Learning from a Biased Sample

Roshni Sahoo
rsahoo@stanford.edu

Lihua Lei
lihualei@stanford.edu

Stefan Wager
swager@stanford.edu

Abstract

The empirical risk minimization approach to data-driven decision making requires access to training data drawn under the same conditions as those that will be faced when the decision rule is deployed. However, in a number of settings, we may be concerned that our training sample is biased in the sense 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 conditional 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 a case study on prediction of mental health scores from health survey data 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 [6, 39, 73]. 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^* [73].

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 cause decision rules learned via empirical risk minimization on the training data to incur high risk on the target distribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ML x OR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

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 online health surveys, Q is the nationwide adult distribution, whereas P is the distribution over survey respondents who 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. A popular solution is to define a robustness set $\mathcal{S}$, a family of distributions that are related to the training distribution P, that likely contains the true target distribution Q; and then to use distributionally robust optimization (DRO) [5, 61] to learn a decision rule that minimizes the worst-case risk over $\mathcal{S}$, i.e.

$$\operatorname*{argmin}_h \sup_{Q\in\mathcal{S}(P)} \mathbb{E}_Q\left[L(h(X), Y)\right]. \tag{3}$$

Applying DRO effectively hinges on choosing an appropriate robustness set: One that is large enough to contain the true target distribution but at the same time is not overly conservative.

Our approach starts by proposing a model for sampling bias that then induces natural robustness sets tailored to challenges arising from learning data-driven decision rules in settings like those highlighted above. Our proposed model of sampling bias, conditional Γ-biased sampling, responds to this insight by allowing arbitrary sampling bias along observed covariates X but bounding the amount of bias due to unobservables. Formally, the model is an extension of the one used in [2] and [46] to the setting where there are covariates X that may affect whether a sample is selected. Here, $\Gamma \geq 1$ captures the allowed strength of sampling bias, and larger values of Γ allow for more bias. Note that, with $\Gamma = 1$ (i.e., no sample selection based on unobservables), this model corresponds to unconfounded sample selection model that is widely studied in the literature on generalizability (e.g., [65, 71, 72]).

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 conditional 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, the (X, Y)-marginal of $\tilde{Q}$ conditionally on $S = 1$ is equal to P, and

$$\frac{\mathbb{P}_{\tilde{Q}}\left[S = 1 \mid X = x, Y = y\right]}{\mathbb{P}_{\tilde{Q}}\left[S = 1 \mid X = x\right]} \in [\Gamma^{-1}, \Gamma] \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}. \tag{4}$$

The main contribution of this paper is a method for robust loss minimization under conditional Γ-biased sampling. To operationalize this goal, we define a robustness set $\mathcal{S}_\Gamma(P, Q_X)$ that consists of all distributions that can generate P via conditional Γ-biased sampling and have covariate distribution Q_X. Notably, this robustness set places restrictions on the conditional distribution $Y|X$ instead of the joint distribution or covariate distribution. We then seek to learn

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

for arbitrary feature distributions Q_X.

Our proposed method, Rockafellar-Uryasev (RU) Regression, involves learning h^*_Γ via (penalized) empirical minimization of the loss L^Γ_{RU}, which we refer to as the RU loss. We use this name because because results of [55] play a key role in our derivation of this loss function. This paper investigates RU Regression both theoretically and empirically.

2 Rockafellar-Uryasev Regression

We propose a method for solving the DRO problem (5) that arises from the assumption of conditional Γ-biased sampling. Our first theorem reformulates (5) as the minimizer of the expectation of a convex function over data drawn from the training distribution P.

Theorem 2.1. *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 $Y \subseteq \mathbb{R}$. Suppose that $P_{Y|X=x}$ is absolutely continuous with respect to Lebesgue measure for every $x \in \mathcal{X}$. 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,*

$$L^{\Gamma}_{RU}(z, a, y) = \Gamma^{-1} L(z, y) + (1 - \Gamma^{-1})a + (\Gamma - \Gamma^{-1})(L(z,\, y) - a)_+, \tag{6}$$

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

$$\{h^*_\Gamma(\cdot),\, \alpha^*_\Gamma(\cdot)\} \in \underset{(h,\, \alpha) \in L^2(P_X,\, \mathcal{X}) \times L^2(P_X,\, \mathcal{X})}{\operatorname{argmin}} \mathbb{E}_P \left[L^{\Gamma}_{RU}(h(X),\, \alpha(X),\, Y) \right] \tag{7}$$

is also a solution to (5) *for any Q_X that is absolutely continuous with respect to P_X, i.e., $Q_X \ll P_X$, and $\sup_{x \in \mathcal{X}} \frac{dP_X(x)}{dQ_X(x)} < \infty$.*

We next demonstrate how the general regularity properties established above translate into convergence guarantees for RU regression in the familiar setting where h^*_Γ is known to belong to a Hölder class defined in Appendix C. Optimal estimation in Hölder classes is a widely studied problem [13]; and, in particular the minimax-optimal rate of convergence for nonparametric regression over the Hölder class of p-smooth functions in d dimensions is known to be $O_P(n^{-\frac{p}{2p+d}})$ [63]. Thus, studying the behavior of RU regression in this setting provides a transparent benchmark for statistical properties empirical minimization with the RU loss. Our main result below, Theorem 2.2, will demonstrate that—up to log factors—RU Regression matches the minimax rate of convergence of [63] for nonparametric regression with p-smooth functions. We omit the proof in this paper for the sake of length constraint. A roadmap can be found in Appendix C and the detailed proofs can be found online [58].

Theorem 2.2. *Suppose that Assumptions 1, 2, 3, 4, 5, 6, 7 in Appendix C hold. Let $(\hat{h}_n, \hat{\alpha}_n)$ be the sieve empirical risk estimator defined in* (26) *with Let $J_n \asymp (\frac{n}{\log n})^{\frac{1}{2p+d}}$. Then $(\hat{h}_n, \hat{\alpha}_n)$ achieves*

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

Furthermore, if $Q_X \ll P_X$ and $\sup_{x \in \mathcal{X}} dQ_X(x) \,/\, dP_X(x) < \infty$, then the same rate of convergence holds over $L^2(Q_X, \mathcal{X})$ also.

3 An Experiment: Learning from an Online Health Survey

This case study is motivated by well-known challenges of working with online health surveys. As described in Example 1, online health surveys are used for population health measurement but suffer heavily from nonresponse. In particular, [36] use data from various surveys to train models to predict the prevalence of mental health conditions, and they find that a prediction model trained on the Household Pulse Survey [9, HPS], an online health survey conducted by the Census Bureau across the United States, overestimates the prevalence of mental health conditions compared to a model trained on the Behavioral Risk Factor Surveillance System [10, BRFSS], a telephone survey conducted by the CDC across the United States. In this case study, we use RU Regression to train prediction models on HPS data and assess whether this approach improves generalization to the BRFSS data.

For our training data, we use survey responses from the 2021 HPS ($n_1 = 1,121,213$). For the target data, we use survey responses from the 2021 BRFSS ($n_2 = 423,807$). While both surveys aim to be representative of the United States adult population, HPS is a Census Bureau Experimental Data Product with only a 2-10 % response rate, while BRFSS has a response rate of 44%. There is a concern that, given the low response rate of the HPS, responders to the online survey may be materially different along unobserved attributes than non-responders (and thus the general population) [8, 36]. This motivates our choice to treat the BRFSS responses as a (near-)true target population Q, while we consider the HPS responses as drawn from a potentially biased population P.

