arXiv:1810.08750v2 [stat.ML] 20 Nov 2018

Learning Models with Uniform Performance via Distributionally Robust Optimization

John C. Duchi[1] Hongseok Namkoong[2]

Stanford University

Departments of [1]Statistics, [1]Electrical Engineering, and [2]Management Science and Engineering

{jduchi,hnamk}@stanford.edu

Abstract

A common goal in statistics and machine learning is to learn models that can perform well against distributional shifts, such as latent heterogeneous subpopulations, unknown covariate shifts, or unmodeled temporal effects. We develop and analyze a distributionally robust stochastic optimization (DRO) framework that learns a model that provides good performance against perturbations to the data-generating distribution. We give a convex optimization formulation for the problem, providing several convergence guarantees. We prove finite-sample minimax upper and lower bounds, showing that distributinoal robustness sometimes comes at a cost in convergence rates. We give limit theorems for the learned parameters, where we fully specify the limiting distribution so that confidence intervals can be computed. On real tasks including generalizing to unknown subpopulations, fine-grained recognition, and providing good tail performance, the distributionally robust approach often exhibits improved performance.

1 Introduction

In many applications of statistics and machine learning, we wish to learn models that achieve uniformly good performance over almost all input values. This is important for safety- and fairness-critical systems such as medical diagnosis, autonomous vehicles, criminal justice and credit evaluations, where poor performance on the tails of the inputs leads to high-cost system failures. Methods that optimize average performance, however, often produce models that suffer low performance on the "hard" instances of the population. For example, standard regressors obtained from maximum likelihood estimation can lose their predictive power on certain regions of covariates [57], so that high average performance comes at the expense of low performance on minority subpopulations. In this work, we propose and study a procedure that explicitly optimizes performance on tail inputs that suffer high loss.

Modern datasets incorporate heterogeneous (but latent) sub-populations, and a natural goal is to perform well across all of these [57, 65, 21]. While many statistical models show strong average performance, their performance often deteriorates on minority groups underrepresented in the dataset. For example, speech recognition systems are inaccurate for people with minority accents [4]. In numerous other applications—such as facial recognition, automatic video captioning, language identification, academic recommender systems—performance varies significantly over different demographic groupings, such as race, gender, or age [38, 42, 18, 68, 76].

In addition to latent heterogeneity in the population, distributional shifts in covariates [71, 8] or unobserved confounding variables (e.g. unmodeled temporal effects [39]) can contribute to changes in the data generating distribution. Performance of machine learning models degrades significantly on domains that are different from what the model was trained on [39, 17, 27, 67, 77]. Domain adaptation [71, 8, 9] and multi-task learning methods [24] can be effective in situations where

(potentially unlabeled) data points from the target domain are available. The reliance on *a priori* fixed target domains, however, is restrictive, as the shifted target distributions are usually unknown before test time and it is impossible to collect data from the targets.

To mitigate these challenges, we consider unknown distributional shifts, developing and analyzing a loss minimization framework that is explicitly robust to local changes in the data-generating distribution. Concretely, let $\Theta \subseteq \mathbb{R}^d$ be the parameter (model) space, P_0 be the data generating distribution on the measure space $(\mathcal{X}, \mathcal{A})$, X be a random element of $\mathcal{X}$, and $\ell : \Theta \times \mathcal{X} \to \mathbb{R}$ be a loss function. Rather than minimizing the average loss $\mathbb{E}_{P_0}[\ell(\theta; X)]$, we study the *distributionally robust* stochastic optimization problem

$$\underset{\theta \in \Theta}{\text{minimize}} \left\{ \mathcal{R}_f(\theta; P_0) := \sup_{Q \ll P_0} \left\{ \mathbb{E}_Q[\ell(\theta; X)] : D_f\left(Q \| P_0\right) \le \rho \right\} \right\}, \tag{1}$$

where the hyperparameter $\rho > 0$ modulates the size of the distributional shift. Here,

$$D_f\left(Q \| P_0\right) := \int f\left(\frac{dQ}{dP_0}\right) dP_0$$

is the f-divergence [3, 26] between Q and P_0, where $f : \mathbb{R}_+ \to \overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\}$ is a convex function satisfying $f(1) = 0$ and $f(t) = +\infty$ for any $t < 0$.

The worst-case risk (1) upweights regions of $\mathcal{X}$ with high losses $\ell(\theta; X)$. Consequently, the worst-case formulation (1) optimizes performance on the tails, as measured by the loss on "hard" examples. Thus, in our motivating scenarios of distribution shift or latent subpopulations, as long as the alternative distribution remains ρ-close to the data-generating distribution P_0, the model $\theta^\star \in \Theta$ that minimizes the worst-case formulation (1) guarantees reasonable performance across such a priori unknown perturbations.

While the motivation of robustness to underlying distributional shifts is appealing, the formulation (1) is not explicit about appropriate choices for f, which leaves nontrivial modeling freedom. In some situations (see Section 2), we can give explicit calculations suggesting appropriate choices of f. Given the challenge of characterizing the "right" choices of f, we begin our study in Section 3 with experiments to substantiate the intuition that the worst-case formulation (1) hedges against heterogeneous subpopulations, covariate shift, and other latent confounding. This motivates our subsequent theoretical study. We view characterizing the "right" choices of f for different scenarios as an important open question. Letting $\widehat{P}_n$ denote the empirical measure on $X_i \overset{\text{iid}}{\sim} P_0$, our approach to minimizing objective (1) is via the plug-in estimator

$$\widehat{\theta}_n \in \underset{\theta \in \Theta}{\text{argmin}} \left\{ \mathcal{R}_f(\theta; \widehat{P}_n) := \sup_{Q \ll \widehat{P}_n} \left\{ \mathbb{E}_Q[\ell(\theta; X)] : D_f(Q \| \widehat{P}_n) \le \rho \right\} \right\}. \tag{2}$$

Our experimental and theoretical work demonstrates that the distributionally robust estimator $\widehat{\theta}_n$ improves performance on the tails of the data-generating distribution and provides better performance uniformly across the input space $\mathcal{X}$.

The main theoretical component of this work is to study finite sample and asymptotic properties of the plug-in estimator (2). We first provide an efficiently minimizable (finite-dimensional) dual formulation (Section 2). We give convergence guarantees for the plug-in estimator (2) (Section 4), and prove that it is rate optimal (Section 5), thereby providing finite-sample minimax bounds on the optimization problem (1). Because the formulation (1) protects against gross departures from the average loss, we observe a degradation in (worst-case) rates that is effectively a consequence

of needing to estimate high moments of random variables. More quantitatively, our convergence guarantees show that for f-divergences with $f(t) \asymp t^k$ as $t \to \infty$, where $k \in (1, \infty)$, the empirical minimizer $\widehat{\theta}_n$ satisfies

$$\mathcal{R}_f(\widehat{\theta}_n; P_0) - \inf_{\theta \in \Theta} \mathcal{R}_f(\theta; P_0) = O_P\left(n^{-\frac{1}{k_* \vee 2}}\right),$$

where $k_* = \frac{k}{k-1}$ (Section 4). We provide minimax lower bounds matching these rates in n.

These worst-case (minimax) guarantees do not necessarily reflect the typical behavior of the estimators, so we complete our theoretical analysis in Section 6 with an asymptotic analysis. The estimator $\widehat{\theta}_n$ is consistent under mild (and standard) regularity conditions (Section 6.1). Under suitable differentiability conditions on $\mathcal{R}_f$, $\widehat{\theta}_n$ is asymptotically normal at the typical $\sqrt{n}$-rate, allowing us to obtain calibrated confidence intervals (Section 6.2).

Related Work Distributional shift problems arise in many guises across statistics, machine learning, and optimization; we give a necessarily abridged survey of the many strains of work and their repsective foci. Work in domain adaptation seeks models that receive data from one domain and are tested on a specified target; typical approach is to reweight the distribution P_0 to make it "closer" to the known target distribution P_{target} [71, 43, 14, 73, 74, 78]. In this vein, one interpretation of the worst-case formulation (1) is as importance-weighted loss minimization without a known target domain—that is, without assuming even unlabeled data from the target domain. The formulation (1) is more conservative than most domain adaptation methods, as it considers shifts in the joint distribution of predictors X and target variable Y instead of covariate shifts.

Other scenarios naturally give rise to structural distributional changes. Time-varying effects are a frequent culprit [39], and time-varying-coefficient models are effective when time indices are available [35, 22]. When one believes there may be latent subpopulations, mixture model approaches can model latent membership directly [2, 36, 56, 23]. In contrast, our worst-case approach (1) does not directly represent (or require) such latent information, and—especially in the case of mixture models—can maintain convexity because of the focus on uniform performance guarantees.

When we know and can identify heterogeneous populations within the data (i.e. the data-generating mechanism explicitly provides group identities), Bühlmann, Meinshausen, and colleagues, connect methods that achieve good performance on all subpopulations with causal interventions. In this vein, Meinshausen and Bühlmann [57] study maximin effects on heterogeneous datasets and learn linear models that maximize (relative) performance over the worst (observed) subgroup, which has connections to minimax regret in linear models [33, 10, 65, 21]. By studying additive (worst-case) perturbations to covariate vectors, Rothenhäusler et al. [66] learn linear structural equation models. Without access to information about particular subpopulations, our more worst-case formulation (1) is likely more conservative than their approaches, but (as we see in our experimental evaluation) still achieves good performance.

In the optimization literature, there is a substantial body of work on distributionally robust optimization problems. A number of authors study worst-case regions arising out of moment conditions on the data vector X [29, 47, 13]. Other work [11, 32, 59, 51, 52] studies a scenario similar to our f-divergence formulation (1). In this line of research, the empirical plug-in procedure (2) with radius ρ/n provides a finite sample confidence set for the population objective $\mathbb{E}_{P_0}[\ell(\theta; X)]$; the focus there is on the true distribution P_0 and does not consider distributional shifts. Duchi et al. [32] and Lam and Zhou [52] show how such approximations correspond to generalized empirical likelihood [60] confidence bounds on $\mathbb{E}_{P_0}[\ell(\theta; X)]$.

An alternative to our f-divergence based sets—that is, $\{Q : D_f(Q\|P_0) \le \rho\}$—are Wasserstein distance balls [85, 61, 86, 69, 15, 16, 34, 72]. Such approaches are satisfying as Wasserstein distances W satisfy $W(\widehat{P}_n, P_0) \to 0$ as $n \to \infty$, providing direct certificates; Wasserstein balls also allow worst-case distributions with different support from the data-generating distribution P_0. This power, however, means that tractable reformulations are only available under restrictive scenarios [69, 34, 72], and they remain computationally challenging. Most guarantees [16, 34, 69] for these problems also consider approximation only of the canonical (population) loss $\mathbb{E}_{P_0}[\ell(\theta; X)]$, and the (necessarily) slow convergence of Wasserstein distances [37] means that the best rates of convergence are $O(n^{-1/d})$, prohibitively slow for most applications.

Notation For a sequence of random variables $Z_1, Z_2, \ldots$ in a metric space $\mathcal{Z}$, we say $Z_n \overset{d}{\rightsquigarrow} Z$ if $\mathbb{E}[h(Z_n)] \to \mathbb{E}[h(Z)]$ for all bounded continuous functions h. We write $Z_n \xrightarrow{p} Z$ for convergence in probability. For some space $\mathcal{Z}$, we let $\ell^\infty(\mathcal{Z})$ the space of bounded real-valued functions on $\mathcal{Z}$, equipped with the supremum norm. We let $D_{\chi^2}(Q\|P) = \frac{1}{2}\int (dQ/dP - 1)^2 \, dP$ be the χ^2-divergence between Q and P. For a random variable $Z \sim P$, we let $\operatorname{ess\,sup}_P Z$ denote its essential supremum under P. We make the dependence on the underlying measure explicit when we write expectations (e.g. $\mathbb{E}_P[X]$), except for when the base distribution is P_0. For $k \in (1, \infty)$, we let $k_* := k/(k-1)$. The Frobenius norm of a matrix A is $\|A\|_{\mathrm{F}}$. We write $\nabla\ell(\theta; X)$, where we always mean differentiation with respect to the parameter vector $\theta \in \mathbb{R}^d$.

2 Formulation

We begin our discussion by presenting dual reformulations for the worst-case objective $\mathcal{R}_f(\theta; P_0)$. The dual form gives a convex procedure for computing the empirical plug-in estimator (2) and makes explicit the role that $t \mapsto f(t)$ plays in defining such a *risk-averse* version of the usual average loss $\mathbb{E}_{P_0}[\ell(\theta; X)]$. By taking the dual of the inner supremum in the distributionally robust problem (1), we obtain a single convex minimization problem in place of the original minimax formulation. We draw upon the dual form heavily in the rest of this paper, both for statistical and computational reasons. Using the likelihood ratio $L(x) := dQ(x)/dP_0(x)$ to reformulate our distributionally robust problem (1), the worst-case objective $\mathcal{R}_f(\theta; P_0)$ is

$$\mathcal{R}_f(\theta; P_0) = \sup_{L \ge 0} \left\{ \int_{\mathcal{X}} \ell(\theta; x) L(x) dP_0(x) \mid \int f(L(x)) dP_0(x) \le \rho, \; \mathbb{E}_{P_0}[L(X)] = 1 \right\}, \tag{3}$$

where the supremum is over measurable functions.

We now provide a dual reformulation of the quantity (3). Ben-Tal et al. [11] present an identical result for discrete (finitely supported) distributions, and we extend it to the infinite dimensional setting here. Recall that f^* denotes the Fenchel conjugate $f^*(s) := \sup_t\{st - f(t)\}$. Without further comment, we always treat the perspective transformation of f^* as

$$\lambda f^*\left(\frac{t}{\lambda}\right) = \begin{cases} \lambda f^*\left(\frac{t}{\lambda}\right) & \text{if } \lambda > 0 \\ \mathbf{I}\left(t \le 0\right) & \text{if } \lambda = 0 \\ +\infty & \text{if } \lambda < 0, \end{cases} \tag{4}$$

which is equal to the standard closure of the perspective plus the closed convex function $t \mapsto \mathbf{I}\left(t \le 0\right)$, so that it is closed [41, Prop. IV.2.2.2]. With this definition, we have the following duality result, whose proof we provide in Section A.1.

Proposition 1. *Let P be an arbitrary probability measure on $(\mathcal{X}, \mathcal{A})$. Then, for any $\rho > 0$, we have for all $\theta \in \Theta$*

$$\sup_{Q \ll P} \{\mathbb{E}_Q[\ell(\theta; x)] : D_f(Q\|P) \le \rho\} = \inf_{\lambda \ge 0, \eta \in \mathbb{R}} \left\{ \mathbb{E}_P \left[\lambda f^* \left(\frac{\ell(\theta; X) - \eta}{\lambda} \right) \right] + \lambda\rho + \eta \right\}. \tag{5}$$

Moreover, if the supremum on the left hand side is finite, there are finite $\lambda(\theta) \ge 0$ and $\eta(\theta) \in \mathbb{R}$ attaining the infimum on the right hand side.

