{P_X}\left[T^{cd}_{3,X}(h(X), \alpha(X))\psi_1(X)\phi_1(X)\right]. \\
L''_{\alpha\alpha}(h, \alpha; \phi_1, \phi_2) &= L''_{3,\alpha\alpha}(h, \alpha; \phi_1, \phi_2) \\
&= (\Gamma - \Gamma^{-1})\mathbb{E}_{P_X}\left[T^{dd}_{3,X}(h(X), \alpha(X))\phi_1(X)\phi_2(X)\right].
\end{aligned}$$

D.9 Proof of Lemma 26

Define $L_1(h, \alpha) = \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},1}(h(X), Y)\right]$. From Lemma 25, we have that $L_1(h, \alpha)$ is twice Gâteaux differentiable in h with

$$\begin{aligned}
L''_{1,h}(h, \alpha; \psi, \psi) &= \Gamma^{-1}\mathbb{E}_{P_X}\left[T^{cc}_{1,X}(h(X)) \cdot (\psi(X))^2\right]. \\
&\geq \Gamma^{-1} \cdot C_{L,l}||\psi||^2_{L^2(\mathcal{X}, P_X)}
\end{aligned}$$

for $\psi \in L^2(P_X, \mathcal{X})$. The last line follows from Assumption 2, where we assume that ℓ is strongly convex. Thus, we have that $\mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},1}(h(X), Y)\right]$ is $\Gamma^{-1} \cdot C_{L,l}$-strongly convex in h.

D.10 Proof of Lemma 27

Let $L_3(h, \alpha) = \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},3}(h(X), \alpha(X), Y)\right]$. By Lemma 25, we have that $\mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},3}(h(X), \alpha(X), Y)\right]$ is Gâteaux differentiable in α with

$$L'_{3,\alpha}(h, \alpha; \phi) = (\Gamma - \Gamma^{-1}) \cdot \mathbb{E}_X\left[T^{d}_{3,X}(h(X), \alpha(X)) \cdot \phi(X)\right].$$

We aim to verify the strict convexity of $\alpha \mapsto \mathbb{E}_P\left[L^{\Gamma}_{\mathrm{RU},3}(h(X), \alpha(X), Y)\right]$ via Lemma 27. So, we must show that for $\alpha_1, \alpha_2 \in \mathcal{A}$ that differ on a set of positive measure, we have that

$$\mathbb{E}_{P_X}\left[(T^{d}_{3,X}(h(X), \alpha_1(X)) - T^{d}_{3,X}(h(X), \alpha_2(X)) \cdot (\alpha_1(X) - \alpha_2(X))\right] > 0.$$

From Lemma 23, we have that

$$T^{d}_{3,x}(h(x), \alpha(x)) = \begin{cases} t^{d}_{3,x}(h(x), \alpha(x)) & \alpha(x) > 0 \\ -1 & \alpha(x) \leq 0 \end{cases},$$

where

$$t^{d}_{3,x}(h(x), \alpha(x)) = -1 + P_{Y|X=x}(h(x) + \ell_1^{-1}(\alpha(x))) - P_{Y|X=x}(h(x) + \ell_2^{-1}(\alpha(x))).$$

By the definition of ℓ_1^{-1}, ℓ_2^{-1} from Lemma 23, we have that

$$\ell_1^{-1}(\alpha(x)) > \ell_2^{-1}(\alpha(x)).$$

Under Assumption 3, we have that $P_{Y|X=x}$ is strictly increasing, so

$$P_{Y|X=x}(h(x) + \ell_1^{-1}(\alpha(x))) - P_{Y|X=x}(h(x) + \ell_2^{-1}(\alpha(x))) > 0,$$

which implies that

$$t^d_{3,x}(h(x), \alpha(x)) > -1. \tag{90}$$

Under Assumption 2, ℓ_1^{-1} is strictly increasing and ℓ_2^{-1} is strictly decreasing. We realize that if $\alpha_1(x) > \alpha_2(x)$, then

$$P_{Y|X=x}(h(x) + \ell_1^{-1}(\alpha_1(x))) > P_{Y|X=x}(h(x) + \ell_1^{-1}(\alpha_2(x)))$$
$$P_{Y|X=x}(h(x) + \ell_2^{-1}(\alpha_1(x))) < P_{Y|X=x}(h(x) + \ell_2^{-1}(\alpha_2(x))),$$

so

$$t^d_{3,x}(h(x), \alpha_1(x)) > t^d_{3,x}(h(x), \alpha_2(x)). \tag{91}$$

Let $D = \{x \in \mathcal{X} \mid \alpha_1(x) \neq \alpha_2(x)\}$. Now, we compute