The covariates $X \in \mathbb{R}^d$ ($d = 44$) include individual-level demographic features such as age, gender, education level, income, race/ethnicity, household size, and state-level characteristics such as unemployment rate and proportion of the state with private health insurance, corresponding to the state of the individual. The mental health indicator Y is the PHQ-4 score, which is the 4-item Patient

Figure 1: We report the risk obtained by robust models trained on HPS training set and evaluated on the BRFSS dataset and a held-out HPS test set when h is a neural network. Bootstrap standard errors are computed with 5000 bootstrap samples.

Health Questionnaire (PHQ-4) screening scale of anxiety-depression [40]. The BRFSS 2021 does not measure the PHQ-4 and instead measures 30 day prevalence of anxiety-depression; however, we are able to impute PHQ-4 scores for BRFSS 2021 respondents using a conversion formula learned from the Depression and Anxiety Module of the 2018 BRFSS survey that measures both outcomes on a subset of respondents.

Following [37], we use Poisson regression to model PHQ-4 scores $Y \in \{0, 1, 2, \dots 12\}$. The loss function is the Poisson negative log likelihood:

$$L(h(X), Y) = -h(X) \cdot Y + \exp(h(X)).$$

We consider 6 approaches to learning h on the HPS 2021 data such as to obtain low Poisson loss on the target BRFSS 2021 data. First, we consider both conditional and unconditional RU regression, with h learned over both flexible neural networks as above and as a linear function (see Remark 4). Second, we consider the transductive-type setting where we get to observe the distribution of the features X_i on BRFSS (but not that of the outcomes) during training, and use this information to conduct weighted RU regression as described in Section C.3 (again using both the flexible neural network representation and the linear class for h). We estimate covariate weights via probabilistic classification [66, 44] on a subset of the covariates from the HPS and BRFSS. We split the HPS 2021 dataset into train, validation, and an additional test set with 403636, 269091, and 448486 samples, respectively. The BRFSS 2021 dataset consists of 423807 samples, only used at test-time.

Results are shown in Figure 1. We learn a robust neural network h. Again, we find that Weighted RU Regression and RU Regression obtain lower BRFSS risk than Unconditional RU Regression for each value of Γ. We find that Weighted RU Regression and RU Regression perform comparably, which is expected because the use of covariate weights should not have a large effect if the model class is sufficiently flexible. Across all experiments, we find that as Γ increases, the RU Regression variants reduce the BRFSS (target) risk, at the cost of increasing the HPS (train) risk. Once Γ becomes too large, we observe that the BRFSS risk begins to increase also.

4 Conclusion

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 [55].

References

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

[2] 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.

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

[4] 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.

[5] 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.

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

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

[8] Valerie C Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. Unrepresentative big surveys significantly overestimated us vaccine uptake. *Nature*, 600(7890):695–700, 2021.

[9] US Census Bureau. Measuring household experiences during the coronavirus pandemic, 2021.

[10] CDC. Behavioral risk factor surveillance system survey data, 2021.

[11] 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.

[12] 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.

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

[14] 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.

[15] Victor Chernozhukov, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis. Long story short: Omitted variable bias in causal machine learning. Technical report, National Bureau of Economic Research, 2022.

[16] 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.

[17] 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.

[18] 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.

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

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

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

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

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

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

[25] Tobias Freidling and Qingyuan Zhao. Sensitivity analysis with the r^2-calculus. *arXiv preprint arXiv:2301.00040*, 2022.

[26] Pascal Geldsetzer. Use of rapid online surveys to assess people's perceptions during infectious disease outbreaks: a cross-sectional survey on covid-19. *Journal of medical Internet research*, 22(4):e18790, 2020.

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

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

[29] 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.

[30] 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.

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

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

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

[34] 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, 2016.

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

[36] Ronald C Kessler, Wai Tat Chiu, Irving H Hwang, Victor Puac-Polanco, Nancy A Sampson, Hannah N Ziobrowski, and Alan M Zaslavsky. Changes in prevalence of mental illness among us adults during compared with before the covid-19 pandemic. *Psychiatric Clinics*, 45(1):1–28, 2022.

[37] Ronald C Kessler, Christopher J Ruhm, Victor Puac-Polanco, Irving H Hwang, Sue Lee, Maria V Petukhova, Nancy A Sampson, Hannah N Ziobrowski, Alan M Zaslavsky, and Jose R Zubizarreta. Estimated prevalence of and factors associated with clinically significant anxiety and depression among us adults during the first year of the covid-19 pandemic. *JAMA Network Open*, 5(6):e2217223–e2217223, 2022.

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

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

[40] Kurt Kroenke, Robert L Spitzer, Janet BW Williams, and Bernd Löwe. An ultra-brief screening scale for anxiety and depression: the phq–4. *Psychosomatics*, 50(6):613–621, 2009.

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

[42] Patrick G Lyons, Mackenzie R Hofford, C Yu Sean, Andrew P Michelson, Philip RO Payne, Catherine L Hough, and Karandeep Singh. Factors associated with variability in the performance of a proprietary sepsis prediction model across 9 networked hospitals in the us. *JAMA internal medicine*, 183(6):611–612, 2023.

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

[44] Aditya Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning*, pages 304–313. PMLR, 2016.

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

[46] 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.

[47] 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.

[48] 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.

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

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

[51] 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.

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

[53] Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019.

[54] 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.

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

[56] Andrzej Ruszczyński and Alexander Shapiro. Risk averse optimization. In Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński, editors, *Lectures on Stochastic Programming: Modeling and Theory*, chapter 6, pages 223–305. Society for Industrial and Applied Mathematics, 3rd edition, 2021.

[57] 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.

[58] Roshni Sahoo, Lihua Lei, and Stefan Wager. Learning from a biased sample. *arXiv preprint arXiv:2209.01754*, 2022.

[59] Joshua A Salomon, Alex Reinhart, Alyssa Bilinski, Eu Jing Chua, Wichada La Motte-Kerr, Minttu M Rönn, Marissa B Reitsma, Katherine A Morris, Sarah LaRocca, and Tamer H Farag. The us covid-19 trends and impact survey: Continuous real-time measurement of covid-19 symptoms, risks, protective behaviors, testing, and vaccination. *Proceedings of the National Academy of Sciences*, 118(51):e2111454118, 2021.

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

[61] Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

[62] 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.

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

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

[65] 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.

[66] Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul Von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 2008.

[67] 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.

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

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

[70] Aleksandr Filippovich Timan. *Theory of approximation of functions of a real variable*. MacMillan, 1963.

[71] 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.

[72] 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.

[73] Vladimir N Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., 1995.

[74] 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.

[75] Andrew Wong, Erkin Otles, John P Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, and Carleen Penoza. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA internal medicine*, 181(8):1065–1070, 2021.

[76] 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.

[77] 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.

[78] Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183, 2023.

A More background

A.1 Examples

Example: Nonresponse in Online Health Surveys Large-scale online health surveys are a recently popularized tool for public health surveillance [26, 9, 59]. While these surveys are cheap to deploy, they suffer from high levels of nonresponse, which can yield biased predictions of the outcome. For example, [36] find that online surveys, such as the Household Pulse Survey (HPS), yield implausibly high estimates for the prevalence of anxiety and depression during the pandemic, compared to a telephone survey called Behavioral Risk Factor Surveillance System (BRFSS) survey. They

hypothesize that HPS respondents differ from members of the general population in their unmeasured psychological characteristics, as well as geographic-demographic characteristics. Similarly, [8] find that online surveys overestimate vaccine uptake and hypothesize that online surveys may be unrepresentative with respect to political partisanship, which has been found to be correlated with vaccine behavior and with survey response. Thus, data-driven rules learned via empirical risk minimization with data from online health surveys may not generalize well to real-world settings.