For convex losses $\theta \mapsto \ell(\theta; X)$, the dual form (5) is jointly convex in (θ, η, λ). While interior point methods [20] are powerful tools for solving such problems, they may be slow in settings where n, the sample size, and d, the dimension of $\theta \in \Theta$, are large. More direct methods can directly solve the primal form, including gradient descent or stochastic gradient algorithms [58, 59].

Cressie-Read and Rényi-divergence families Much of our development in this paper centers around a general family of divergences, known as the Cressie-Read or Rényi divergences, with applications in statistics and information theory. The *Rényi α-divergence* [84] between distributions P and Q is

$$D_\alpha(P\|Q) := \frac{1}{\alpha - 1} \log \int \left(\frac{dP}{dQ} \right)^\alpha dQ,$$

with the limit as $\alpha \to 1$ satisfies $D_1(P\|Q) = D_{\rm kl}(P\|Q)$. These form a natural collection of measures of divergence of probability distributions; for analytical reasons, we use the equivalent Cressie-Read family of f-divergences [25]. These are parameterized by $k \in (-\infty, \infty) \setminus \{0, 1\}$, setting

$$f_k(t) := \frac{t^k - kt + k - 1}{k(k-1)} \quad \text{so} \quad f_k^*(s) := \frac{1}{k} \left[((k-1)s + 1)_+^{k_*} - 1 \right], \tag{6}$$

where $k_* = \frac{k}{k-1}$. We let $f_k(t) = +\infty$ for $t < 0$, and we define f_1 and f_0 as their respective limits as $k \to 0, 1$. The family of divergences (6) includes χ^2-divergence ($k = 2$), empirical likelihood $f_0(t) = -\log t + t - 1$, and KL-divergence $f_1(t) = t \log t - t + 1$. We let

$$\mathcal{R}_k(\theta; P) := \sup_{Q \ll P} \{\mathbb{E}_Q[\ell(\theta; X)] : D_{f_k}(Q\|P) \le \rho\} \tag{7}$$

be the obective associated with f_k. While most of our results generalize to other values of k, we focus temporarily on $k \in (1, \infty)$ for ease of exposition. By minimizing out $\lambda \ge 0$ in the original dual form (5), we obtain a simplified dual formulation for the Cressie-Read family (6):

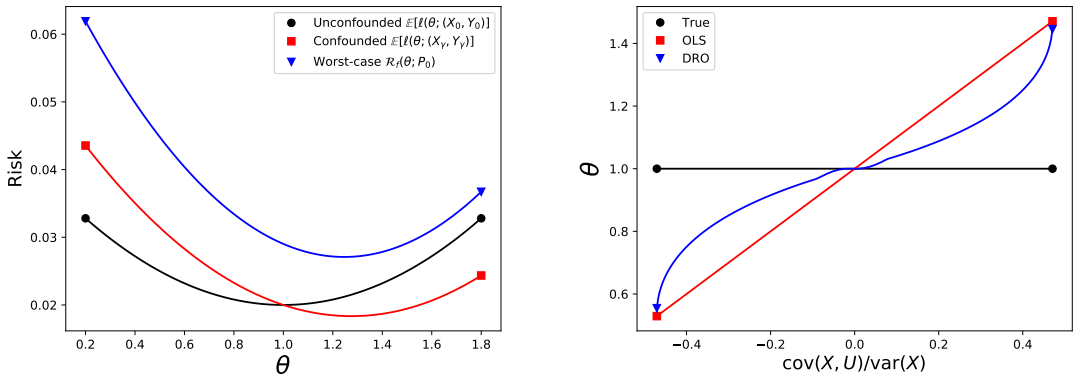
Lemma 1. *Let P be an arbitrary probability measure on $(\mathcal{X}, \mathcal{A})$. Then, for $k \in (1, \infty)$ and $k_* = k/(k-1) \in (1, \infty)$, and any $\rho > 0$, we have for all $\theta \in \Theta$*

$$\mathcal{R}_k(\theta; P) = \inf_{\eta \in \mathbb{R}} \left\{ c_k(\rho) \mathbb{E}_P \left[(\ell(\theta; X) - \eta)_+^{k_*} \right]^{\frac{1}{k_*}} + \eta \right\}. \tag{8}$$

where $c_k(\rho) := (k(k-1)\rho + 1)^{\frac{1}{k}}$.

We draw upon the lemma heavily in following sections; see Section A.2 for the proof.

The simplified dual form (8) shows that the worst-case objective $\mathcal{R}_k(\theta; P)$ only penalizes losses above some value $\eta^\star$. The $L^{k_*}(P)$-norm upweights these tail values of $\ell(\theta; x)$, giving a worst-case objective that focuses on "hard" regions of $\mathcal{X}$. The dual form (8) also makes explicit the relationship between the growth of $t \mapsto f_k(t)$ and the worst-case objective $\mathcal{R}_k(\theta; P)$: as f_k grows for large t



(a) Risk at $\gamma = .3$ (b) Parameters

Figure 1. Regression under confounding ($\sigma = .2$). We use f_2-divergence and choose ρ as given in expression (9) with ($k = 2$). Figure (a) plots the unconfounded risk $\mathbb{E}[\ell(\theta; (W_0, Y))] = \frac{1}{2}\mathbb{E}[(Y - \theta W_0)^2]$, confounded risk $\mathbb{E}[\ell(\theta; (W_\gamma, Y))] = \frac{1}{2}\mathbb{E}[(Y - \theta W_\gamma)^2]$ used by OLS, and the distributionally robust risk $\mathcal{R}_{f_2}(\theta; P_{0,\gamma})$ for $\gamma = .3$. Figure (b) plots the true parameter $\theta_{\rm true} = 1$, the OLS estimate $\theta_{\rm OLS} = 1 + \frac{\mathrm{Cov}(W,U)}{\mathrm{Var}(W)}$, and the DRO solution (1) while varying $\frac{\mathrm{Cov}(W,U)}{\mathrm{Var}(W)} = \frac{\gamma}{1+\gamma^2}$ (x-axis).

($k \uparrow \infty$), the f-divergence ball $\{Q : D_f(Q\|P) \le \rho\}$ shrinks, and the risk measure $\mathcal{R}_k(\theta; P)$ becomes less conservative (smaller). The dual form (8) quantifies this with the $L^{k_*}(P)$-norm of the loss above the quantile η determines.

There is a well-developed literature on coherent risk measures that define negative utility functions giving "sensible" risk preference [5, 63, 50, 70], and our worst-case objective is a coherent risk measure. In this sense, the distributionally robust problem (1) is a risk-averse formulation of the canonical stochastic optimization problem of minimizing $\mathbb{E}_{P_0}[\ell(\theta; X)]$. Indeed, Krokhmal [50] proposes the dual form (8) as a higher order generalization of the classical conditional value-at-risk [63], which corresponds to $\mathcal{R}_k(\theta; P)$ defined with $k = \infty$ (or $k_* = 1$) in our notation.

2.1 Examples

While—as we note in the introduction—we do not provide precise recommendations for the choice of f-divergence, it is instructive to consider a few examples for motivation. We begin with a generic description, specializing more in the final two.

Example 1 (Generic distributional shift)**:** Consider data in pairs (X, Y), where X is a feature (covariate) vector and Y a dependent variable (e.g. label) we wish to model from X. Let U be a latent (unobserved) confounding variable, and assume that the pair (X, Y) jointly follows $P_0(\cdot \mid U = u)$. For a probability measure μ on U, let $P_\mu((X, Y) \in A) := \int P_0((X, Y) \in A \mid U = u) d\mu(u)$. In this case, we have the essentially tautological one-to-one correspondence

$$\{P \mid D_f(P\|P_0) \le \rho\} = \left\{P_\mu \mid \int f\left(\frac{dP_\mu(x,y)}{dP_0(x,y)}\right) dP_0(x,y) \le \rho\right\}.$$

That is, the robustness set consists of a family of distributional interventions on the hidden variable. We leave characterizing the precise form of such interventions as an open question. $\diamond$

Example 2 (Regression under confounding)**:** Consider a structural equation model $Y = X + U$,

where we wish to predict Y from X (or intervene on X to increase Y), but U is an unobserved confounder. To make this concrete, assume that

$$X = X_\gamma = \gamma U + Z \quad \text{where} \quad U, Z \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2).$$

Although our data comes from the γ-confounded distribution $(X_\gamma, Y) \sim P_{0,\gamma} =: P_0$, our goal is to learn the "true" relationship $\theta_{\text{true}} = 1$ between Y and X_γ, that is, the effect on Y if we intervene and set X to a particular value. The ordinary least squares (OLS) estimator suffers the bias

$$\theta_{\text{ols}} = \theta_{\text{true}} + \frac{\text{Cov}(X_\gamma, U)}{\text{Var}(X_\gamma)} = 1 + \frac{\text{Cov}(X_\gamma, U)}{\text{Var}(X_\gamma)} = 1 + \frac{\gamma}{1+\gamma^2}.$$

Now we consider the accuracy of the distributionally robust formulation—with squared loss $\ell(\theta; (x, y)) = \frac{1}{2}(y - \theta x)^2$—when the uncertainty region includes the unconfounded model $Y = X$. To that end, recall the Cressie-Read divergence (6), $f_k(t) = \frac{t^k - kt + k - 1}{k(k-1)}$, with $k \in (1, \infty)$. The f_k-divergence between the data-generating confounded distribution $(X_\gamma, Y) \sim P_{0,\gamma}$ and the unconfounded distribution $(X_0, Y) \sim P_{0,0}$ is

$$D_{f_k}\left(P_{0,0} \| P_{0,\gamma}\right) = \frac{1}{k(k-1)} \left((1 - 2(k-1)\gamma^2)^{-\frac{1}{2}} - 1 \right) \quad \text{for} \quad |\gamma| < \frac{1}{\sqrt{2(k-1)}}. \tag{9}$$

Setting $\rho = D_{f_k}(P_{0,0} \| P_{0,\gamma})$, the worst-case risk $\mathcal{R}_{f_k}(\theta; P_{0,\gamma})$ then upper bounds the unconfounded risk $\frac{1}{2}\mathbb{E}[(Y - \theta X_0)^2]$ (see Figure 1(a)). In Figure 1(b), we show the distributionally robust solution $\theta_{\text{dro}} = \operatorname{argmin} \mathcal{R}_{f_k}(\theta; P_{0,\gamma})$ as γ, as well as the OLS solution θ_{ols} and the "true" value $\theta_{\text{true}} = 1$. By accounting for certain confounding mixture possibilities, as in Example 1, the robust solution estimates θ_{true} more accurately. $\diamond$

Example 3 (Worst-case minority performance and CVaR)**:** For $c > 1$, the conditional value-at-risk [63] (CVaR) is

$$\text{CVaR}_c(\theta; P_0) := \inf_{\eta \in \mathbb{R}} \left\{ c\mathbb{E}_{P_0}\left[(\ell(\theta; X) - \eta)_+ \right] + \eta \right\} = \mathcal{R}_{f_{\infty,c}}(\theta; P_0),$$

where $f_{\infty,c}$ is given by $f_{\infty,c}(t) = 0$ if $0 \le t \le c$ and $f_{\infty,c}(t) = \infty$ otherwise. In this case, a calculation [70, Example 6.19] shows that

$$\begin{aligned} \mathcal{P} &:= \left\{ P \mid D_{f_{\infty,c}}(P \| P_0) \le \rho \right\} \\ &= \left\{ P \mid \text{ there exists } Q,\ \alpha \in [1/c, 1] \text{ s.t. } P_0 = \alpha P + (1-\alpha) Q \right\}. \end{aligned}$$

That is, the uncertainty set exactly corresponds to distributions with minority sub-populations of size at least $1/c$, and thus $\text{CVaR}_c(\theta; P_0) = \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\theta; X)]$ is the expected loss of the worst $1/c$-sized subpopulation. $\diamond$

3 Empirical Analysis

As this paper proposes and argues for alternatives to empirical risk minimization and standard M-estimation—the workhorses of much of machine learning and statistics [82, 83, 45]—it is important that we justify our approach. To that end, we first provide a number of experiments that illustrate

the empirical properties of the distributionally robust formulation (1). We test our plug-in estimator (2) on a variety of tasks involving real and simulated data, and compare its performance with the standard empirical risk minimizer

$$\widehat{\theta}_n^{\rm erm} \in \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}_{\widehat{P}_n}[\ell(\theta; X)].$$

For concreteness, we focus on the Cressie-Read (equivalently Rényi) family (6) of divergences with $k \in (1, \infty)$. In our experiments, we focus on three related challenges for which we believe the divergence-based approach (2) is well-suited:

1. Domain adaptation and distributional shifts, in which we fit predicotrs on a training distribution differing from the test

2. Performance on tail losses, where we measure quantiles of a model's loss rather than its expected losses

3. Data coming from multiple heterogeneous subpopulations, where we study performance on each subpopulation (or worst-case subpopuilations).

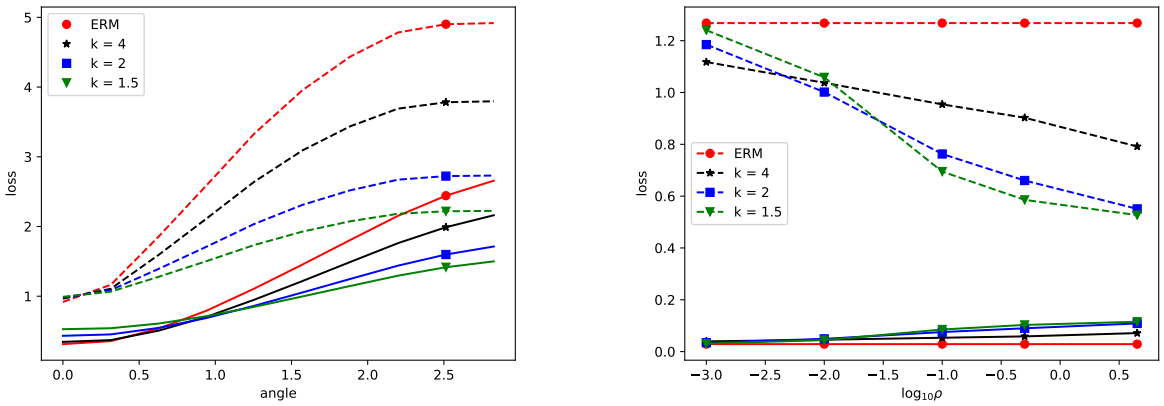
If our intuition on the development of the distributionally robust risk is accurate, we would expect results of roughly the following form: as we decrease k in the Cressie-Read divergence (6), $f_k(t) \propto t^k - 1$, we expect the solutions to exhibit more robustness while trading against empirical performance, as the set $\{P : D_f(P\|P_0) \le \rho\}$ gets larger. Thus, such models should have better tail behavior or generalization on difficult sub-populations. We expect increasing ρ to exhibit similar effects. We shall see the ways this intuition bears out in our experiments.