$$\mathbb{E}_{P_X}\left[(T^d_{3,X}(h(X), \alpha_1(X)) - T^d_{3,X}(h(X), \alpha_2(X)) \cdot (\alpha_1(X) - \alpha_2(X))\right] \tag{92a}$$
$$= \mathbb{E}_{P_X}\left[(T^d_{3,X}(h(X), \alpha_1(X)) - T^d_{3,X}(h(X), \alpha_2(X)) \cdot (\alpha_1(X) - \alpha_2(X))\mathbb{I}(D)\right] \tag{92b}$$
$$= \mathbb{E}_{P_X}\left[((t^d_{3,X}(h(X), \alpha_1(X)) - t^d_{3,X}(h(X), \alpha_2(X))) \cdot (\alpha_1(X) - \alpha_2(X)) \cdot \mathbb{I}(S_{\alpha_1,0} \cap S_{\alpha_2,0} \cap D)\right] \tag{92c}$$
$$+ \mathbb{E}_{P_X}\left[(t^d_{3,X}(h(X), \alpha_1(X)) + 1)(\alpha_1(X) - \alpha_2(X)) \cdot \mathbb{I}(S_{\alpha_1,0} \cap S^c_{\alpha_2,0} \cap D)\right] \tag{92d}$$
$$+ \mathbb{E}_{P_X}\left[(-1 - t^d_{3,X}(h(X), \alpha_2(X)))(\alpha_1(X) - \alpha_2(X)) \cdot \mathbb{I}(S^c_{\alpha_1,0} \cap S_{\alpha_2,0} \cap D)\right]. \tag{92e}$$

The first line holds because $(T^d_{3,x}(h(x), \alpha_1(x)) - T^d_{3,x}(h(x), \alpha_2(x)) \cdot (\alpha_1(x) - \alpha_2(x)) = 0$ on D^c. The decomposition into (92c), (92d), (92e) holds because $T^d_{3,x}(h(x), \alpha_1(x)) - T^d_{3,x}(h(x), \alpha_2(x) = 0$ when $\alpha_1(x) \leq 0$ and $\alpha_2(x) \leq 0$.

Since we have that $\alpha_1, \alpha_2 \in \mathcal{A}$ and D has positive measure, we can show that $P(S^c_{\alpha_1,0} \cap S^c_{\alpha_2,0} \cap D) < P(D)$. We consider two cases 1) $S_{\alpha_1,0} \cap D$ has positive measure and 2) $S_{\alpha_1,0} \cap D = \emptyset$. Suppose $S_{\alpha_1,0} \cap D$ has positive measure, then clearly

$$P(S^c_{\alpha_1,0} \cap S^c_{\alpha_2,0} \cap D) \leq P(S^c_{\alpha_1} \cap D) < P(D).$$

If $S_{\alpha_1,0} \cap D$ empty, this means that $\alpha_1(x) \leq 0$ for all $x \in D$. At the same time, we have that for all $\alpha \in \mathcal{A}$, $\alpha(x) \geq 0$ for every $x \in \mathcal{X}$. So, we must have that $\alpha_1 = 0$ on D. We must have that $\alpha_2(x) > 0$ on D, because α_1, α_2 must differ on D and $\alpha_2(x) \geq 0$ for all $x \in \mathcal{X}$. So, this means that $S_{\alpha_2,0} \cap D$ has positive measure, so

$$P(S^c_{\alpha_1,0} \cap S^c_{\alpha_2,0} \cap D) \leq P(S^c_{\alpha_2,0} \cap D) < P(D).$$

So, at least at least one of the sets $S_{\alpha_1,0} \cap S_{\alpha_2,0} \cap D$, $S_{\alpha_1,0} \cap S^c_{\alpha_2,0} \cap D$, $S^c_{\alpha_1,0} \cap S_{\alpha_2,0} \cap D$ has positive measure.

Suppose $S_{\alpha_1,0} \cap S_{\alpha_2,0} \cap D$ has positive measure. WLOG, if $\alpha_1(x) > \alpha_2(x)$, then $T^d_{3,x}(h(x), \alpha_1(x) - T^d_{3,x}(h(x), \alpha_2(x)) > 0$. In addition, if $\alpha_1(x) < \alpha_2(x)$, then $T^d_{3,x}(h(x), \alpha_1(x) - T^d_{3,x}(h(x), \alpha_2(x)) < 0$. Then (92c) must be positive. We can use a similar argument to verify that (92d) will be positive if $S_{\alpha_1,0} \cap S^c_{\alpha_2,0} \cap D$ has positive measure and (92e) will be positive if $S^c_{\alpha_1,0} \cap S_{\alpha_2,0}$ has positive measure. Thus, we conclude that

$$\mathbb{E}_{P_X}\left[(T^d_{3,X}(h(X), \alpha_1(X)) - T^d_{3,X}(h(X), \alpha_2(X)) \cdot (\alpha_1(X) - \alpha_2(X))\right] > 0$$

so $\alpha \mapsto \mathbb{E}_P\left[L^\Gamma_{\mathrm{RU},3}(h(X), \alpha(X), Y)\right]$ is strictly convex on $\mathcal{A}$.