Example: Site Selection in Development of Medical Risk Models. Various risk predictors are widely used to guide both clinical practice and hospital logistics. Bias may arise if a risk model that is trained using data from one hospital is then deployed at another hospital, and the two hospitals have different patient populations. For example, the Epic Sepsis Model (ESM), a proprietary sepsis prediction model deployed at hundreds of US hospitals, generates automated alerts to warn clinicians that patients may be developing sepsis. In an external validity study, [75] found that ESM performed much worse (AUC, 0.63) on Michigan Medicine hospitalization data than the reported performance by Epic Systems (AUC, 0.73). Followup analysis by [42] suggests that this performance gap may be driven by differences in sepsis presentation and comorbidities among patient populations at different hospitals. Patient populations may differ along observable attributes, as well as unobservable attributes.

Example: Self-Selection in Randomized Trials. In randomized trials for estimating treatment effects, participants often volunteer or apply to be a part of the study. For instance, [4] measure the effect of a vocational training program on labor market outcomes in a randomized trial. However, participants were not randomly sampled from the target population; they needed to apply to be a part of the study. Similarly, [74] describes that the effectiveness of anti-depressants is assessed in randomized trials involving volunteers. In such studies, participants may differ from non-participants in fundamental ways, and so data-driven rules learned using data collected from study participants may again fail to generalize to the full population.

A.2 Conditional vs. Unconditional Distributional Robustness

Our contribution fits broadly within a large existing literature on distributionally robust optimization (DRO) [5, 61]. Most existing work in this space has either focused on constructing global (or unconditional) robustness sets about the joint distribution over (X, Y) [20, 19, 30, 45, 47, 52, 57], or just the covariate distribution over X [19]. We believe, however, that addressing the challenges arising in our motivating examples requires working with conditional robustness sets, e.g., as induced by Definition 1, that let us specifically focus on bias along unobservables.

To give a concrete example of unconditional DRO, [20] consider the problem of learning

$$h^* = \operatorname*{argmin}_h \sup \Big\{ \mathbb{E}_Q \left[L(h(X), Y) \right] : D_f(Q|P) \leq \Gamma \Big\}, \quad D_f(Q|P) = \int f\Big(\frac{dQ}{dP}\Big) dP, \tag{8}$$

where D_f is an f-divergence. One can verify that, if we consider an "improper" f-divergence

$$f(z) = \begin{cases} 0 & \Gamma^{-1} \leq z \leq \Gamma \\ \infty & \text{else,} \end{cases} \tag{9}$$

then this robustness set is consistent with unconditional Γ-biased sampling, which is an analogue of Definition 1 defined below. As discussed in Section 2 below, DRO under this unconditional robustness set can also be solved via an augmented convex formulation which we refer to as Unconditional RU regression—and will use as a baseline for our approach throughout.

Definition 2. Let $\Gamma \geq 1$. For any pair of distributions P and Q over (X, Y), we say that Q can generate P under *unconditional* 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, the (X, Y)-marginal of $\tilde{Q}$ conditionally on $S = 1$ is equal to P, and

$$\Gamma^{-1} \cdot \mathbb{P}_{\tilde{Q}} \left[S = 1 \right] \leq \mathbb{P}_{\tilde{Q}} \left[S = 1 \mid X = x, Y = y \right] \leq \Gamma \cdot \mathbb{P}_{\tilde{Q}} \left[S = 1 \right] \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}. \tag{10}$$

In unconditional Γ-biased sampling, there is no distinction between observables and unobservables, and together they can only affect an individual's probability of selection at most Γ. When $\Gamma = 1$

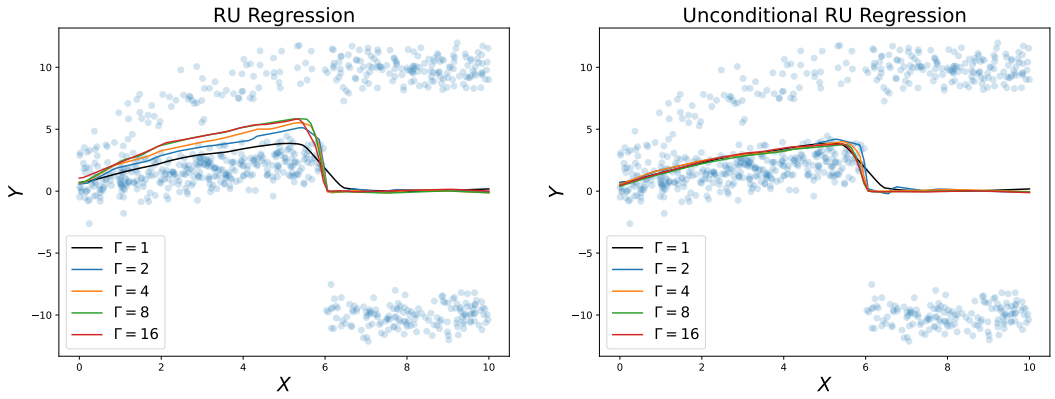

Figure 2: Comparison of (Conditional) RU Regression and Unconditional RU Regression in an example where the optimal model has a heteroscedastic loss distribution.

in unconditional Γ-biased sampling, $P = Q$ because all units have the same probability of being selected; whereas, as noted above, when $\Gamma = 1$ our conditional Γ-biased sampling model reduces to unconfounded sample selection.

We argue that conditional models for sample selection are a better fit than unconditional ones in many settings—including the examples discussed above. First, when there is sampling bias along both observables and unobservables, the bias parameter Γ required for the conditional restriction (4) to hold will generally be smaller than the unconditional restriction (10) because the conditional restriction does not need to account for sampling bias along the observed X. Thus, in applications where Γ is chosen based on substantive information, it's likely that the use of the conditional restriction (4) will enable using a smaller Γ—and thus require less conservatism under the data-collection distribution.

A second, more subtle issue with methods induced by the unconditional robustness set (10) is that they do not always provide robustness that is useful in practice. This issue is illustrated in Figure 2, using a simple one-dimensional simulation example. Here, when X_i is small, accurate prediction is possible but data is split in two unequal-sized bands—and we may worry that the relative size of these bands is unrepresentative due to sampling bias. When X_i is large, outcomes are not predictable and large errors are unavoidable. In this setting, our proposed method, RU Regression, does what we want it to do, i.e., the larger we make Γ, the more it shifts towards making predictions that are accurate over both bands of data in case their relative importance changes. On the other hand, the baseline method targeting (10), Unconditional RU Regression, is essentially unaffected by changing Γ. Qualitatively (and in a sense that will be made precise through our the formal arguments), this is because pessimism under (10) leads the method to upweight the "unpredictable" region where X_i is large relative to the "predictable" region where X_i is small—but in the end this isn't useful for robust prediction since we already had the ability to flexibly react to X_i during prediction.

We see both of these phenomena play out in our experiments presented in Section 3. Across both a semi-synthetic experiment and a real-world evaluation, we find that robust learning under the conditional restriction (4) enables better tradeoffs between accuracy under the data-collection distribution and the target distribution than methods motivated by the unconditional restriction (10).