We begin with simulation experiments that touch on all three of our challenges in Section 3.1. We then investigate the three challenges on different real-world datasets. We begin in Section 3.2 by studying domain adaptation in the context of predictors trained to recognize handwritten digits, then tested on typewritten digit recognition tasks. In Section 3.3, we study tail prediction performance in a crime prediction problem. In our final experiment, in Section 3.4, we study a fine-grained recognition problem, where a classifier must label images as one of 120 different dog breeds; this highlights a combination of items 2 and 3 on tail performance and sub-population 2performance.

To efficiently solve the empirical worst-case problem (2) for the Cressie-Read family (6), we employ two approaches. For small datasets (small n and d), we solve the dual form (8) directly using a (conic) interior point solver; we extended the open-source Julia package `convex.jl` to implement power cone solvers [80] (the package now contains our implementation). For larger datasets (e.g. $n \approx 10^3 - 10^5$ and $d \approx 10^2 - 10^4$), we apply gradient descent with backtracking Armijo line-searches [20]. The probability vector $Q^* = \{q_i^*\}_{i=1}^n \in \mathbb{R}_+^n$ achieving the supremum in the definition (7) is unique as long as the loss vector $[\ell(\theta; X_i)]_{i=1}^n$ is non-constant, which it is in all of our applications, so $\mathcal{R}_k$ is differentiable [41, Theorem VI.4.4.2] with

$$\nabla\mathcal{R}_k(\theta, \widehat{P}_n) = \sum_{i=1}^n q_i^* \nabla\ell(\theta; X_i) \quad \text{where} \quad Q^* = \operatorname*{argmax}_{Q: D_{f_k}(Q\|\widehat{P}_n)\le\rho} \left\{ \sum_{i=1}^n q_i \ell(\theta; X_i) \right\}. \tag{10}$$

We use a fast bisection method [59] to compute Q^* at every iteration of our first-order method; the bisection code is publicly available at `https://github.com/hsnamkoong/robustopt`.



(a) Classification (b) Regression

Figure 2. (a) Hinge losses (average and 90th percentile in solid and dashed lines, respectively) under distributional shifts from $\theta_0^\star$ to $\theta_t^\star = \theta_0^\star \cdot \cos t + v \cdot \sin t$. The horizontal axis indexes perturbation t. (b) Losses on minority group (real-line) and majority group (dotted-line) under the distribution (12). We define the minority group as those with $X^1 \le z_{.95}$.

3.1 Simulation

Our first experiments use simulated data, where we fit linear models both for binary classification and prediction of a real-valued signal. We train our models with different values of f-divergence power k and tolerance ρ, testing them on perturbations of the data-generating distribution.

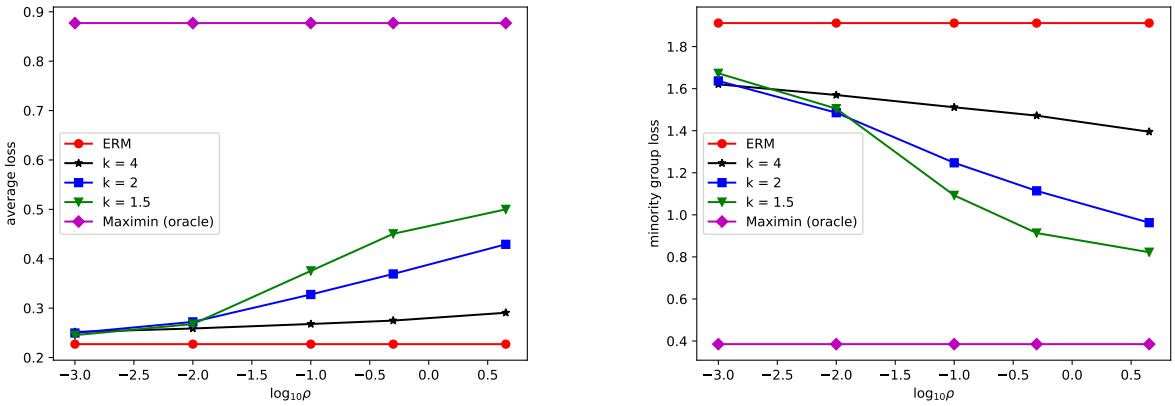
3.1.1 Domain adaptation and distributional shifts

We investigate distributional shifts via a binary classification experiment using the hinge loss $\ell(\theta; (x, y)) = \left(1 - yx^\top\theta\right)_+$, where $y \in \{\pm 1\}$ and $x \in \mathbb{R}^d$ with $d = 5$. In this case, fix a vector $\theta_0^\star \in \mathbb{R}^5$, chosen uniformly on the unit sphere, and generate training data

$$X \stackrel{\text{iid}}{\sim} \mathsf{N}(0, I_d) \quad \text{and} \quad Y \mid X = \begin{cases} \text{sign}(X^\top\theta_0^\star) & \text{w.p. } 0.9 \\ -\text{sign}(X^\top\theta_0^\star) & \text{w.p. } 0.1. \end{cases} \tag{11}$$

We train our models on $n_{\text{train}} = 100$ training data points, where we use $\rho = .5$ and vary values of $k \in \{1.5, 2, 4\}$ for our distributionally robust procedure (2). To simulate distributional shift, we take a uniformly random vector $v \perp \theta_0^\star$, $v \in \mathbb{S}^{d-1}$, and for $t \in [0, \pi]$ define $\theta_t^\star = \theta_0^\star \cdot \cos t + v \cdot \sin t$, so that $\theta_\pi^\star = -\theta_0^\star$. For each perturbation, we generate $n_{\text{test}} = 100,000$ test examples using the same scheme (11) with $\theta_t^\star$ replacing $\theta_0^\star$.

We measure both average and 90%-quantile losses for our problems. Based on our intuition, we expect that the lower k is (recall that $f(t) \propto t^k$), the better performance the fit model should exhibit on high quantiles of the loss, with potential decrease in average performance. Moreover, for $t = 0$, we should see that ERM and large k solutions exhibit the best average performance, with growing t reversing this behavior. Figure 2 bears this intuition out. In Figure 2(a), we plot the average loss (solid line) and the 90%-quantile of the losses (dotted line) on the shifted test sets, where the horizontal axis displays the rotation value $t \in [0, \pi]$. From the plot, we see precisely our predictions: the distributionally robust solution $\widehat{\theta}_n$ has worse *mean* loss on the original



(a) Average loss

(b) Loss on minority group

Figure 3: Figures (c) and (d) plots average and minority group losses under the distribution (13).

distribution than empirical risk minimization (ERM), but it achieves significantly smaller loss on the distributional shifts. The ordering of the mean performance of the different solutions inverts as the perturbation grows: for $t = 0$, that is, no perturbation, the least robust method (empirical risk minimization) has the best performance, while the most robust method (corresponding to $k = \frac{3}{2}$ or $k_* = 3$) eventually achieves the best performance for large perturbation.

3.1.2 Tail performance

We transition now to regression experiments, investigating performance on rare losses, where the goal is to predict $y \in \mathbb{R}$ from $x \in \mathbb{R}^d$ and we use loss $\ell(\theta; (x, y)) = \frac{1}{2}(y - x^\top\theta)^2$. In this case, we again take $d = 5$ and generate data $X \stackrel{\text{iid}}{\sim} N(0, I_d)$, $\varepsilon \sim N(0, .01)$,

$$Y = \begin{cases} X^\top\theta^\star + \varepsilon & \text{if } \; X^1 \le z_{.95} = 1.645 \\ X^\top\theta^\star + X^1 + \varepsilon & \text{otherwise,} \end{cases} \tag{12}$$

where in each experiment we choose $\theta^\star$ uniformly on the unit sphere $\mathbb{S}^{d-1}$ and X^1 denotes the first coordinate of X. (We use very small noise to highlight the more precise transition between average-case and higher percentiles.) As the effect of X^1 changes only 5% of the time (when it is above $z_{.95}$), we expect ERM to have poor performance on rare events when $X^1 \ge 1.645$—or in the tails generally. In addition, a fully robust solution in our framework is to use $\theta^{\text{rob}} = \theta^\star + \frac{1}{2}e_1$, as this minimizes worst-case expected loss across the two cases (12); we expect that for high robustness parameters (ρ large) the robust model should have worse average performance but about half of the losses at higher quantiles. We simulate $n_{\text{train}} = 2000$ training data points, and train the distributionally robust solution (2) with $\rho \in \{.001, .01, .1, .5, 4.5\}$, and $k \in \{1.5, 2, 4\}$. In Figure 2(b), we plot the mean loss under the data generation scheme (12) as solid lines and the 90%-quantile as a dotted line. We see once again, as our intuition suggests, that the robust solutions trade tail performance for average-case performance. The tail performance (90th percentile losses) improve with increasing robustness level ρ, with slight degradation in average case performance.

3.1.3 Performance on different subgroups

For our final small-scale simulation, we study item 3 (subpopulation performance) by considering a two-dimensional regression problem with two subgroups. In this case, we define the parameters $\theta_1^\star = (1, 1)$ and $\theta_2^\star = (0, -1)$, and we generate

$$Y = X^\top \theta_G^\star + \varepsilon \tag{13}$$

where $X \sim N(0, I_2)$, $\varepsilon \sim \mathsf{N}(0, .01)$, and $G = 1$ with probability 0.9 and $G = 2$ otherwise, where the variables are independent. Both the distributionally robust procedure (2) and ERM are oblivious to the label G, where $G = 1$ is the majority group, and $G = 2$ is the minority group. In addition to the ERM solution, we also consider the maximin effects estimator [57] as a benchmark,

$$\widehat{\theta}_n^{\text{maximin}} = \operatorname*{argmax}_\theta \min_{g=1,2} \left\{ 2\theta^\top \widehat{\Sigma}_{n,g} \theta_g^\star - \theta^\top \widehat{\Sigma}_{n,g} \theta \right\}$$

where $\widehat{\Sigma}_{n,g}$ is the empirical covariance matrix of the X_i such that $G_i = g$, and which maximizes the explained variance for each group [57]. The estimator $\widehat{\theta}_n^{\text{maximin}}$ is an oracle method requiring knowledge of the labels G_i and the group-specific regressors $\theta_g^\star$ for $g = 1, 2$.

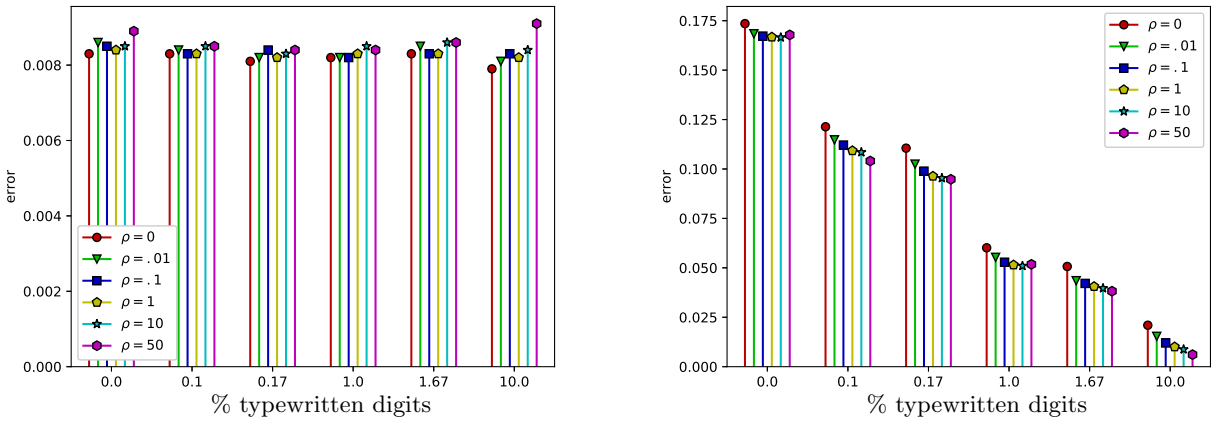
We simulate $n_{\text{train}} = 1000$ training data points. In Figure 3(a) and (b), we plot the average and average minority group losses for the different methods, respectively. Here the robust methods interpolate between the empirical risk minimizing (ERM) solution—which has the best average loss and worst minority group loss—and the maximin estimator $\widehat{\theta}_n^{\text{maximin}}$, which sacrifices performance on the average loss for strong minority group performance. The distributionally robust estimators $\widehat{\theta}_n$ exhibit tradeoffs between the two regimes, improving performance on the minority population at smaller degradation in the average loss. The parameters ρ and k allow flexibility in achieving these tradeoffs, though they of course must be set appropriately in applications.

3.2 Domain generalization for classification and digit recognition

Minority	All Digits		Digit 9 (hard)		Digit 6 (hard)		Digit 3 (easy)	
proportion	ERM	$\rho = 50$	ERM	$\rho = 50$	ERM	$\rho = 50$	ERM	$\rho = 50$
0	17.35	16.78	30.12	25.98	35.63	38.39	6.69	6.69
0.1	12.14	10.4	21.95	17.03	21.06	14.27	6.89	6.99
0.17	11.05	9.48	19	10.83	19.69	12.8	6.89	7.19
1	6.01	5.18	10.73	5.81	7.97	7.97	4.92	3.54
1.67	5.07	3.82	9.35	4.13	6.59	5.91	4.63	3.54
10	2.1	0.61	3.44	0.59	1.77	0.39	2.66	0.69

Table 1: Test error on type-written digits (%)

We now transition to experiments with real datasets. In this first of our real experiments, we consider a multi-class digit classification example, investigating domain generalization, though we conflate this with item 3 (multiple subpopulations). We construct our training set as a mixture of MNIST hand-written digits [30] (majority population) and type-written digits consisting of different fonts [28] (minority population). We fix the number of training examples, and vary the minority proportions of type-written digits from 0–10% of the training data. In the MNIST hand-written training dataset comprising of $n_{\text{train}} = 60,000$ digits, we replace $n \in \{0, 6, 10, 60, 100, 600\}$ images per digit by randomly drawn digits from the type-written dataset (with the same label).



(a) MNIST hand-written Digits

(b) All type-written Digits

(c) Type-written Digit 3 (easy class)

(d) Type-written Digit 9 (hard class)

Figure 4. (a) Test error on the *hand-written digits* (MNIST test dataset). (b)–(d) Test errors on *type-written digits*. Models were trained on data consisting of MNIST hand-written digits with 0–10% replaced by type-written digits. The horizontal axis of each plot denotes percentage of type-written digits (relative to handwritten) in training. Each of the six lines represents a different value of ρ used in training, where $\rho = 0$ corresponds to empirical risk minimization (ERM). (b) Classification error on entire test set of type-written digits. (c) Classification error on digit 3 of the type-written digits. (d) Classification errors for digit 9 of the type-written digits.