D.11 Proof of Lemma 28

Recall that we defined the sieve *with truncation* Θ_m and sieve *without truncation* $\tilde{\Theta}_m$. In addition, define $\tilde{\pi}_m : \Theta \to \tilde{\Theta}_J$ to be the projection of a function $\theta \in \Theta$ onto $\tilde{\Theta}_J$.

By Lemma 5, the truncation is a contraction map to the true minimizer, so

$$||\pi_m(\theta^*) - \theta^*||_{L^2(P_X,\mathcal{X})} \leq ||\tilde{\pi}_m(\theta^*) - \theta^*||_{L^2(P_X,\mathcal{X})}.$$

Now, we verify the conditions of Lemma 18 to show that the right side of the above inequality converges to zero as $m \to \infty$. First, we note that $\tilde{\Theta}$ is a Hilbert space (with the $L^2(P_X, \mathcal{X})$ norm). Second, we note that

$$\tilde{\pi}_m(\theta^*) = \sum_{i=1}^{m} \langle \theta^*, \phi_i \rangle \phi_i,$$

where $\{\phi_i\}$ is an infinite-dimensional basis for Θ. Since $\tilde{\pi}_m(\theta^*)$ is a partial sum of the Fourier-Bessel series, we have that $\tilde{\pi}_m(\theta^*) \to \theta^*$. This implies that $||\pi_m(\theta^*) - \theta^*||_{L^2(P_X,\mathcal{X})}$, as well.

D.12 Proof of Lemma 31

First, by the Mean Value Theorem, we have that for any $z \in \mathbb{R}$,

$$|\ell'(z)| = |\ell'(z) - \ell'(0)| \leq |\ell''(\tilde{z})| \cdot |z|,$$

where $\tilde{z}$ is between z and 0. By Assumption 4, $|\ell''(\tilde{z})| \leq C_{L,u}$, so

$$|\ell'(z)| \leq C_{L,u} \cdot |z|. \tag{93}$$

Again, by the Mean Value Theorem, we have that for any $h \in \Lambda_c^p(\mathcal{X})$ and $x \in \mathcal{X}$,

$$\begin{aligned} |L(h(x), y) - L(h^*(x), y)| &= |\ell(y - h(x)) - \ell(y - h^*(x))| \\ &= |\ell'(y - (\lambda(x) \cdot h(x) + (1 - \lambda(x)) \cdot h^*(x)))| \cdot |h(x) - h^*(x)| \qquad \lambda(x) \in [0, 1]. \end{aligned}$$

We can define $\bar{L}(x, y) = |\ell'(y - (\lambda(x) \cdot h(x) + (1 - \lambda(x)) \cdot h^*(x)))|$. Now, we aim to verify that there exists some $0 < M < \infty$ such that

$$\sup_{x \in \mathcal{X}} \mathbb{E}_{P_{Y|X}} \left[\bar{L}(X, Y)^2 \mid X = x \right] < M.$$

We apply (93).

$$\begin{aligned} \mathbb{E}_{P_{Y|X}} \left[\bar{L}(x, Y)^2 \mid X = x \right] &= \mathbb{E}_{P_{Y|X}} \left[(\ell'(Y - (\lambda(x) \cdot h(x) + (1 - \lambda(x) \cdot h^*(x)) \cdot h^*(x))))^2 \mid X = x \right] \\ &= \mathbb{E}_{P_{Y|X}} \left[(\ell'(Y - (\lambda(x) \cdot h(x) + (1 - \lambda(x)) \cdot h^*(x))) \cdot h^*(x))^2 \mid X = x \right] \\ &= \mathbb{E}_{P_{Y|X}} \left[C_{L,u}^2 \cdot ((Y - (\lambda(x) \cdot h(x) + (1 - \lambda(x)) \cdot h^*(x))) \cdot h^*(x))^2 \mid X = x \right] \\ &\lesssim \mathbb{E}_{P_{Y|X}} \left[Y^2 \mid X = x \right] + h(x)^2 + h^*(x)^2 \\ &\lesssim \sup_{x \in \mathcal{X}} \mathbb{E}_{P_{Y|X}} \left[Y^2 \mid X = x \right] + c^2 \\ &< \infty. \end{aligned}$$

The last two lines follow from Assumption 5 and 6. Assumption 5 gives that $h, h^* \in \Lambda_c^p(x)$, so $|h(x)| \leq c$ and $|h^*(x)| \leq c$. Assumption 6 gives that $\sup_{x \in \mathcal{X}} \mathbb{E}_{P_{Y|X}} \left[Y^2 \mid X = x \right]$ is finite. Thus, $\sup_{x \in \mathcal{X}} \mathbb{E}_{P_{Y|X}} \left[\bar{L}(X, Y)^2 \mid X = x \right] < \infty$.