Remark 1. Another advantage of the conditional restriction is that, in some settings, it is realistic to assume knowledge of the true population covariate distribution Q_X at train-time; see, e.g., our application to health surveys with sampling bias in Section 3. This then enables us to a suite of well known reweighting techniques to adjust for any shift along measurable attributes [65, 71, 72]. In contrast, in our examples, the target conditional distribution $Q_{Y|X}$ is always unknown at train-time, so the only tool available to the analyst (if they cannot collect more data) is to posit a model on the shift due to unobservables. In settings like these, the conditional Γ-biased sampling model places an assumption only on the part of the problem that is truly unidentified from data.

A.3 Related Work

Our proposed model of sampling bias, conditional Γ-biased sampling, builds on previous models for sampling bias [2, 46], 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, previous 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 [2] and [46] that includes covariates in such a way that we allow the unobserved probability of sample selection π to be arbitrarily affected by the covariates X but bounds on the amount of unexplained variation in π_i. Also, unlike [2] and [46], we focus on learning a robust decision rules rather than on partial identification of moments of Q.

Our model is also connected to the broader literature on sensitivity analysis in causal inference [1, 18, 25, 33, 49, 76], 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 [68] 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.

As discussed above, our work fits within the broader DRO literature [5, 61], but because the vast majority of that literature focuses on robustness to global or unconditional shifts the resulting methods and analytic techniques are not directly applicable to our setting. We do note, however, that there are a handful of recent works that also consider robustness sets that place restrictions on conditional shifts [22, 51, 69]. [22] 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 $\mathcal{X}$. [51] leverages access to noisy proxies of unobserved variables for learning models that are robust to shifts in the distribution of unobservables. [69] studies how to evaluate the worst-case loss under a parametric robustness set, which consists of interpretable, conditional shifts. Our work differs from [51, 69] in that we do not make any fine-grained assumptions on the nature of the shift, such as 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 [18, 33, 49, 76]. These works use robust optimization for a different goal of obtaining partial identification bounds on the treatment effect. Most related to our work, [18] consider a model that places pointwise bounds on the conditional odds ratio and uses results of [55] to obtain formulas for partial identification bounds.

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 [3], [6], [21], [24], [35], [39], [43], [50], [64], [67], [77] and [78]. A recurring theme of this line of work is in choosing loss functions $L(\cdot)$ that captures relevant aspects of various decision tasks [6]. 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 [35], 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 [35] 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.

B More About RU Regression

Lemma B.1. *Let P, Q be the distributions over (X, Y). Suppose that $P_{Y|X=x}, Q_{Y|X=x}$ are absolutely continuous with respect to Lebesgue measure for every $x \in \mathcal{X}$. Q can generate P via conditional Γ-biased sampling if and only if*

$$\Gamma^{-1} \leq \frac{dQ_{Y|X=x}(y)}{dP_{Y|X=x}(y)} \leq \Gamma, \quad \forall x \in \mathcal{X}, y \in \mathcal{Y} \tag{11}$$

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

Proof of Theorem 2.1. The first claim regarding convexity of L^{Γ}_{RU} follows immediately using the standard rules for composing convex functions [7]. We focus on the second claim of Theorem 2.1. To this end, we start by reducing the worst-case population risk minimization problem in (5) to separate worst-case conditional risk minimization for each $x \in \mathcal{X}$ using the following lemma. □

Lemma B.2. *A function* $h \in L^2(P_X, \mathcal{X})$ *solves* (5) *if and only if,* h *solves the following almost surely for* $X \sim P_X$*:*

$$h(X) = \operatorname*{argmin}_{h\in\mathbb{R}} \sup \Big\{ \mathbb{E}_{Q_{Y|X}} \left[L(h, Y) \mid X \right] : Q \in S_\Gamma(P, Q_X) \Big\}. \tag{12}$$

We can apply Lemma B.1 to view the inner maximization of (12) as the maximization of a linear function subject to convex constraints. The optimization variable in the inner problem is $dQ_{Y|X=x}$ and the constraints require that:

1. $dQ_{Y|X=x}$ is a valid probability distribution for $x \in \mathcal{X}$, i.e. $\int dQ_{Y|X=x}(y) = 1$, and
2. the Γ-biased sampling condition (11) holds.

Moreover, the worst-case conditional distribution $dQ^*_{Y|X=x}$ must satisfy the Γ-biased sampling condition, i.e., $dP_{Y|X=x}(y) \,/\, dQ^*_{Y|X=x}(y) \in \{\Gamma^{-1}, \Gamma\}$ because the supremum of a convex function over a closed, bounded, convex set exists and is achieved at some extreme point of the feasible set and $P_{Y|X=x}$ is absolutely continuous with respect to Lebesgue measure.

We next show that this supremum admits a simple characterization. Let $F_{x;h(x)}(z)$ be the c.d.f. of $L(h(x), Y)$ where Y is distributed according to $P_{Y|X=x}$, i.e., $F_{x;h(x)}(z)$ is the distribution over the conditional losses when $X = x$, let $q^L_\eta(x; h(x))$ be the η-th quantile of the distribution over the conditional losses when $X = x$,

$$q^L_\eta(x; h(x)) = F^{-1}_{x;h(x)}(\eta), \tag{13}$$

and let

$$\eta(\Gamma) = \frac{\Gamma}{\Gamma + 1}. \tag{14}$$

Then, the worst-case distribution can be written as

$$dQ^*_{Y|X=x}(y) = \begin{cases} \Gamma \cdot dP_{Y|X=x}(y) & \text{if } L(h(x), y) \geq q^L_{\eta(\Gamma)}(x; h(x)) \\ \Gamma^{-1} \cdot dP_{Y|X=x}(y) & \text{o.w.} \end{cases} \tag{15}$$

To verify that this is in fact the worst-case distribution, note that here we assign a weight Γ to the values of y that yield values of $L(h(x), y)$ that exceeds $q^L_\eta(x; h(x))$ for some $\eta \in (0, 1)$ and weight Γ^{-1} to the values of y that yield values of $L(h(x), y)$ which fall above this threshold. And the worst-case distribution must do this; otherwise, there would exist a distribution $dQ_{Y|X=x}$ that obtains higher risk than $dQ^*_{Y|X=x}$. The choice of quantile η is set to ensure that $dQ^*_{Y|X=x}$ is a valid probability distribution; we pick η that satisfies $\Gamma^{-1}(1 - \eta) + \Gamma \cdot \eta = 1$. Solving this equation yields $\eta(\Gamma)$ as defined in (14).

Next, we can use (15) to verify that

$$\begin{aligned} &\sup\{\mathbb{E}_{Q_{Y|X}} \left[L(h(X), Y) \mid X = x \right] : Q \in S_\Gamma(P, Q_X)\} \\ &= \mathbb{E}_{P_{Y|X}} \left[L(h(X), Y) \Big(\Gamma^{-1} + (\Gamma - \Gamma^{-1}) \mathbb{I}(L(h(X), Y) \geq q^L_{\eta(\Gamma)}(X; h(X)) \Big) \mid X = x \right], \end{aligned} \tag{16}$$

and so (12) can be rewritten as

$$\min_{h(x)\in\mathbb{R}} \mathbb{E}_{P_{Y|X}} \left[L(h(x), Y) \Big(\Gamma^{-1} + (\Gamma - \Gamma^{-1}) \mathbb{I}(L(h(x), Y) \geq q^L_{\eta(\Gamma)}(X; h(x)) \Big) \mid X = x \right]. \tag{17}$$

Thus, we can focus on the optimization problem in (17).