Our classifiers have no knowledge of whether a given image is hand-written or type-written, and our goal is to learn models that perform uniformly well across both majority (hand-written) and minority (type-written) subpopulations. We compare our procedure (2) with $k = 2$ against the ERM solution $\widehat{\theta}_n^{\text{erm}}$, where we vary ρ and the latent minority proportion. We evaluate our classifiers on both hand-written and type-written digits on held-out tests sets.

For $y \in \{0, \ldots, 9\}$ and $x \in \mathbb{R}^d$, we use the multi-class logistic regression loss

$$\ell(\theta; (x, y)) = \log \left(\sum_{i=0}^{k} \exp \left((\theta_i - \theta_y)^\top x \right) \right),$$

where $\theta_i \in \mathbb{R}^d$ we arbitrarily set $\theta_0 = \mathbf{0}$ as it is redundant. For our feature vector X, we use the $d = 4509$-dimensional output of the final fully connected layer of LeNet [54] after 10^4 stochastic gradient steps on the training dataset (see [46] for detailed hyper-parameter settings). We constrain our parameter matrix $[\theta_0, \ldots, \theta_9] \in \mathbb{R}^{d \times 10}$ to lie in the Frobenius norm ball of radius $r = 5$, chosen by cross validation on ERM ($\rho = 0$); this advantages ERM.

Returning to the justification for our development, we believe that the robust models should exhibit better performance on test data different from the training data than ERM models. This prediction is mostly consistent with the data, though the effects are not always strong. In Figure 4, we plot the classification errors over the minority proportion as we vary ρ (so that $\rho = 0$ corresponds to empirical risk minimization), summarizing the classification errors in Table 1. In Figure 4(a), we observe virtually the same performance on the hand-written test set (majority) across different radii ρ (error below 1%, with a decrease in accuracy of at most .1–.2%). On a test set of all typed digits (Figure 4(b)), the robust solutions exhibit a 1–2% improvement over the non-robust (ERM) solution in each mixture of typewritten digits (minority proportions) into the training data, which is larger than the persistent .1–.2% degradation on handwritten recognition. The trend of robust improvements on typewritten digits is more pronounced on the harder classes: the gap between $\widehat{\theta}_n^{\rm erm}$ and $\widehat{\theta}_n$ widens up to 9% on the digit 9 (see Table 1 and Fig. 4(d)). We observe that $\widehat{\theta}_n$ consistently performs well on the latent minority (type-written) subpopulation by virtue of upweighting the hard instances in the training set.

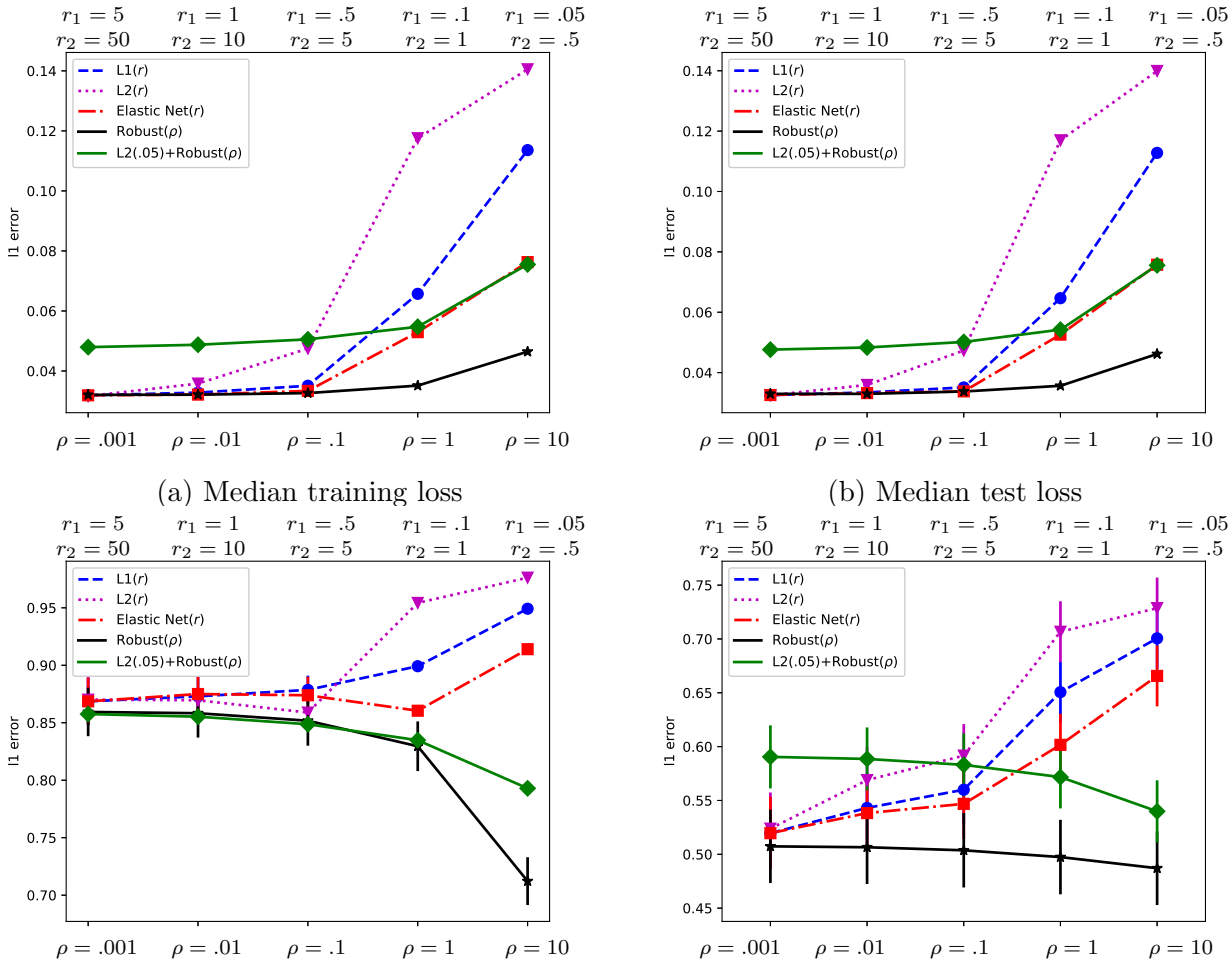
3.3 Tail performance in a regression problem

We consider a linear regression problem using the `communities and crime` dataset [62, 6], studying the peformance of distributionally robust methods on tail losses. Given a 122-dimensional attribute vector X describing a community, the goal is to predict per capita violent crimes Y (see [62]). We use the absolute loss $\ell(\theta; (x, y)) = |\theta^\top x - y|$ and compare method (2) with constrained forms of lasso, ridge, and elastic net regularization [88], taking constraint sets Θ of the form

$$\Theta = \left\{\theta \in \mathbb{R}^d : a_1 \left\|\theta\right\|_1 + a_2 \left\|\theta\right\|_2 \le r\right\}.$$

We vary a_1, a_2, and r: for ℓ_1-constraints we take $a_1 = 1, a_2 = 0$ and vary $r_1 \in \{.05, .1, .5, 1, 5\}$; for ℓ_2-constraints we take $a_1 = 0, a_2 = 1$ and vary $r_2 \in \{.5, 1, 5, 10, 50\}$; for elastic net we take $a_1 = 1, a_2 = 10$ and set $r = r_1 + r_2$. We compare these regularizers with the distributionally robust procedure (2) with $k = 2$, and the same procedure coupled with the ℓ_2-constraint ($a_1 = 1, a_2 = 0$) with $r = .05$, where we vary $\rho \in \{.001, .01, .1, 1, 10\}$.

In Figure 5, we plot the quantiles of the training and test losses with respect to different values of regularization or ρ. The horizontal axis in each figure indexes our choice of regularization value. We observe that $\widehat{\theta}_n$ shows very different behavior than other regularizers; $\widehat{\theta}_n$ attains median losses similar or slightly higher than the regularized ERM solutions, and achieves much smaller loss on the tails of the inputs. As ρ grows, the robust solution exhibits increasing median loss—though slowly—and decreasing maximal loss. To validate our experiments, we made 50 independent random partitions of our dataset with $n = 2118$ samples. For each random partition, we divide the dataset into training set with $n_{\rm train} = 1800$ and a test set with $n_{\rm test} = 318$.



(a) Median training loss

(b) Median test loss

(c) Maximal training loss

(d) Maximal test loss

Figure 5. Median and maximal loss $|Y - Z^\top \theta|$ evaluated on training and test datasets. Values of the x-axis corresponds to different indices for the values of ρ and r, so that "x-axis = 1" for the ℓ_1-constrained problem corresponds to $r = 5$, and for the distributionally robust method (2) it corresponds to $\rho = .001$. Error bars correspond to standard error.

3.4 Fine-grained recognition and challenging sub-groups

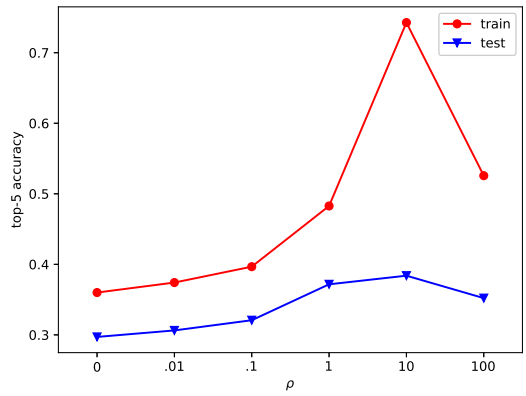
Finally, we consider the fine-grained recognition task of the `Stanford Dogs` dataset [48], where the goal is to classify an image of a dog into one of 120 different breeds. There are 20,580 images, $n_{\text{train}} = 12{,}000$ training examples, with 100 training examples for each class. We use the default histogram of SIFT features in the dataset [81], resulting in vectors $x \in \mathbb{R}^d$ with $d = 12{,}000$.

We train 120 one-versus-rest classifiers, one for each of the classes, and combine their predictions by taking the k predictions with largest scores for a given example x. For each binary classification problem, we use the binary logistic loss, regularized with lasso (in constrained form) so that

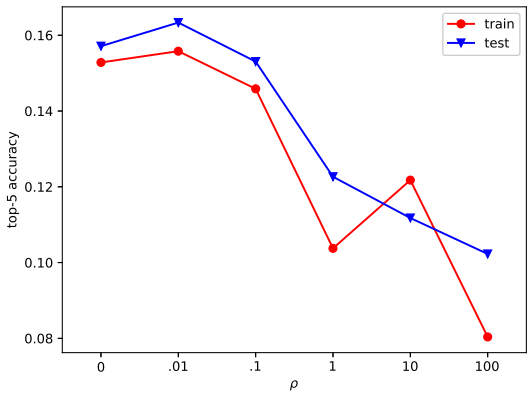
$$\Theta_{\text{one-vs-rest}} = \left\{\theta \in \mathbb{R}^d : \|\theta\|_1 \le r\right\}.$$

Thus, for each class i, we represent a pair (x, y) by $y = 1$ if the image is of breed i, and -1 otherwise, fitting a binary classifier θ_i for each class. We use $r = 1.0$ for all of our methods based on cross-validation for ERM ($\rho = 0$). As we predict using the k highest scores, we measure performance with respect to top-k accuracy, which counts the number of test examples in which the true label was among these k predictions. Based on our intuitions on robustness and subpopulation performance, we expect that for large ρ in the robustness set $\{P : D_f(P\|P_0) \leq \rho\}$, we should have better performance on challenging classes, sacrificing performance on easier classes. We also expect that for large ρ, the variance in the top-5 accuracy across classes should be smaller—we expect more uniform performance. We do not necessarily expect that accuracies should improve as ρ increases.

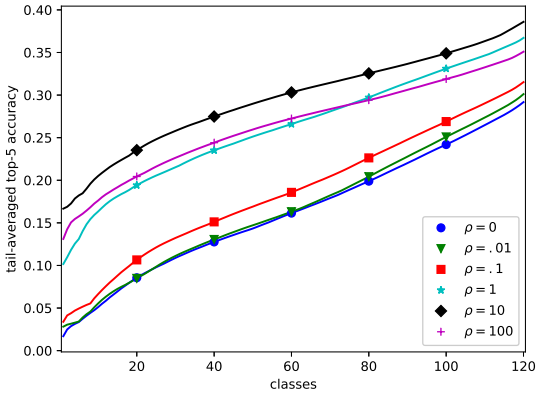
In Figure 6, we present our results. We use top-5 accuracy; top-1 and top-3 accuracies are similar. Overall accuracy improves moderately as ρ grows (Figure 6(a)). The *standard deviation* of the top-5 accuracy across the classes decreases as ρ increases (Figure 6(b)), which is consistent with our hypothesis that the robust formulations should yield more uniform performance. In Figure 6(c), we plot the accuracy averaged over the first c-classes that suffer the lowest accuracy under each model, varying c on the horizontal axis; the accuracy at $c = 120$ is simply the average top-5 accuracy of the models. For c small, meaning for classes on which the respective models perform most poorly, we observe that the ensemble of one-vs-rest $\widehat{\theta}_n$'s outperform the ensemble of ERM solutions $\widehat{\theta}_n^{\text{erm}}$'s. In Figure 6(d), we plot the accuracy averaged over the first c-classes that have the lowest accuracy under the ERM model. In this case, $\widehat{\theta}_n$ improves performance on classes that ERM does poorly on; such tail-performance improves monotonically with ρ up to $\rho = 10$; we conjecture the degradation for higher ρ is a consequence of overly conservative estimates. Because of the conflation of increasing ρ with improved overall performance and robustness that the figures illustrate, it is somewhat hard to draw conclusions from this experiment, and it is in a sense inconclusive. With that said, Figure 6(c) shows that the gap between the robust classifier performance and non-robust classifier goes from .17 vs. .03 (hardest class accuracy) to .38 vs. .28 (overall accuracy), so that relative performance gains of the robust approach seem largest on the hardest classes.



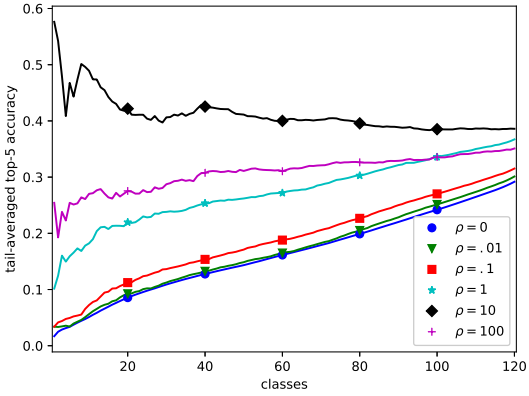
(a) Overall top-5 accuracy



(b) Standard deviation of top-5 accuracy



(c) Test top-5 on worst c classes



(d) Test top-5 accuracy on worst ERM classes

Figure 6. (a) Top-5 error against ρ on train and test. (b) Standard deviation of top-5 accuracy across 120 different classes against ρ. (c) Test top-5 accuracy on the worst-c classes under each model, i.e. c classes with lowest accuracy under each model. (d) Test top-5 accuracy on the worst-c classes ordered by accuracy of *ERM* model ($\rho = 0$).

4 Convergence Guarantees

Our empirical expermience in the previous section gives convincing evidence for the potential statistical benefits of the distributionally robust estimator (2). As a consequence, in our view it is important to develop some of its theoretical properties, so we investigate its performance under a variety of conditions on the f-divergence. In this section, we provide finite sample convergence guarantees. Recalling the definition (7) of worst-case risk $\mathcal{R}_k(\theta; P_0)$ corresponding to the Cressie-Read family of divergences (6), we show that the empirical minimizer $\widehat{\theta}_n$ for the plug-in (2) satisfies $\mathcal{R}_f(\widehat{\theta}_n; P_0) - \inf_{\theta\in\Theta} \mathcal{R}_f(\theta; P_0) \leq C n^{-\frac{1}{k_* \vee 2}}$ with high probability, where $k_* = \frac{k}{k-1}$ and C is a problem dependent constant. As we show in Section 5, the $n^{-1/(k_* \vee 2)}$ rate is optimal in n. The departure from parametric rates as the uncertainty set becomes large, meaning $k \downarrow 1$ or $k_* = \frac{k}{k-1} \uparrow \infty$, is a consequence of the fact that in the worst case, it is challenging to estimate L^q-norms of random

variables X for $q > 2$; that is, the minimax rate for such estimation is $n^{-1/q}$.

Throughout this section, we assume that for any $\theta \in \Theta$ and $x \in \mathcal{X}$, we have $\ell(\theta; x) \in [0, M]$, and restrict attention to the Cressie-Read family of divergences (6) with $k \in (1, \infty)$. We first show pointwise concentration of the finite sample objective $\mathcal{R}_k(\theta; \widehat{P}_n)$ to its population counterpart $\mathcal{R}_k(\theta; P_0)$ by using convex concentration inequalities [19, 75] and then carefully bounding the expected risk $\mathbb{E}[\mathcal{R}_k(\theta; \widehat{P}_n)]$, which is a biased estimator of the population risk $\mathcal{R}_k(\theta; P_0)$.

Theorem 2. *Assume that $\ell(\theta; x) \in [0, M]$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$, and define $c_k := (k(k-1)\rho + 1)^{1/k}$. For a fixed $\theta \in \Theta$ and $t > 0$, with probability at least $1 - 2e^{-t}$*

$$\left| \mathcal{R}_k(\theta; \widehat{P}_n) - \mathcal{R}_k(\theta; P_0) \right| \leq c_k M n^{-\frac{1}{k_* \vee 2}} \left(\frac{2}{k} + \sqrt{t} \right).$$

We provide a proof in Section B.1. We may replace the assumption $\ell(\theta; x) \in [0, M]$ with $\sup_x \ell(\theta; x) - \inf_x \ell(\theta; x) \leq M$ without any change to the conclusion or argument of the theorem.

Given the pointwise concentration result (Theorem 2), we can use a simple covering argument to obtain its uniform counterpart. Our uniform guarantees rely on covering numbers for the model class $\{\ell(\theta; \cdot) : \theta \in \Theta\}$ as the notion of complexity (e.g. [83]). Recall that for a set V, a collection $v_1, \ldots, v_N$ is an *ϵ-cover* of V in norm $\|\cdot\|$ if for each $v \in \mathcal{V}$, there exists v_i such that $\|v - v_i\| \leq \epsilon$. The *covering number* of V with respect to $\|\cdot\|$ is

$$N(V, \epsilon, \|\cdot\|) := \inf \{ N \in \mathbb{N} \mid \text{there is an } \epsilon\text{-cover of } V \text{ with respect to } \|\cdot\| \}.$$

For $\mathcal{F} := \{\ell(\theta, \cdot) : \theta \in \Theta\}$ equipped with sup-norm $\|h\|_{L^\infty(\mathcal{X})} := \sup_{x \in \mathcal{X}} |h(x)|$, we have the following uniform concentration result.

Corollary 1. *Let $\ell(\theta; x) \in [0, M]$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$. Then for any $t > 0$, with probability at least $1 - 2N(\mathcal{F}, \frac{t}{3}, \|\cdot\|_{L^\infty(\mathcal{X})})e^{-t}$*

$$\sup_{\theta \in \Theta} \left| \mathcal{R}_k(\theta; \widehat{P}_n) - \mathcal{R}_k(\theta; P_0) \right| \leq c_k M n^{-\frac{1}{k_* \vee 2}} \left(\frac{2}{k} + \sqrt{6t} \right).$$

Corollary 1 and an application of the triangle inequality immediately yield the next corollary.

Corollary 2. *Let $\ell(\theta; x) \in [0, M]$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$. Then for any $t > 0$, with probability at least $1 - 2N(\mathcal{F}, \frac{t}{3}, \|\cdot\|_{L^\infty(\mathcal{X})})e^{-t}$*

$$\mathcal{R}_k(\widehat{\theta}_n; P_0) \leq \inf_{\theta \in \Theta} \mathcal{R}_k(\theta; P_0) + 2c_k M n^{-\frac{1}{k_* \vee 2}} \left(\frac{2}{k} + \sqrt{6t} \right).$$

As an example, let $\theta \mapsto \ell(\theta; x)$ be L-Lipschitz for all $x \in \mathcal{X}$, with respect to some norm $\|\cdot\|$ on Θ. Assuming $D := \sup_{\theta, \theta' \in \Theta} \|\theta - \theta'\| < \infty$, a standard bound [83, Chapter 2.7.4] is

$$N\left(\mathcal{F}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{X})}\right) \leq N\left(\Theta, \frac{\epsilon}{L}, \|\cdot\|\right) \leq \left(1 + \frac{DL}{\epsilon}\right)^d.$$

If there exists $\theta_0 \in \Theta$ and $M_0 > 0$ such that $|\ell(\theta_0; x)| \leq M_0$ for all $x \in \mathcal{X}$, we have $|\ell(\theta; X)| \leq LD + M_0$, and Corollary 2 implies that

$$\mathcal{R}_k(\widehat{\theta}_n; P_0) \leq \inf_{\theta \in \Theta} \mathcal{R}_k(\theta; P_0) + 2c_k n^{-\frac{1}{k_* \vee 2}} (DL + M_0) \left(\frac{2}{k} + \sqrt{6t} \right)$$

with probability at least $1 - 2\exp(-t + d\log(1 + \frac{3DL}{t}))$. Replacing covering numbers in the above guarantees with tighter quantities such as Rademacher averages or their localized variants [7] is a topic of future research.

5 Lower Bounds

To complement our uniform upper bounds, in this section, we provide minimax lower bounds showing they are rate optimal, though developing optimal dimension-dependent bounds remains open. For a collection $\mathcal{P}$ of distributions and f-divergence f, we define the minimax rate

$$\mathfrak{M}_n(\mathcal{P}, f, \ell) := \inf_{\widehat{\theta}(X_1^n)} \sup_{P_0 \in \mathcal{P}} \mathbb{E}_{P_0}\left[\mathcal{R}_f\left(\widehat{\theta}(X_1^n); P_0\right) - \inf_{\theta \in \Theta} \mathcal{R}_f\left(\theta; P_0\right)\right] \tag{14}$$

where the outer infimum is over all $(X_1, \ldots, X_n)$-measurable functions and the inner supremum is over probability measures in $\mathcal{P}$, where the loss is implicit in the risk $\mathcal{R}_f$. Whenever $f(t) \lesssim t^k$ as $t \uparrow \infty$, we show there exist losses for which $n^{-1/(k_* \vee 2)}$ is a lower bound on the minimax distributionally robust risk (14). Thus there is a necessary transition from parametric $\sqrt{n}$-type rates to n^{1/k_*} when k is small—that is, we seek more distributionally robust protection.

We divide our lower bounds into bounds on our ability to even estimate the risk $\mathcal{R}_f(\theta; P_0)$ and lower bounds on the actual minimax risk (14), which build out of these results (Sections 5.1 and 5.2, respectively). Within each section, we initially present our results for the Cressie-Read family (6) with $k \in (1, \infty)$, allowing explicit constants, then providing lower bounds for general f-divergences using the same techniques. The rough intuition for our approach is as follows: we consider Bernoulli random variables Z supported on $\{0, M\}$, where the probability that $Z = M$ is small, though this probability has substantial influence on the risk $\mathcal{R}_f$. This highlights the reason for the potentially slow rates of convergence: one must sometimes observe rarer events to estimate or optimize the risk $\mathcal{R}_f$.

5.1 Lower bounds on estimation

We first establish a lower bound for estimating the worst-case objective $\mathcal{R}_k(\theta; P_0)$ under the Cressie-Read family of divergences (6). For the rest of this subsection, we fix an arbitrary $\theta \in \Theta$, and consider $Z(x) := \ell(\theta; x)$, abusing notation by writing

$$\mathcal{R}_k(Z) := \sup_{Q \ll P_0} \{\mathbb{E}_Q[Z] : D_f\left(Q \| P_0\right) \le \rho\}.$$

The proof of the following result—which we give in Section C.1—uses Le Cam's method [87, 53]. Our proof uses that if Z takes two values $z_1 < z_2$, then $\mathcal{R}_k(Z) = z_2$ holds if and only if P_0 places enough mass on z_2; we compute the precise threshold at which the worst-case region contains a point mass, quantifying the fundamental difficulty in estimating $\mathcal{R}_k(Z)$.

Theorem 3. *Let $\rho > 0$ be arbitrary but fixed and $\mathcal{P}$ be the collection of Bernoulli random variables taking values on $\{0, M\}$. Define $c_k := (1 + k(k-1)\rho)^{1/k}$, $p_k := (1 + k(k-1)\rho)^{-1/(k-1)}$, and $\beta_k = \frac{k(k-1)\rho}{2(1+k(k-1)\rho)}$. Then*

$$\inf_{\widehat{R}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P_0}\left|\widehat{R}(Z_1^n) - \mathcal{R}_k(Z)\right| \ge M \max\left\{\frac{1}{8k_* p_k}\left(\sqrt{\frac{p_k(1-p_k)}{8n}} \wedge \frac{1}{2}(1-p_k) \wedge p_k\right), \right.$$
$$\left. \frac{1}{8}\beta_k^{\frac{1}{k}} c_k \left(\frac{1}{4n} \wedge p_k \wedge (1 - (1-\beta_k)^{1-k_*} p_k)\right)^{\frac{1}{k_*}}\right\}$$

where $Z \sim P$ and $Z_1^n \stackrel{\text{iid}}{\sim} P$, and the outer infimum is over $\widehat{R} : \{0, M\}^n \to \mathbb{R}$.

For general f-divergences we can provide a similar result, showing that the growth of the function f defining the divergence D_f fundamentally determines worst-case rates of convergence; when $f(t)$ grows slowly as $t \uparrow \infty$, the robust formulation (1) is conservative, so rates of convergence are slower. For these results, we assume for simplicity that f is strictly convex at $t = 1$, meaning that $f(\lambda t_0 + (1-\lambda)t_1) < \lambda f(t_0) + (1-\lambda)f(t_1)$ whenever $t_0 < 1 < t_1$. To state our results, we require some additional notation. For $p, q \in [0, 1]$, define the binary divergence

$$h_f(q; p) := pf\left(\frac{q}{p}\right) + (1-p)f\left(\frac{1-q}{1-p}\right).$$

As f is strictly convex at $t = 1$, for $q \geq p$ the function $q \mapsto h_f(q; p)$ is strictly increasing on its domain and continuous, so there exists a unique

$$q(p) := \sup_{q \geq p}\{q : h_f(q; p) \leq \rho\}. \tag{15}$$

We then have

Proposition 4. *Let $f : (0, \infty) \to \mathbb{R} \cup \{+\infty\}$ be strictly convex at $t = 1$ and $\mathcal{P}$ be the collection of distributions on Z supported on $\{0, M\}$. Assume there exists $p \in (0, 1)$ such that f is $\mathcal{C}^1$ in a neighborhood of $\frac{q(p)}{p}$ and $\frac{1-q(p)}{p}$. Then for any such p,*

$$\liminf_{n\to\infty} \inf_{\widehat{R}} \sup_{P\in\mathcal{P}} \sqrt{n}\mathbb{E}_{P_0}\left[\left|\widehat{R}(Z_1^n) - \mathcal{R}_f(Z)\right|\right] \geq M\frac{\sqrt{p(1-p)}}{8}\frac{-\partial_p h_f(q(p); p)}{\partial_q h_f(q(p); p)} > 0$$

where the infimum is over $\widehat{R} : \{0, M\}^n \to \mathbb{R}$.

See Section C.2 for the proof. The final ratio is indeed positive, as by (strict) convexity of f and the joint convexity of h_f, $\partial_q h_f(q(p); p) > 0 \in \partial_q h_f(p; p)$ and $\partial_p h_f(q(p); p) < 0 \in \partial_p h_f(q(p); q(p))$.

Under a slightly different condition that the asymptotic growth of f is at most t^k, we can give an $\Omega(n^{-1/k_*})$ lower bound for $k_* = \frac{k}{k-1}$. Letting $f^{-1}(s) := \inf\{t \in [0, 1] : f(t) \leq s\}$, for all $m > 0$ we define

$$C_{f,\rho,m} := \frac{m}{\rho}\left(1 \wedge \left(\frac{\rho}{2m}\right)^{-k_*}\left(1 - f^{-1}\left(\frac{\rho}{2}\right)\right)^{k_*}\right)^{-1}. \tag{16}$$

Then we have the following result, whose proof we provide in Section C.3.

Proposition 5. *Let $\mathcal{P}$ be the collection of distributions on Z supported on $\{0, M\}$. For some $m > 0$ and $k \in (1, \infty)$, assume $f(t) \leq mt^k$ for all $t \geq \left\{(n \vee C_{f,\rho,m})\rho m^{-1}\right\}^{\frac{1}{k}}$. Then*

$$\inf_{\widehat{R}} \sup_{P\in\mathcal{P}} \mathbb{E}_P\left|\widehat{R}(Z_1^n) - \mathcal{R}_f(Z)\right| \geq \frac{M}{16}\left(\frac{\rho}{m}\right)^{\frac{1}{k}}\left(\frac{1}{n \vee C_{f,\rho,m}}\right)^{\frac{1}{k_*}}$$

where the infimum is over $\widehat{R} : \{0, M\}^n \to \mathbb{R}$.