We realize that the objective in (17) is closely related to the conditional value-at-risk (CVaR) [55], 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

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

Applying the CVaR definition, we realize that

$$\mathbb{E}_{P_{Y|X}}\left[L(h(X),Y)\mathbb{I}(L(h(X),Y) > q_\eta^L(X;h(X))) \mid X = x\right] = (1-\eta(\Gamma))\cdot\text{CVaR}_{\eta(\Gamma)}(L(h(x),Y)). \tag{18}$$

Substituting (18) into (17) and simplifying gives the following problem

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

We are now ready to use the influential result of [55, 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:[1]

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

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 $\text{CVaR}_{\eta(\Gamma)}(L(h(x),Y))$ from (19) as

$$\text{CVaR}_{\eta(\Gamma)}(L(h(x),Y)) = \min_{\alpha(x)\in\mathbb{R}} \alpha(x) + \frac{1}{1-\eta(\Gamma)}\mathbb{E}_{P_{Y|X}}\left[(L(h(x),Y)-\alpha(x))_+|X=x\right]. \tag{21}$$

Furthermore, by Theorem 2 of [55], any minimizer of the following joint optimization also minimizes the CVaR. In particular,

$$\min_{h\in H}\text{CVaR}_\eta(L(h,Y)) = \min_{(h,\alpha)\in H\times\mathbb{R}} \alpha + (1-\eta)^{-1}\mathbb{E}_Y\left[(L(h,Y)-\alpha)_+\right]. \tag{22}$$

Applying this theorem to (19), we have that

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

The last line follows from the definition of L_{RU}^Γ in (6). In other words, (19) can be written as the augmented conditional risk minimization

$$\min_{h(x),\alpha(x)\in\mathbb{R}} \mathbb{E}_{P_{Y|X}}\left[L_{\text{RU}}^\Gamma(h(x),\alpha(x),Y) \mid X = x\right]. \tag{23}$$

Functions $h_\Gamma^*, \alpha_\Gamma^*$ that solve (23) also solve (7) for every $x \in \text{supp}(P_X)$ almost surely. In addition, any minimizer of (23) also solves (12) for any $x \in \text{supp}(P_X)$ almost surely. Since $Q_X \ll P_X$ and Lemma B.2 holds, we have that functions that minimize (12) almost surely for any $x \in \text{supp}(P_X)$ also minimize (5). □

The derivation of RU Regression reveals that the optimal robust decision rule is agnostic to the target covariate distribution Q_X as long as it is absolutely continuous with respect to the training covariate distribution P_X and has $\sup_{x\in\mathcal{X}} dP_X(x)\,/\,dQ_X(x) < \infty$. This is because the optimal rule in fact minimizes the worst-case conditional loss for $x \in \text{supp}(P_X)$ almost surely. Note that this universality via reduction to conditional risk minimization is only applicable when the decision rule h and auxiliary function α can represent the conditionally-optimal decision rule. This setting arises

[1] A similar argument is made in [56, Example 6.19].

in Section C.2 when we consider learning over sieve spaces that eventually span $L^2(P_X,\, \mathcal{X})$, and in Section 3 when we consider joint optimization of deep neural networks to learn the solution of (7). On the other hand, if we want to learn the optimal decision rule h over a constrained function class, conditional and population risk minimization may no longer equivalent; and, as discussed in Section C.3, and additional weighting correction that depends on Q_X will be needed.

Remark 2. The techniques used to prove Theorem 2.1 can also be applied to study robust learning under the unconditional Γ-biased sampling model (9), resulting in the statement in Corollary B.1 below.[2] We refer empirical minimization with the resulting objective (24) as Unconditional RU Regression. While (Conditional) RU Regression learns the optimal robust decision rule under the assumption of conditional Γ-biased sampling, Unconditional RU Regression learns the optimal robust decision rule under the assumption of unconditional Γ-biased sampling. The main difference between RU Regression and Unconditional RU Regression is that in Unconditional RU Regression, we only learn a one-dimensional auxiliary parameter α, while in RU Regression we must fit an auxiliary function $\alpha(X)$.

Corollary B.1. *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 $\Gamma > 1$. Any solution*

$$\{h^*_\Gamma(\cdot),\, \alpha^*_\Gamma\} \in \underset{(h,\, \alpha) \in L^2(P_X,\, \mathcal{X}) \times \mathbb{R}}{\operatorname{argmin}} \mathbb{E}_P \left[L^\Gamma_{RU}(h(X),\, \alpha,\, Y) \right] \tag{24}$$

is also a solution to (8) *where f is given by* (9)*.*

C Large-Sample Theory

In the previous section, we showed that the minimax decision rule under conditional 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}.

One challenge in doing so is that $L^\Gamma_{\text{RU}}(z,\, a,\, y)$ is not strongly convex in (z, a); and in fact, it is not even strongly convex in expectation when $a < 0$. The following results show, however, that the population RU risk has a unique minimizer—and is strongly convex and smooth in a neighborhood around the minimizer. These properties enable us to obtain estimation guarantees when the optimal robust decision rule lies in a p-Hölder space, a class of smooth functions. 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 conditional Γ-biased sampling. We conclude this section by providing a weighted version of RU Regression for learning the optimal robust decision rule over a constrained function class.

C.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 use 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 assumptions, we can show that the population RU risk has a unique minimizer over $L^2(P_X, \mathcal{X}) \times L^2(P_X, \mathcal{X})$.

Assumption 1. $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$ is compact. In particular, there exists a constant B such that $0 < B < \infty$ and $\mathcal{Y} \subset [-B, B]$.

[2] While this result is conceptually similar to the results of [20], the choice of f that corresponds to the robustness set we consider is discontinuous and unbounded, so the formal results (and proof strategies) of [20] do not apply.

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

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 $\mathcal{Y}$. As a consequence, we can define

$$P_{\min,\Gamma} := \inf_{c\in[1-\frac{\eta(\Gamma)}{2},1+\frac{\eta(\Gamma)}{2}],x\in\mathcal{X}} p_{Y|X=x}(q_c^Y(x)), \tag{25}$$

where $q_c^Y(x)$ denotes the c-th quantile of $P_{Y|X=x}$, and note $P_{\min,\Gamma} > 0$. We assume that there exists $P_{\max}$ such that $\sup_{x\in\mathcal{X},y\in\mathcal{Y}} p_{Y|X=x}(y) \le P_{\max}$, where $0 < P_{\max} < \infty$.

Theorem C.1. *Under Assumptions 1, 2, 3,* $\mathbb{E}_P\left[L_{RU}^\Gamma(h(X), \alpha(X), Y)\right]$ *has a unique minimizer* $(h_\Gamma^*, \alpha_\Gamma^*)$ *over* $L^2(P_X, \mathcal{X}) \times L^2(P_X, \mathcal{X})$. *In addition,*

$$\alpha_\Gamma^*(x) = q_{\eta(\Gamma)}^L(x; h_\Gamma^*(x)),$$

and there exist positive constants M^-, M_Γ^+ *such that*

$$M^- < \alpha_\Gamma^*(x) < M_\Gamma^+ \quad \forall x \in \mathcal{X}.$$

In particular, M^- *depends on* $P_{\max}$ *and loss function L, and* M_Γ^+ *depends on* B, Γ *and loss function* L.