5.2 Lower bounds on optimization

The lower bounds we provide on optimization are a bit different, though the techniques grow out of those in the previous section on estimating the risk $\mathcal{R}_f$. We consider linear losses, which makes the situation closest to the estimation of the risk results above (as we roughly must still estimate kth norms of random variables), providing analogous lower bounds for optimizing the worst-case

objective $\mathcal{R}_f(\cdot; P_0)$. To study the minimax risk for optimization, we use a standard distance-like quantity for proving lower bounds in stochastic optimization [1, 31], then construct a reduction from distributionally robust optimization to hypothesis testing.

We begin by considering the lower bound for the Cressie-Read family (6). See Section C.4 for the proof of the lower bound.

Theorem 6. *Let $\ell(\theta; x) = \theta x$ where $\theta \in \Theta = [-M, M]$ and $x \in [-1, 1]$, and f_k be the kth order Cressie-Read function (6). Define $c_k := (1 + k(k-1)\rho)^{1/k}$, $p_k := (1 + k(k-1)\rho)^{-1/(k-1)}$, and $\beta_k = \frac{k(k-1)\rho}{2(1+k(k-1)\rho)}$. Then*

$$\mathfrak{M}_n(\mathcal{P}, f_k, \ell) \geq M \max \left\{ \frac{1}{16 k_* p_k} \left(\sqrt{\frac{p_k(1-p_k)}{n}} \wedge \frac{1}{2}(1-p_k) \wedge (1-2p_k) \wedge p_k \right), \right. \\ \left. \frac{1}{16} \beta_k^{\frac{1}{k}} c_k \left(\frac{1}{4n} \wedge p_k \wedge (1-p_k) \wedge (1-(1-\beta_k)^{1-k_*} p_k) \right)^{\frac{1}{k_*}} \right\}.$$

For general f-divergences, we can show a similar standard $\Omega(n^{-1/2})$ lower bound for optimization. We defer the proof of this result to Section C.5.

Proposition 7. *Let the conditions on f of Proposition 4 hold, and let $\ell(\theta; x) = \theta x$ and $\mathcal{P}$ be the collection of distributions supported on $[-M, M]$. Then*

$$\liminf_{n \to \infty} \sqrt{n} \mathfrak{M}_n(\mathcal{P}, f, \ell) \geq M \frac{\sqrt{p(1-p)}}{16 q(p)} \frac{-\partial_p h_f(q(p); p)}{\partial_q h_f(q(p); p)} > 0.$$

Our final minimax lower bound applies to f-divergences with $f(t) = O(t^k)$ as $t \to \infty$, where we can prove a lower bound on optimization of $\mathcal{R}_f(\cdot; P_0)$ with rate n^{-1/k_*}. Recalling the definition (16) of $C_{f,\rho,m}$, we obtain the following result. See Section C.6 for the proof.

Proposition 8. *Let $\ell(\theta; x) = \theta x$ and $\mathcal{P}$ be the collection of distributions supported on $[-M, M]$. If there exists $m > 0$ such that $f(t) \leq m t^k$ for all $t \geq \{(n \vee C_{f,\rho,m}) \rho m^{-1}\}^{\frac{1}{k}}$, then for $\ell(\theta; x) = \theta x$ with $\theta \in \Theta = [-M, M]$ and $x \in [-1, 1]$*

$$\mathfrak{M}_n(\mathcal{P}, f, \ell) \geq \frac{M}{16} \left(\frac{\rho}{m} \right)^{\frac{1}{k}} \left\{ \left(\frac{1}{n \vee C_{f,\rho,m}} \right)^{\frac{1}{k_*}} \wedge \left(\frac{\rho}{2m} \right)^{\frac{1}{k_*}} \left(\left(\frac{2}{3} \right)^{k-1} \wedge \left(\frac{1}{2} \right)^{\frac{1}{k_*}} \frac{2m}{\rho} \right) \right\}.$$

This result shows that—at least in terms of rates in n—there is a tradeoff between degree of robustness, as measured by the asymptotic growth of the function f defining the robustness set $\{P : D_f(P \| P_0) \leq \rho\}$, and worst-case convergence rates. In this sense, we see that our finite sample convergence guarantees of Section 4 are sharp.

6 Asymptotics

In the previous two sections, we studied convergence properties for the robust formulation (1) that hold uniformly over collections of data generating distributions P_0, showing that robustness can incur nontrivial statistical cost. In this section, by contrast, we turn to pointwise asymptotic properties of the empirical plug-in (2), applying to a fixed distribution P_0. This allows two contributions.

First, we prove a general consistency result for convex losses. Second, while the minimax convergence rates in the previous section exhibit a departure from classical parametric rates, we show that under appropriate regularity conditions the typical $\sqrt{n}$-rates of convergence and asymptotic normality guarantees are possible.

6.1 Consistency

In this section, we give a general set of convergence results, relying on the powerful theory of epi-convergence in variational analysis [64, 49]. Our first results shows that $\mathcal{R}_f(\theta; \widehat{P}_n)$ is a (pointwise) consistent estimator of its population counterpart $\mathcal{R}_f(\theta; P_0)$ under mild conditions on f. See Section D.1 for the proof.

Proposition 9. *Let f be finite on (t_0, ∞) for some $t_0 < 1$. For any $\theta \in \Theta$, if $\mathbb{E}[f^*(|\ell(\theta; X)|)] < \infty$ then $\mathcal{R}_f(\theta; \widehat{P}_n) \stackrel{a.s.}{\to} \mathcal{R}_f(\theta; P_0) < \infty$.*

We now provide sufficient conditions for parameter consistency in the distributionally robust estimation problem (2). The main assumption is that the loss functions are closed and the non-robust population risk is coercive. (Weaker sufficient conditions are possible, but in our view, a bit esoteric.)

Assumption A (Coercivity)**.** *For each $x \in \mathcal{X}$, the function $\theta \mapsto \ell(\theta; x)$ is closed convex, and $\mathbb{E}_{P_0}[\ell(\theta; X)] + \mathbf{I}(\theta \in \Theta)$ is coercive.*

It is possible to replace the convexity assumption with a Glivenko-Cantelli property on the collection $\{f^*(\ell(\theta; \cdot))\}_{\theta \in \Theta}$; for example, if $\theta \mapsto \ell(\theta; X)$ is continuous and Θ is compact, then a similar consistency result holds, although computation of the plug-in (2) may be difficult. The coercivity assumption guarantees the existence and compactness of the set of optima for $\mathcal{R}_f(\theta; P_0)$.

Before providing our consistency result, we define a small amount of additional notation. The *inclusion distance*, or the *deviation*, from a set A to B is

$$d_\subset(A, B) := \sup_{y \in A} \operatorname{dist}(y, B) = \inf_\epsilon \left\{\epsilon \geq 0 : A \subset \{y : \operatorname{dist}(y, B) \leq \epsilon\}\right\}.$$

Now, for any $\varepsilon \geq 0$ and distribution P on $\mathcal{X}$, define the sets of ε-approximate minimizers

$$S_P(\Theta, \varepsilon) := \left\{\theta \in \Theta \mid \mathcal{R}_f(\theta; P) \leq \inf_{\theta \in \Theta} \mathcal{R}_f(\theta; P) + \varepsilon\right\},$$

where we let $S_P(\Theta) = S_P(\Theta, 0)$ for shorthand. Then we have the following consistency result, whose proof we provide in Section D.2.

Proposition 10. *Let f be finite on (t_0, ∞) for some $t_0 < 1$, and assume that $\mathbb{E}[f^*(|\ell(\theta; X)|)] < \infty$ on some neighrborhood of $S_{P_0}(\Theta)$. Let Assumption A hold. Then*

$$\inf_{\theta \in \Theta} \mathcal{R}_f(\theta; \widehat{P}_n) \stackrel{a.s.}{\to} \inf_{\theta \in \Theta} \mathcal{R}_f(\theta; P_0),$$

and for any sequence $\varepsilon_n \downarrow 0$, with probability 1 we have $S_{\widehat{P}_n}(\Theta, \varepsilon_n) \neq \varnothing$ eventually and

$$d_\subset\left(S_{\widehat{P}_n}(\Theta, \varepsilon_n), S_{P_0}(\Theta)\right) \to 0.$$

6.2 Asymptotic normality

In this section, we provide a central limit result for the empirical optimizer $\widehat{\theta}_n \in \operatorname{argmin}_\theta \mathcal{R}(\theta; \widehat{P}_n)$ under appropriate smoothness conditions on the risk. Given that in the general formulation of our problem, the distribution P_0 and the supremum over P near P_0 act as nuisance parameters, it seems challenging to give the most generic conditions under which asymptotic normality of $\widehat{\theta}_n$ should hold. Accordingly, we assume somewhat simpler conditions that allow an essentially classical treatment, and we are thus somewhat brief.

We begin with a smoothness assumption.

Assumption B (Smoothness and growth)**.** *For some $k > 1$, the function f satisfies $\liminf_{t\to\infty} f(t)/t^k > 0$. Let $k_* = \frac{k}{k-1}$. There exists a neighborhood U of $\theta^\star$ such that*

1. *There exists $L : \mathcal{X} \to \mathbb{R}_+$ such that $|\ell(\theta_0; x) - \ell(\theta_1; x)| \le L(x) \|\theta_0 - \theta_1\|_2$ for all $\theta_i \in U$, where $\mathbb{E}[L(X)^{2k_*}] < \infty$.*
2. *$\mathbb{E}[|\ell(\theta^\star; X)|^{2k_*}] < \infty$.*
3. *The function $\theta \mapsto \ell(\theta; x)$ is differentiable on U.*

Recalling the dual (5), for shorthand define

$$g_P(\theta, \lambda, \eta) := \lambda \mathbb{E}_P \left[f^* \left(\frac{\ell(\theta; X) - \eta}{\lambda} \right) \right] + \rho\lambda + \eta.$$

We make the following additional

Assumption C (Strong identifiability)**.** *The objective g_{P_0} is $\mathcal{C}^2$ in a neighborhood of $(\theta^\star, \lambda^\star, \eta^\star) = \operatorname{argmin} g_{P_0}$ with positive definite Hessian. Additionally, $P_0(\ell(\theta^\star; \eta^\star) - \eta^\star > 0) > 0$.*

The second condition of Assumption C guarantees that $\lambda^\star > 0$.

In general, verifying Assumption C may be somewhat challenging. Let us provide a few conditions sufficient for uniqueness of $\lambda^\star$ and $\eta^\star$, along with differentiability, for the Cressie-Read divergences.

Lemma 2. *Assume that f is the Cressie-Read divergence (6) with parameter $k \in (1, \infty)$, and let $\theta_0 \in \Theta$. If $\ell(\theta_0; X)$ is non-constant under the distribution P and $\mathbb{E}_P[|\ell(\theta; X)|^{k_*}] < \infty$ near θ_0, then $(\lambda_0, \eta_0) = \operatorname{argmin}_{\lambda \ge 0, \eta} g_P(\theta_0, \lambda, \eta)$ is unique.*

See Appendix E.1 for a proof. Sufficient conditions for differentiability are similar to the classical conditions for asymptotic normality of quantile estimators [82]; for example, if $\ell(\cdot; X)$ is $\mathcal{C}^2$ near some θ_0 and $P(\ell(\theta; X) = \eta) = 0$ for θ, η near θ_0, η_0, then the dual formulation g_P is $\mathcal{C}^2$ in a neighborhood of $(\theta_0, \eta_0, \lambda_0)$ whenever $\lambda_0 > 0$.

With this brief preliminary discussion in place, we turn to providing an asymptotic normality result on $\widehat{\theta}_n$.

Theorem 11. *Let Assumptions B and C hold. Then any sequence of estimators $\widehat{\theta}_n$ satisfying $\mathcal{R}(\widehat{\theta}_n; \widehat{P}_n) \le \inf_\theta \mathcal{R}(\theta; \widehat{P}_n) + O_P(1/n)$ satisfies*

$$\sqrt{n} \left(\widehat{\theta}_n - \theta^\star \right) \overset{d}{\rightsquigarrow} \mathsf{N} \left(0, V \operatorname{Cov} \left(f^{*\prime} \left(\frac{\ell(\theta^\star; X) - \eta^\star}{\lambda^\star} \right) \nabla \ell(\theta^\star; X) \right) V \right) \tag{17}$$

where V is the first d-by-d block of $\left(\nabla^2 g(\theta^\star, \lambda^\star, \eta^\star) \right)^{-1} \in \mathbb{R}^{(d+2)\times(d+2)}$.

See Section E.2 for the proof.

As the positive definiteness of $\nabla^2 g_k(\theta^\star, \lambda^\star, \eta^\star)$ in Assumption C is hard to verify in most modeling scenarios, we can relax the assumption to positive definiteness of the Hessian of the map $(\eta, \theta) \mapsto c_k(\mathbb{E}_{P_0}[(\ell(\theta; X) - \eta)_+^{k_*}])^{\frac{1}{k_*}} + \eta$ at $(\theta^\star, \eta^\star)$, which is the dual objective g_k with λ minimized out. We omit the proof with this relaxed condition for brevity, as it is quite involved. Letting $B = (\ell(\theta^\star; X) - \eta^\star)_+$, under Assumption B and the randomness conditions of Lemma 2, this relaxed condition holds if

$$(k-1)\mathbb{E}B^{k_*-2}\left(\mathbb{E}B^{k_*}\mathbb{E}B^{k_*-2} - (\mathbb{E}B^{k_*-1})^2\right)\mathbb{E}[B^{k_*-1}\nabla^2\ell(\theta^\star; X)] \\ - \left(\mathbb{E}B^{k_*-1}\right)^2\mathbb{E}[B^{k_*-2}j\nabla\ell(\theta^\star; X)]\mathbb{E}[B^{k_*-2}\nabla\ell(\theta^\star; X)]^\top \succ 0, \tag{18}$$

and $k \in (1, 2)$. For $k = 2$, the relaxed condition holds if in addition to the bound (18) holding, there is a neighborhood of $(\theta^\star, \eta^\star)$ such that $\mathbb{P}(\ell(\theta; X) = \eta) = 0$.

7 Discussion and further work

We have presented a collection of statistical problems that arise out of a distributionally robust formulation of M-estimation, whose purpose is roughly to obtain uniformly small loss and protect against rare but large losses. While our results give convergence guarantees, and our experimental results suggest the potential of these approaches in a number of prediction problems, numerous questions remain.

In our view, the most important limitation is guidance in the choices of the robustness set, that is, $\{P : D_f(P\|P_0) \le \rho\}$. The analytic consequences of our choices are nice in that they allow explicit dual calculations and algorithmic development; in the case in which the radius ρ is instead shrinking with as ρ/n, asymptotic and non-asymptotic considerations [59, 32, 11, 51, 52] show that the robustness provides a type of regularization by variance of the loss when f is smooth, no matter that choice of f. In our setting, such limiting similarity is not the case, and it may be unrealistic to assume a user of the approach can justify the appropriate choice of f.