Building on this characterization of the minimizer, we can show in a $||\cdot||_\infty$-ball about the minimizer, the population RU loss is strongly convex and smooth. To show smoothness, we require the loss function L to be D-smooth (have second derivative upper bounded by D) for some constant $0 < D < \infty$. Theorem C.2 below implies that in a neighborhood about the minimizer,

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

and this property will be useful for establishing nonparametric estimation guarantees in the following section.

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

Theorem C.2. *Let* $\mathcal{C}_\delta = \{(h, \alpha) \in L^2(P_X, \mathcal{X}) \times L^2(P_X, \mathcal{X}) \mid ||(h, \alpha) - (h_\Gamma^*, \alpha_\Gamma^*)||_\infty < \delta\}$. *Under Assumptions 1, 2, 3, 4, there exists* $0 < \delta < M^-$ *and positive constants* κ_1, κ_2 *such that* $\mathbb{E}_P\left[L_{RU}^\Gamma(h(X), \alpha(X), Y)\right]$ *is* κ_1*-strongly convex and* κ_2*-smooth in* (h, α) *on* $\mathcal{C}_\delta$, *where strong convexity and smoothness are defined using the norm on the product space* $L^2(P_X, \mathcal{X}) \times L^2(P_X, \mathcal{X})$. *In addition,* κ_1 *depends on constants* $C, P_{\min,\Gamma}, P_{\max}, \Gamma, M^-, M_\Gamma^+$, *and loss function L, and* κ_2 *depends on constants* $P_{\max}, D, \Gamma, M^-, M_\Gamma^+$, *and loss function L.*

C.2 Estimation Guarantees under Hölder-Smoothness

Definition 3. The class of p-Hölder smooth functions over $\mathcal{X} \subset \mathbb{R}^d$, $\Lambda^p(\mathcal{X})$, is defined as follows. Let β be a d-tuple of nonnegative integers, and set $|\beta|_1 = \beta_1 + \beta_2 + \cdots + \beta_d$. Let D^β denote the differential operator defined by $D^\beta = \partial^{|\beta|_1} / (\partial x_1^{\beta_1} \ldots \partial x_d^{\beta_d})$. Let $C^m(\mathcal{X})$ be the space of all m-times differentiable real-valued functions on $\mathcal{X}$. Let $p = m + \gamma$, where m is an integer $m \ge 0$ and $\gamma \in (0, 1]$. The Hölder space $\Lambda^p(\mathcal{X})$ consists of all functions $h \in C^m(\mathcal{X})$ for which the norm

$$||h||_{\Lambda^p(\mathcal{X})} = \sum_{|\beta|_1\le m} ||D^\beta h||_\infty + \sum_{|\beta|_1=m} \sup_{\substack{x,x'\in\mathcal{X},\\ x\neq x'}} \frac{|D^\beta h(x) - D^\beta h(x')|}{|x - x'|_2^\gamma}$$

is finite. Furthermore, the p-Hölder ball with radius c is $\Lambda_c^p(\mathcal{X}) = \{h \in \Lambda^p(\mathcal{X}) \mid ||h||_{\Lambda^p(\mathcal{X})} \le c\}$.

A first question we need to address is: If we assume that our estimation target $h_\Gamma^*(x)$ is p-smooth in the Hölder sense, what does this imply about the auxiliary parameter $\alpha_\Gamma^*(x)$ that emerged from our RU regression construction? Theorem C.1 showed that $\alpha_\Gamma^*(x)$ is a conditional quantile of the losses, and so any smoothness of α_Γ^* will depend on the smoothness of the conditional quantile function of the losses—which in turn depends on the smoothness of the conditional distribution of $Y|X$, the smoothness of the loss function ℓ, and the smoothness of h_Γ^*. It is thus not a-priori obvious that $\alpha_\Gamma^*(\cdot)$ should generally inherit regularity properties from $h_\Gamma^*(\cdot)$; however, as shown below, it does hold that if $h_\Gamma^*(\cdot)$ is p-smooth then $\alpha_\Gamma^*(\cdot)$ will also be p-smooth under mild additional assumptions.

Assumption 5. Let $\mathcal{Y} - \mathcal{Y} = \{y, y' \in \mathcal{Y} \mid y - y'\}$. We assume that the optimal robust predictor is smooth $h^*_\Gamma \in \Lambda^p_c(\mathcal{X})$, the loss function is smooth $\ell \in \Lambda^p_c(\mathcal{Y} - \mathcal{Y})$, and that the conditional outcome distribution is smooth $P_{Y|X} \in \Lambda^{p+1}_c(\mathcal{X} \times \mathcal{Y})$.

Lemma C.1. *Suppose Assumptions 1, 2, 3, 4, 5 hold. The optimal auxiliary function $\alpha^*_\Gamma \in \Lambda^p_{c'}(\mathcal{X})$ for some constant $c' > 0$ that depends on $p, c, d, M^-, M^+_\Gamma, P_{\min,\Gamma}, P_{\max}$, and the loss function L.*

This result motivates learning $h(\cdot)$ and $\alpha(\cdot)$ by running RU regression over a function class that can effectively represent p-smooth functions. The full class of p-smooth functions is an infinite dimensional space that is challenging to optimize over directly. For this reason, we instead consider the method of sieves [27], where we optimize the empirical risk over a sequence of finite-dimensional sieve spaces $\mathcal{H}_1 \times \mathcal{A}_1 \subseteq \cdots \subseteq \mathcal{H}_J \times \mathcal{A}_J \subseteq \ldots$, whose span provides increasingly sharp approximation to all p-smooth functions as the sieve index J increases. Empirical risk minimization over the sieve space can then be written as

$$(\hat{h}_n, \hat{\alpha}_n) \in \underset{(h,\alpha)\in\mathcal{H}_{J_n}\times\mathcal{A}_{J_n}}{\operatorname{argmin}} \widehat{\mathbb{E}}_P \left[L_{\text{RU}}(h(X), \alpha(X), Y) \right], \tag{26}$$

where J_n corresponds to the size of the sieve basis for a given sample size.

Standard choices of sieves for approximating smooth functions include polynomials and univariate splines [13]. For technical reasons, it is helpful to constrain our sieve functions to take values only within a bounded interval; and to accomplish this we follow the truncation strategy of [33]. Formal definitions of truncated polynomial and/or univariate spline sieves used in our analysis can be found online [58].

Obtaining the optimal rate of convergence for sieve estimation requires balancing the estimation error and sieve approximation error [13]. Estimation error is given by the error between the empirical RU risk minimizer in the sieve space and the population RU risk minimizer in the sieve space, and can be bounded using the metric entropy of the sieve space. Sieve approximation error is the error that arises from projecting the minimizer over the infinite-dimensional model space $(h^*_\Gamma, \alpha^*_\Gamma) \in \Lambda^p(\mathcal{X}) \times \Lambda^p(\mathcal{X})$ onto a finite-dimensional sieve. To get a handle on the sieve approximation error, our proofs adapt the result from [70] that

$$\inf_{(h,\alpha)\in\tilde{\mathcal{H}}_{J_n}\times\tilde{\mathcal{A}}_{J_n}} ||(h, \alpha) - (h^*_\Gamma, \alpha^*_\Gamma)||_\infty = O(J_n^{-p}),$$

where $\tilde{\mathcal{H}}_{J_n} \times \tilde{\mathcal{A}}_{J_n}$ denotes a (non-truncated) polynomial or univariate spline sieve.

Assumption 6. 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 $x \in \mathcal{X}$.

Assumption 7. We assume that $\sup_{x\in\mathcal{X}} \mathbb{E}_{P_{Y|X}} \left[Y^2 \mid X = x \right] < \infty$.

C.3 Extension to Constrained Function Classes

Thus far, we have focused on learning a robust decision rule without any constraints on the functional form of h. In some setting, however, it may be of interest to learn robust prediction rules with functional form constraints imposed on h (e.g., we want to find the best robust linear or tree-shaped predictor under our Γ-biased sampling model). In constrained function classes, population risk minimization is no longer equivalent to conditional risk minimization, because the optimal decision rule from the constrained class cannot perfectly minimize the risk conditionally for every $x \in \mathcal{X}$. However, the following results shows that a weighted minimizer of the RU loss, where the weights are given by the density ratio between the target and train covariate distribution, still identifies the optimal robust decision rules within a constrained function class.

Corollary C.1. *Suppose that $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ are drawn i.i.d. from P. Assume that $P_{Y|X=x}$ is absolutely continuous with respect to Lebesgue measure for all $x \in \mathcal{X}$. For any h,*

$$\sup_{Q\in\mathcal{S}_\Gamma(P,Q_X)} \mathbb{E}_Q \left[L(h(X), Y) \right] = \inf_{\alpha\in L^2(P_X,\mathcal{X})} \mathbb{E}_P \left[r(X) \cdot L^\Gamma_{\text{RU}}(h(X), \alpha(X), Y) \right], \tag{27}$$

where $r(x) = \frac{dQ_X(x)}{dP_X(x)}$ and Q_X is a distribution with the same support and $\sup_{x\in\mathcal{X}} \frac{dP_X(x)}{dQ_X(x)} < \infty$.

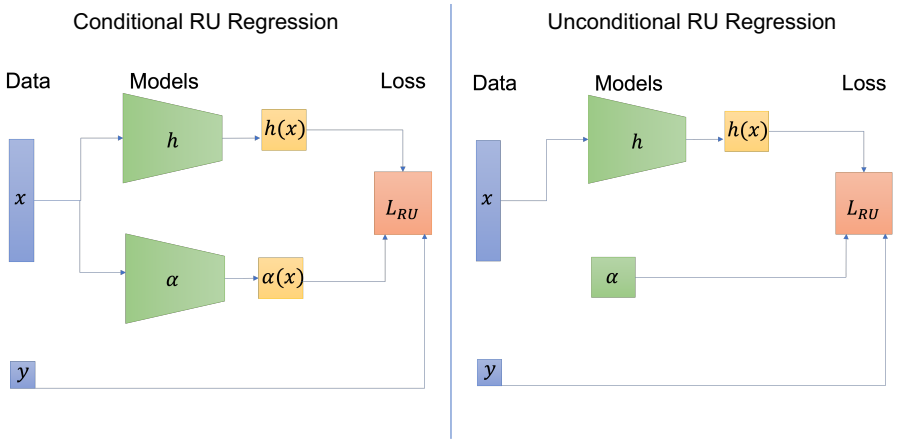

Figure 3: Model architecture for (Conditional) RU Regression and Unconditional RU Regression. Notably, conditional RU Regression requires fitting an auxiliary neural network $\alpha : \mathcal{X} \to \mathbb{R}$, while unconditional RU regression requires fitting a one-dimensional auxiliary parameter $\alpha \in \mathbb{R}$.

The above result suggests that if the covariate density ratio r is known and we aim to learn a robust decision rules from a function class $\mathcal{H}$, which may not necessarily be $L^2(P_X, \mathcal{X})$, then we can consider the following weighted risk minimization problem

$$\inf_{h \in \mathcal{H}} \sup_{Q \in \mathcal{S}_\Gamma(P, Q_X)} \mathbb{E}_P \left[r(X) \cdot L(h(X), Y) \right] = \inf_{(h,\alpha) \in \mathcal{H} \times L^2(P_X, \mathcal{X})} \mathbb{E}_P \left[r(X) \cdot L^\Gamma_{\mathrm{RU}}(h(X), \alpha(X), Y) \right]. \tag{28}$$

Thus, Weighted RU Regression can be applied to learning a robust decision rule from a constrained function class $\mathcal{H}$. Unlike the setting where the decision rule can take value in $L^2(P_X, \mathcal{X})$, when the decision rule is restricted to $\mathcal{H}$, the optimal robust decision rule is not agnostic to the target covariate distribution Q_X, and so Weighted RU Regression can only be applied if the target covariate distribution is identifiable. This limitation is inherent to any distributionally robust optimization approach that places restrictions on the conditional distribution $Y|X$ instead of the joint distribution (X, Y).

Remark 3. One subtle aspect of the above result is that even though we have constrained the function class that h comes from, Weighted RU Regression still requires optimizing α over a flexible class. One can check that α^*_Γ that minimizes the right side of (28) corresponds to a conditional quantile of the losses incurred under h^*_Γ. Restricting h to take value in a simple function class does not necessarily guarantee that the optimal α^*_Γ takes values in that class. Constraining the function class of α without making further assumptions on the data distribution may introduce bias due to misspecification.

C.4 Other experiments

C.5 Implementing RU Regression

We implement our baselines and proposed method using gradient-based optimization of neural networks [28]. From a statistical perspective, neural networks can be seen as a practical sieve-like methods that automate the selection of relevant basis functions [14, 23, 60]. 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 [54] to instantiate, train, validate, and test the neural networks. RU 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 $\alpha : \mathcal{X} \to \mathbb{R}$. The model architecture of RU Regression is visualized in the left side of Figure 3. Since RU Regression is a joint optimization problem over both (h, α), 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 (6). The third term of the RU loss depends on $(L(h(X), Y) - \alpha(X))_+$, which can be represented using the ReLU (rectified linear unit) activation function available in Pytorch. We compute the gradient of the RU loss with respect to h and α and update the parameters of both networks described above using the Adam optimizer [38]. We similarly implement Unconditional RU Regression using gradient-based optimization; the only implementation difference being that the auxiliary function $\alpha(X)$ in RU Regression is now replaced with an auxiliary parameter $\alpha \in \mathbb{R}$.

One limitation of using overparametrized neural networks to implement RU Regression is the potential for overfitting to the training data. Recent works have observed that it is possible for neural networks to interpolate the training data and obtain zero training loss. When the model can interpolate the training data, DRO approaches that explicitly or implicitly (like RU Regression) reweight the training data may not necessarily yield improved robustness because the worst-case risk on the training data also vanishes. To address this, [57] recommend coupling DRO with some form of regularization, such as early stopping or ℓ_2 regularization. In our experiments, we use early stopping. To implement early stopping, we hold out part of our training set as a validation set, evaluate the RU loss obtained on the validation set while training for a fixed number of epochs, and select the model that obtains the lowest RU loss on the validation set.

Remark 4. In Section C.3 we discussed the setting where h is constrained to only take values in a function class $\mathcal{H}$. In this setting, we can still use gradient-based optimization to solve the resulting Weighted RU Regression problem, as long as $\mathcal{H}$ has a tractable differentiable representation. For instance, when $\mathcal{H}$ is the class of linear models, we can represent h as a one-layer neural network and α using a neural network and jointly train both models with the Weighted RU loss.

C.6 Predicting Hospital Length of Stay

Accurate patient length-of-stay predictions are useful for scheduling and hospital resource management [29]. Many recent works study the problem of predicting patient length-of-stay from patient covariates [16, 48, 62]. In this setting, we evaluate the potential of RU regression for sampling-bias-robust length-of-stay prediction using a semi-synthetic experiment designed using the publicly available MIMIC-III dataset [34].