The minimax guarantees show that there are tradeoffs in terms of the robustness we provide, in the sense that larger robustness sets yield more difficult estimation and optimization problems. Our upper and lower bounds, however, only match up to rates in n of n^{-1/k_*}, so that our understanding of higher-dimensional robustness is limited. Obtaining convergence guarantees (Section 4) with scale-sensitive model complexity terms such as Rademacher complexity and its localized variants [7] is a topic of future research.

The robust formulation (1) and its empirical formulation (2) are complementary to traditional robustness approaches in statistics arising out of Huber's work [44, 45]. In the classical notions of robustness, one wishes to obtain an estimate of some parameter θ of a distribution P_0 contaminated by some noise Q; in our case, in contrast, we wish to obtain a parameter that performs well *for all* contaminations Q, at least contaminations nearby in some f-divergence ball. Developing a deeper understanding of the connections and contrasts between classical contamination models and distributional robustness approaches will likely yield fruit.

Two related issues arise when we consider problems with covariates X and a target or label Y. For these problems, the distributionally robust formulation (1) considers shifts in the joint distribution $(X, Y) \sim P_0$. Traditional domain adapation approaches, in contrast, take a fixed conditional distribution $P_{0,Y|X}(y \mid x)$ and consider shifts to the marginal distribution $P_{0,X}$ (covariate shift).

Similarly, in causal or intervential data analyses, one wishes to perturb only the distribution of the covariates X, observing the effect of such interventions on Y. Consequently, connecting these ideas and developing variants of the formulation (1) that only hedge against covariate shift or structural shifts on X may be useful in many scenarios.

Acknowledgments

JCD and HN were partially supported by the SAIL-Toyota Center for AI Research and HN was partially supported Samsung Fellowship. JCD was also partially supported by the National Science Foundation award NSF-CAREER-1553086.

References

[1] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

[2] M. Aitkin and D. B. Rubin. Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society, Series B*, pages 67–75, 1985.

[3] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28:131–142, 1966.

[4] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, and G. Chen. Deep speech 2: end-to-end speech recognition in english and mandarin. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 173–182, 2016.

[5] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

[6] A. Asuncion and D. J. Newman. UCI machine learning repository, 2007. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

[7] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

[8] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems 20*, pages 137–144, 2007.

[9] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

[10] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

[11] A. Ben-Tal, D. den Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[12] D. P. Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, 1973.

[13] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *Mathematical Programming, Series A*, 167(2):235–292, 2018. URL `http://arxiv.org/abs/1401.0212`.

[14] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[15] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *arXiv:1604.01446 [math.PR]*, 2016. URL `https://arxiv.org/abs/1604.01446`.

[16] J. Blanchet, Y. Kang, and K. Murthy. Robust wasserstein profile inference and applications to machine learning. *arXiv:1610.05627 [math.ST]*, 2016.

[17] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128. Association for Computational Linguistics, 2006.

[18] S. L. Blodgett, L. Green, and B. O'Connor. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of Empirical Methods for Natural Language Processing*, pages 1119–1130, 2016.

[19] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: a Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[20] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[21] P. Bühlmann and N. Meinshausen. Magging: maximin aggregation for inhomogeneous large-scale data. *Proceedings of the IEEE*, 104(1):126–135, 2016.

[22] Z. Cai, J. Fan, and R. Li. Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, 95(451):888–902, 2000.

[23] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

[24] R. Caruana. Multitask learning. In *Learning to Learn*, pages 95–133. Springer, 1998.

[25] N. Cressie and T. R. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, pages 440–464, 1984.

[26] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientifica Mathematica Hungary*, 2:299–318, 1967.

[27] H. Daume III and D. Marcu. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126, 2006.

[28] T. E. de Campos, B. R. Babu, and M. Varma. Character recognition in natural images. In *Proceedings of the Fourth International Conference on Computer Vision Theory and Applications*, February 2009.

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

[30] J. S. Denker, W. R. Gardner, H. P. Graf, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, H. S. Baird, and I. Guyon. Neural network recognizer for hand-written zip code digits. In *Advances in Neural Information Processing Systems 1*, 1988.

[31] J. C. Duchi. Introductory lectures on stochastic convex optimization. In *The Mathematics of Data*, IAS/Park City Mathematics Series. American Mathematical Society, 2018.

[32] J. C. Duchi, P. W. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv:1610.03425 [stat.ML]*, 2016.

[33] Y. C. Eldar, A. Ben-Tal, and A. Nemirovski. Linear minimax regret estimation of deterministic parameters with bounded data uncertainties. *IEEE Transactions on Signal Processing*, 52(8): 2177–2188, 2004.

[34] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming, Series A*, to appear, 2017.

[35] J. Fan and W. Zhang. Statistical estimation in varying coefficient models. *Annals of Statistics*, 27(5):1491–1518, 1999.

[36] M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.

[37] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[38] P. J. Grother, G. W. Quinn, and P. J. Phillips. Report on the evaluation of 2d still-image face recognition algorithms. *NIST interagency report*, 7709:106, 2010.

[39] D. J. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14, 2006.

[40] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Springer, New York, 1993.

[41] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I & II*. Springer, New York, 1993.

[42] D. Hovy and A. Søgaard. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, volume 2, pages 483–488, 2015.

[43] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems 20*, pages 601–608, 2007.

[44] P. J. Huber. *Robust Statistics*. John Wiley and Sons, New York, 1981.

[45] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. John Wiley and Sons, second edition, 2009.

[46] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093 [cs.CV]*, 2014.

[47] R. Jiang and Y. Guan. Data-driven chance constrained stochastic program. *Optimization Online*, 2013. URL `http://www.optimization-online.org/DB_FILE/2013/09/4044.pdf`.

[48] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2, page 1, 2011.

[49] A. J. King and R. J. Wets. Epi-consistency of convex stochastic programs. *Stochastics and Stochastic Reports*, 34(1-2):83–92, 1991.

[50] P. A. Krokhmal. Higher moment coherent risk measures. *Quantitative Finance*, 7(4):373–387, 2007.

[51] H. Lam. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1275, 2016.

[52] H. Lam and E. Zhou. The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Operations Research Letters*, 45(4):301–307, 2017.

[53] L. Le Cam and G. L. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2000.

[54] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1 (4):541–551, 1989.

[55] D. Luenberger. *Optimization by Vector Space Methods*. Wiley, 1969.

[56] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, 2004.

[57] N. Meinshausen and P. Bühlmann. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015.

[58] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems 29*, 2016.

[59] H. Namkoong and J. C. Duchi. Variance regularization with convex objectives. In *Advances in Neural Information Processing Systems 30*, 2017.

[60] A. Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120, 1990.

[61] G. Pflug and D. Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442, 2007.

[62] M. Redmond and A. Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

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

[64] R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*. Springer, New York, 1998.

[65] D. Rothenhäusler, N. Meinshausen, and P. Bühlmann. Confidence intervals for maximin effects in inhomogeneous large-scale data. In *Statistical Analysis for High-Dimensional Data*, pages 255–277. Springer, 2016.

[66] D. Rothenhäusler, P. Bühlmann, N. Meinshausen, and J. Peters. Anchor regression: heterogeneous data meets causality. *arXiv:1801.06229 [stat.ME]*, 2018.

[67] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *Proceedings of the European Conference on Computer Vision*, pages 213–226. Springer, 2010.

[68] P. Sapiezynski, V. Kassarnig, and C. Wilson. Academic performance prediction in a gender-imbalanced environment. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, volume 1, pages 48–51, 2017.

[69] S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems 28*, pages 1576–1584, 2015.

[70] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM and Mathematical Programming Society, 2009.

[71] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[72] A. Sinha, H. Namkoong, and J. C. Duchi. Certifiable distributional robustness with principled adversarial training. *arXiv:1710.10571 [stat.ML]*, 2017.

[73] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

[74] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems 21*, pages 1433–1440, 2008.

[75] M. Talagrand. A new look at independence. *Annals of Probability*, 24(1):1–34, 1996.

[76] R. Tatman. Gender and dialect bias in You Tube's automatic captions. In *First Workshop on Ethics in Natural Langauge Processing*, volume 1, pages 53–59, 2017.

[77] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528. IEEE, 2011.

[78] Y. Tsuboi, H. Kashima, S. Hido, S. Bickel, and M. Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155, 2009.

[79] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[80] M. Udell, K. Mohan, D. Zeng, J. Hong, S. Diamond, and S. Boyd. Convex optimization in Julia. In *First Workshop on High Performance Technical Computing in Dynamic Languages*, pages 18–28. IEEE, 2014.

[81] Y. Usui and K. Kondo. The sift image feature reduction method using the histogram intersection kernel. In *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pages 517–520. IEEE, 2009.

[82] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[83] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

[84] T. van Erven and P. Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[85] A. Wald. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, 46(2):265–280, 1945.

[86] D. Wozabal. A framework for optimization under ambiguity. *Annals of Operations Research*, 193(1):21–47, 2012.

[87] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, 1997.

[88] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, 2005.

A Proof of Duality Results

A.1 Proof of Proposition 1

Fix any $\theta \in \Theta$ and let $Z(x) = \ell(\theta; x)$ to simplify notation. Let us consider the likelihood ratio formulation (3). Introducing Lagrange multiplier $\lambda \geq 0$ for the constraint $\int f(L)dP \leq \rho$ and $\eta \in \mathbb{R}$ for $\mathbb{E}_P[L] = 1$, we obtain the Lagrangian

$$\mathcal{L}(L, \lambda, \eta) = \int_{\mathcal{X}} \left[(Z(x) - \eta) L(x) - \lambda f(L(x)) \right] dP(x) + \lambda\rho + \eta.$$

Then taking $L \equiv 1$, we have that $\int f(L)dP = 0$ and $\mathbb{E}_P[L] = 1$, so the extended Slater condition holds. Thus we have (see, e.g., Luenberger [55, Theorem 8.6.1 and Problem 8.7]) that

$$\begin{aligned}
&\sup_{Q \ll P} \left\{ \mathbb{E}_Q[Z] : D_f\left(Q \| P\right) \leq \rho \right\} \\
&= \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \sup_{L \geq 0} \left\{ \int_{\mathcal{X}} \left[\left(Z(x) - \eta \right) L(x) - \lambda f(L(x)) \right] dP(x) + \lambda \rho + \eta \right\}. \qquad (19)
\end{aligned}$$

Next, we wish to interchange the inner supremum over all (measurable) nonnegative functions $L : \mathcal{X} \to \mathbb{R}$ and the integral in the dual (19). In this case, the function $f^* : \mathbb{R} \to \overline{\mathbb{R}}$ is closed convex and continuous on its domain, and we have $\sup_{\ell \geq 0} \{ (\frac{z-\eta}{\lambda}\ell - f(\ell) \} = f^*(\frac{z-\eta}{\lambda})$. Moreover, standard results [41] give that $\ell \in \partial f^*(\frac{z-\eta}{\lambda})$ if and only if $\frac{z-\eta}{\lambda}\ell - f(\ell) = f^*(\frac{z-\eta}{\lambda})$. Now, as $x \mapsto Z(x)$ is measurable and f^* is convex (and hence measurable), the set-valued mapping $x \mapsto \partial f^*(\frac{Z(x)-\eta}{\lambda})$ is measurable [64, Thm. 14.13]. Consequently, assuming that

$$P\left(f^*\left(\frac{Z(X) - \eta}{\lambda} \right) = +\infty \right) = 0,$$

we may construct a measurable function $x \mapsto L_*(x) \in \partial f^*(\frac{Z(x)-\eta}{\lambda})$, in which case the inner supremum (19) becomes

$$\begin{aligned}
\sup_{L \geq 0} \int_{\mathcal{X}} \left[(Z(x) - \eta) L(x) - \lambda f(L(x)) \right] &= \lambda \int_{\mathcal{X}} \left[\frac{Z(x) - \eta}{\lambda} L_*(x) - f(L_*(x)) \right] dP(x) \\
&= \lambda \int_{\mathcal{X}} f^*\left(\frac{Z(x) - \eta}{\lambda} \right) dP(x),
\end{aligned}$$

where we have used $f(t) = +\infty$ for $t < 0$. The attainment result follows from standard infinite dimensional duality results [55, Chapter 8].

Assume now that $P(f^*(\frac{Z(X)-\eta}{\lambda}) = +\infty) > 0$. Using our w.l.o.g. assumption that $f(t) \geq 0$, if $f^*(s) = +\infty$ then $\lim_{t \to \infty} \{st - f(t)\} = \infty$. Then we may take a sequence $L_n : \mathcal{X} \to \mathbb{R}_+$ with $L_n(x) = n$ if $x \in \{x \mid f^*(\frac{Z(x)-\eta}{\lambda}) = +\infty\}$, a measurable set by the above, and $L_n(x) = 0$ otherwise. The integrals $\int [(Z(x) - \eta) L_n(x) - \lambda f(L_n(x))] dP(x) \to \infty$ as $n \uparrow \infty$.

A.2 Proof of Lemma 1

First, we compute the Fenchel conjugate for Cressie-Read family of divergences f_k.