MIMIC-III has data on 19571 patients. The observed covariates X consist of 20 patient attributes (medical measurements and demographic characteristics) recorded within the first 24 hours of hospital stay. The outcome Y is the patient length-of-stay (LoS) in the ICU in days. Our semi-synthetic experiment involves resampling the original MIMIC-III dataset to introduce sampling bias. We then seek to use this biased data to learn a prediction rules that can predict Y with low mean-squared error on the original (unbiased) dataset.

More specifically, we start by splitting the original dataset into train, validation, and test sets consisting of 7045, 4697, and 7829 samples, respectively. We then resample both the train and validation sets with resampling weights $\pi_e : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ to generate distribution shifts; we resample with replacement such as to preserve the nominal sizes of both sets. We consider two weight functions for resampling,

$$\pi_1(x, y) \propto dQ_Y(y), \quad \pi_2(x, y) \propto \frac{1}{(dQ_Y(y))^2}.$$

Histograms of the marginal distribution over Y (LoS) in the target population and the biased training populations are given in Figure 4. Note that the weights π_e are not used by any learning algorithm, they are only used to generate the biased training data and compute evaluation metrics. We learn $h(\cdot)$ via the deep-learning based approach described above (without any covariate reweighting).

On the test set, we report two evaluation metrics. The first metric treats the original MIMIC-III dataset as coming from the true target distribution Q; when we report results under Q, we are effectively evaluating the ability of methods to compensate for the synthetic sampling bias introduced

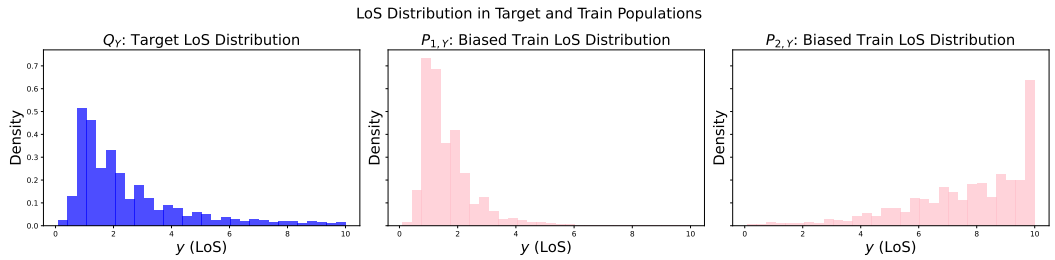

Figure 4: Marginal distributions over Y (LoS) in the target population and synthetic biased train populations.

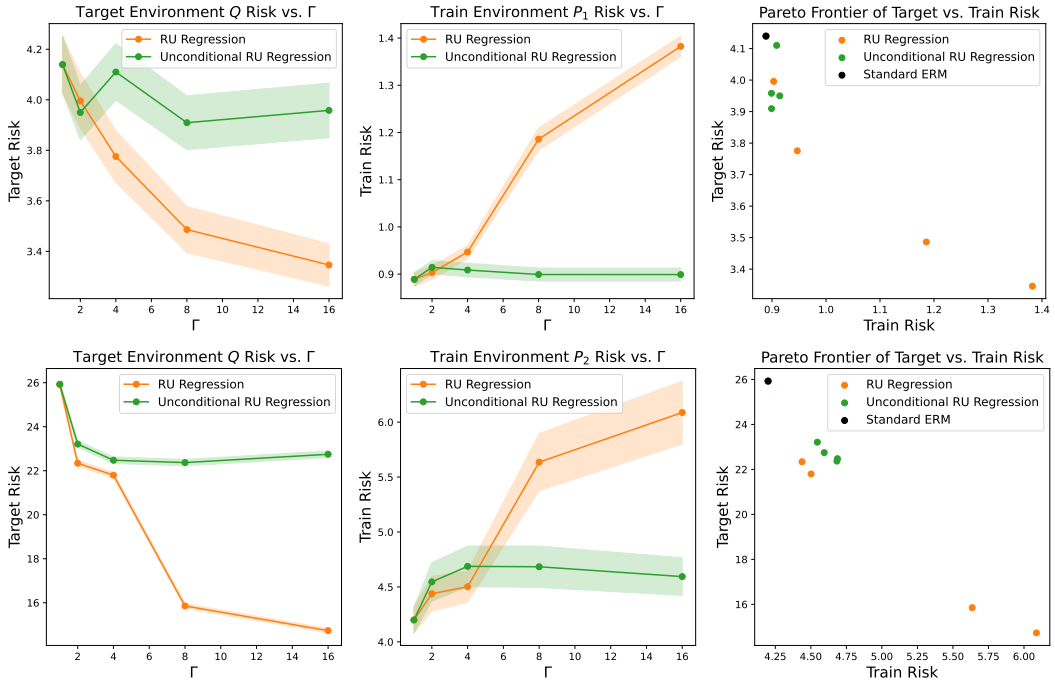

Figure 5: We evaluate RU Regression and Unconditional RU Regression when trained on biased training populations P_1, P_2. Bootstrap standard errors are computed with 5000 bootstrap samples.

by weighted resampling. The second evaluation metric applies weighting to the test set that imitates the sampling bias, thus providing us with an assessment of accuracy under with respect to the training environment P_e; this metric is what's targeted by empirical-risk minimization and other approaches that don't consider potential sampling bias.

We compute these metrics as

$$\begin{aligned}\text{Target Environment } Q \text{ Risk} &= \sum_{i=1}^{n_{\text{test}}} L(h_e(X_i), Y_i) \,/\, n_{\text{test}} \\ \text{Train Environment } P_e \text{ Risk} &= \sum_{i=1}^{n_{\text{test}}} \pi_e(X_i, Y_i) \cdot L(h_e(X_i), Y_i) \,/\, \sum_{i=1}^{n_{\text{test}}} \pi_e(X_i, Y_i).\end{aligned} \tag{29}$$

In all cases, we use squared-error loss for $L(\cdot)$, i.e., we target mean-squared error.

We compare the train and target risk obtained by RU Regression and Unconditional RU Regression models trained on P_1, P_2. In the left and middle plots of Figure 5, we find that as Γ increases, RU Regression has increasing training risk and decreasing target risk. In contrast, as Γ increases, the Unconditional RU Regression model's train risk is relatively constant and its target risk decreases

modestly, even though the worst-case risk increases with Γ. In the right plots, we plot the Pareto frontier between train and target risk for RU Regression and Unconditional RU Regression models for $\Gamma = 2, 4, 8, 16$. We find that RU Regression trades off performance on the training environment for improved target risk, meanwhile Unconditional RU Regression behaves similarly to the Standard ERM model on this frontier. We hypothesize Unconditional RU Regression exhibits this behavior because it implicitly upweights training samples from "hard to learn" regions of the covariate-outcome space, where no model can perform well. Practically, this results in the Unconditional RU Regression model behaving similarly to a model fit via standard ERM in the remaining regions of the covariate space.

D Discussion

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, which is a key limitation of our work. 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 [32], 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.

In other settings, we may have to select Γ without access to any target conditional distribution. In the absence of data from any target conditional distribution, we can only view Γ as a sensitivity parameter that is postulated by the researcher. While there is no true value, we can follow the approaches of [53] and [15] to benchmark Γ using the distribution shift of observables.

Finally, we note that it is interesting to consider how our results relate to the 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 [12, 31], or was chosen by an adversary [11, 17, 41]. 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.