Lemma 3.

$$f_k^*(s) = \frac{1}{k} \left((k-1)s + 1 \right)_+^{k^*} - \frac{1}{k} \qquad (20)$$

Proof Consider the supremum $f^*(s) = \sup_t \{st - f(t)\}$. Then for $t \geq 0$, we have

$$\frac{\partial}{\partial t} \left[st - f_k(t) \right] = s - \frac{1}{k-1}(t^{k-1} - 1).$$

If $s < 0$, then the supremum is attained at $t = 0$, as the derivative above is < 0 at $t = 0$. If $s \geq -\frac{1}{k-1}$, then we solve $\frac{\partial}{\partial t} \left[st - f_k(t) \right] =$ to find $t = \left((k-1)s + 1 \right)^{1/(k-1)}$, and substituting gives

$$st - f(t) = \frac{1}{k} \left((k-1)s + 1 \right)^{\frac{k}{k-1}} - \frac{1}{k}$$

which is our desired result as $1 - 1/k = 1/k_*$. $\square$

From the dual formulation (5), we have

$$\begin{aligned}\sup_{P \ll P_0} \left\{\mathbb{E}_P[Z] \text{ s.t. } D_f\left(P \| P_0\right) \le \rho\right\} &= \inf_{\lambda \ge 0, \eta} \left\{\lambda \mathbb{E}_{P_0} f^*\left(\frac{Z-\eta}{\lambda}\right) + \lambda\rho + \eta\right\} \\ &= \inf_{\lambda \ge 0, \eta} \left\{\frac{(k-1)^{k_*}}{k} \lambda^{1-k_*} \mathbb{E}_{P_0}\left(Z - \eta + \frac{\lambda}{k-1}\right)_+^{k_*} + \lambda(\rho - \frac{1}{k}) + \eta\right\} \\ &= \inf_{\lambda \ge 0, \tilde{\eta}} \left\{(k-1)^{k_*} k^{-1} \mathbb{E}_{P_0}\left(Z - \tilde{\eta}\right)_+^{k_*} \lambda^{1-k_*} + \left(\rho + \frac{1}{k(k-1)}\right)\lambda + \tilde{\eta}\right\}\end{aligned}$$

where the last line followed by setting $\tilde{\eta} := \eta - \frac{\lambda}{k-1}$. Taking derivatives with respect to λ to infimize the preceding expression, we have (noting that $(k_* - 1)/k_* = 1/k$)

$$\lambda = (k-1)(k(k-1)\rho + 1)^{-\frac{1}{k_*}} \left(\mathbb{E}_{P_0}\left(Z - \tilde{\eta}\right)_+^{k_*}\right)^{\frac{1}{k_*}} \tag{21}$$

By substituting into the preceding expression, we find that the supremum is

$$\inf_{\tilde{\eta}} \left(k(k-1)\rho + 1\right)^{\frac{1}{k}} \left(\mathbb{E}_{P_0}\left(Z - \tilde{\eta}\right)_+^{k_*}\right)^{1/k_*} + \tilde{\eta}.$$

B Proof of Upper Bounds

B.1 Proof of Theorem 2

To ease notation, fix any $\theta \in \Theta$ and let $Z(x) = \ell(\theta; x)$, so $\mathcal{R}_k(Z; \widehat{P}_n) = \sup_{p: D_f(p \| \widehat{P}_n) \le \rho} \langle p, Z_1^n \rangle$ by definition. Let $g_n(z_1^n) = \sup_{p: D_f(p \| \widehat{P}_n) \le \rho} \langle p, z_1^n \rangle$ so that $g_n(Z_1^n) = \mathcal{R}_k(Z; \widehat{P}_n)$. We show that g_n is concentrates around its expectation, which we do via convex Lipschitz concentration [19, 75]. The subgradients of g_n are

$$p_* \in \operatorname*{argmax}_{p: D_f(P \| \mathbf{1}/n) \le \rho} \langle p, z \rangle,$$

that is, all vectors $p^* \in \mathcal{P}_{n,k}$ attaining the supremum [40, Corollary 4.4.4]. From the f-divergence constraint, we have that

$$\sum_{i=1}^n p_i^k \le n^{1-k}(k(k-1)\rho + 1) = n^{1-k} c_k^k.$$

When $k \in (1, 2]$, since $\|\cdot\|_k \ge \|\cdot\|_2$, we have $\|p\|_2 \le n^{-\frac{1}{k_*}} (k(k-1)\rho + 1)^{\frac{1}{k}}$. For $k \ge 2$, we have from Holder's inequality that

$$\sum_{i=1}^n p_i^2 \le \left(\sum_{i=1}^n p_i^k\right)^{\frac{2}{k}} n^{1-\frac{2}{k}} \le n^{-1}(k(k-1)\rho + 1)^{\frac{2}{k}} = n^{-1} c_k^2.$$

In particular, we obtain that g is Lipschitz with respect to the ℓ_2-norm with Lipschitz constant $c_k n^{-\frac{1}{k_*}}$ when $k \in (1, 2]$ and $c_k n^{-\frac{1}{2}}$ when $k \ge 2$.

We recall the standard convex Lipschitz concentration inequality for bounded random variables.

Lemma 4 (Boucheron et al. 2013, Theorem 6.10)**.** *Let $h : \mathbb{R}^n \to \mathbb{R}$ be convex or concave and L-Lipschitz with respect to the ℓ_2-norm. Let Z_i be independent random variables with $Z_i \in [a, b]$. For $t \geq 0$,*

$$\mathbb{P}(|h(Z_1^n) - \mathbb{E}[h(Z_1^n)]| \geq t) \leq 2 \exp\left(-\frac{t^2}{2L^2(b-a)^2}\right).$$

Using Lemma 4 with g_n, we obtain via our Lipschitz argument that for $Z_i \in [0, M]$,

$$\mathbb{P}\Big(|\mathcal{R}_k(Z; \widehat{P}_n) - \mathbb{E}_{P_0}[\mathcal{R}_k(Z; \widehat{P}_n)]| \geq t\Big) \leq 2 \exp\left(\frac{-\min\left\{n, n^{2/k_*}\right\} t^2}{2c_k^2 M^2}\right),$$

or equivalently, with probability at least $1 - 2e^{-t}$

$$\left|\mathcal{R}_k(Z; \widehat{P}_n) - \mathbb{E}_{P_0}[\mathcal{R}_k(Z; \widehat{P}_n)\right| \leq c_k M n^{-\frac{1}{k_* \vee 2}} \left(\frac{2}{k} + t\right). \tag{22}$$

To show concentration for the empirical version of our robust problem (1), it remains to show that $\mathbb{E}_{P_0}[\mathcal{R}_k(Z; \widehat{P}_n)] - \mathcal{R}_k(Z; P_0)$ is small. First, we note that from Proposition 1, we have

$$\begin{aligned}
\mathbb{E}_{P_0}\left[\mathcal{R}_k(Z; \widehat{P}_n)\right] &= \mathbb{E}_{P_0}\left[\inf_{\lambda \geq 0, \eta \in \mathbb{R}} \left\{\frac{1}{n}\sum_{i=1}^n \lambda f^*\left(\frac{Z_i - \eta}{\lambda}\right) + \lambda\rho + \eta\right\}\right] \\
&\leq \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \mathbb{E}_{P_0}\left[\frac{1}{n}\sum_{i=1}^n \lambda f^*\left(\frac{Z_i - \eta}{\lambda}\right) + \lambda\rho + \eta\right] = \mathcal{R}_k(Z; P_0),
\end{aligned} \tag{23}$$

so we always have the upper that the expectation of the supremum is less than the supremum of expectations. To see the other direction, the following lemma provides a good lower bound. We defer its proof to Section B.1.1.

Lemma 5. *Let $k_* \in [1, \infty)$ and let Y_i be an i.i.d. sequence of random variables satisfying $\mathbb{E}[|Y|^{2k_*}] \leq C^{k_*}\mathbb{E}[|Y|^{k_*}]$ for some $C \in \mathbb{R}_+$. For any $k_* \in [1, \infty)$, we have*

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n |Y_i|^{k_*}\right)^{\frac{1}{k_*}}\right] \geq \mathbb{E}[|Y|^{k_*}]^{\frac{1}{k_*}} - \frac{2}{k}\sqrt{C} n^{-\frac{1}{k_* \vee 2}} \tag{24}$$

We now wish to apply Lemma 5 to the dual representation (8). First, we show that we can restrict η to be nonnegative in the dual form (8).

Claim 6. *Let P be an arbitrary probability measure on $(\mathcal{X}, \mathcal{A})$. Assume that $\ell(\theta; x) \in [0, M]$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$. Then, for $k \in (1, \infty)$ and $k_* = k/(k-1) \in (1, \infty)$, and any $\rho > 0$, we have for all $\theta \in \Theta$*

$$\mathcal{R}_k(\theta; P) = \inf_{\eta \geq 0}\left\{c_k \mathbb{E}_P\left[(\ell(\theta; X) - \eta)_+^{k_*}\right]^{\frac{1}{k_*}} + \eta\right\}.$$

Proof of Claim Fix an arbitrary $\theta \in \Theta$, and to ease notation, define

$$g_k(\eta; \theta, P) := c_k \mathbb{E}_P\left[(\ell(\theta; X) - \eta)_+^{k_*}\right]^{\frac{1}{k_*}} + \eta.$$

Since $\ell(\theta; x) \geq 0$, $\eta \mapsto g_k(\eta; \theta, P)$ is differentiable on $(-\infty, 0)$ with its derivative given by

$$g_k'(\eta; \theta, P) = -c_k \frac{\mathbb{E}[(\ell(\theta; X) - \eta)^{k_* - 1}]}{(\mathbb{E}[(\ell(\theta; X) - \eta)^{k_*}])^{\frac{1}{k}}} + 1.$$

From Jensen's inequality, we have $\mathbb{E}[(\ell(\theta;X)-\eta)^{k_*-1}] \le \left(\mathbb{E}[(\ell(\theta;X)-\eta)^{k_*}]\right)^{\frac{1}{k}}$ so that $g_k'(\eta;\theta,P) \le -c_k + 1 < 0$ for $\eta \in (-\infty, 0)$. Hence, $\eta \mapsto g_k(\eta;\theta,P)$ is decreasing on $(-\infty,0)$ and the claim follows. □

Now, noting that $(Z-\eta)_+/M \in [0,1]$ for $\eta \ge 0$, we apply Lemma 5 with $C=1$ to obtain

$$\mathbb{E}_{P_0}\left[\left(\mathbb{E}_{\widehat{P}_n}(Z-\eta)_+^{k_*}\right)^{\frac{1}{k_*}}\right] \ge \left(\mathbb{E}_{P_0}(Z-\eta)_+^{k_*}\right)^{\frac{1}{k_*}} - \frac{2}{k}Mn^{-\frac{1}{k_*\vee 2}}$$

for $\eta \ge 0$. Multiplying c_k and adding η on both sides, and taking the infimum over $\eta \ge 0$, we obtain $\mathbb{E}_{P_0}[\mathcal{R}_k(Z;\widehat{P}_n)] \ge \mathcal{R}_k(Z;P_0) - \frac{2c_k}{k}M\max\left\{n^{-\frac{1}{k_*}}, n^{-\frac{1}{2}}\right\}$. Combining this with the bound (23), we have

$$\mathbb{E}_{P_0}[\mathcal{R}_k(Z;\widehat{P}_n)] \le \mathcal{R}_k(Z;P_0) \le \mathbb{E}_{P_0}[\mathcal{R}_k(Z;\widehat{P}_n)] + \frac{2c_k}{k}Mn^{-\frac{1}{k_*\vee 2}}.$$

Combining the above bound with the concentration inequality (22) gives Theorem 2.

B.1.1 Proof of Lemma 5

First, we claim that it suffices to show

$$\begin{aligned}\mathbb{E}[|Y|^q]^{1/q} &\ge \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n |Y_i|^q\right)^{1/q}\right] \\ &\ge \mathbb{E}[|Y|^q]^{1/q} - 2\frac{q-1}{q}\begin{cases}(C^{q/2}\vee 1)\cdot n^{-1/q} & \text{if } q \ge 2\\ (C \vee C^{1-q/2})\cdot n^{-1/2} & \text{if } q < 2.\end{cases}\end{aligned} \tag{25}$$

where the last inequality holds for $n \ge C^q$ when $q \ge 2$. To see how our desired bound (24) follows from (25), we use a quick scaling argument. Let $\alpha > 0$, and note that $\mathbb{E}[|\alpha Y|^{2q}] \le (C\alpha^2)^q\mathbb{E}[|Y|^q]$ by assumption. Let $\sigma_n := \mathbb{E}[(\frac{1}{n}\sum_{i=1}^n |Y_i|^q)^{1/q}]$ and $\sigma = \mathbb{E}[|Y|^q]^{1/q}$ for shorthand. First, if $q \ge 2$, we have $(\alpha^2 C)^{q/2} \ge 1$ if $\alpha \ge C^{-\frac{1}{2}}$, and we obtain

$$\alpha\sigma_n \ge \alpha\sigma - 2\frac{q-1}{q}\alpha^q C^{q/2}n^{-1/q} \;\text{ or }\; \sigma_n \ge \sigma - 2\frac{q-1}{q}C^{q/2}\alpha^{q-1}n^{-1/q}.$$

Choosing $\alpha = C^{-\frac{1}{2}}$ gives the result (24) when $q \ge 2$. For $q < 2$, we similarly obtain that $C\alpha^2 \ge (C\alpha^2)^{1-q/2}$ for $\alpha \ge C^{-\frac{1}{2}}$, whence we have the lower bound

$$\alpha\sigma_n \ge \alpha\sigma - 2\frac{q-1}{q}C\alpha^2 n^{-1/2} \;\text{ or }\; \sigma_n \ge \sigma - 2\frac{q-1}{q}C\alpha n^{-1/2}$$

for $\alpha \ge C^{-\frac{1}{2}}$. Choosing $\alpha = C^{-\frac{1}{2}}$ thus gives the desired result (24).

Now, we proceed to show the bound (25). Let

$$\gamma_n = \operatorname*{argmin}_{\gamma \ge 0}\left\{\frac{1}{(q-1)}\frac{\frac{1}{n}\sum_{i=1}^n |Y_i|^q}{\gamma^{q-1}} + \gamma\right\} = \left(\frac{1}{n}\sum_{i=1}^n |Y_i|^q\right)^{1/q}$$

so that

$$\frac{1}{q}\frac{\frac{1}{n}\sum_{i=1}^n |Y_i|^q}{\gamma_n^{q-1}} + \frac{(q-1)\gamma_n}{q} = \left(\frac{1}{q}+\frac{q-1}{q}\right)\left(\frac{1}{n}\sum_{i=1}^n |Y_i|^q\right)^{1/q} = \left(\frac{1}{n}\sum_{i=1}^n |Y_i|^q\right)^{1/q}.$$

For any $\gamma \geq 0$ we have by the first order inequality for convexity (as the function $\gamma \mapsto 1/\gamma^{q-1} + \gamma$ is convex for $\gamma \geq 0$) that

$$\left(\frac{1}{n}\sum_{i=1}^n |Y_i|^q\right)^{1/q} = \frac{\frac{1}{n}\sum_{i=1}^n |Y_i|^q}{q\gamma_n^{q-1}} + \frac{q-1}{q}\gamma_n$$
$$\geq \frac{\frac{1}{n}\sum_{i=1}^n |Y_i|^q}{q\gamma^{q-1}} + \frac{q-1}{q}\gamma + \left(\frac{q-1}{q} - \frac{(q-1)\frac{1}{n}\sum_{i=1}^n |Y_i|^q}{q\gamma^q}\right)(\gamma_n - \gamma). \tag{26}$$

We now show how to provide a bound on magnitude of the final term in expression (26).

Let $\sigma^q = \mathbb{E}[|Y|^q]$, and choose $\gamma^q = \max\{n^{-\alpha}, \sigma^q\}$, where $\alpha \geq 0$ is a power to be chosen. Then

$$\mathbb{E}\left[\left(\frac{q-1}{q} - \frac{(q-1)\frac{1}{n}\sum_{i=1}^n |Y_i|^q}{q\gamma^q}\right)^2\right] = \left(\frac{q-1}{q}\right)^2 \mathbb{E}\left[\left(1 - \frac{\sigma^q}{\gamma^q} + \frac{\sigma^q}{\gamma^q} - \frac{\frac{1}{n}\sum_{i=1}^n |Y_i|^q}{\gamma^q}\right)^2\right]$$
$$= \left(\frac{q-1}{q}\right)^2 \left[(1 - \sigma^q/\gamma^q)^2 + \frac{1}{\gamma^{2q}n}\mathrm{Var}(|Y|^q)\right],$$

and noting that $\mathrm{Var}(|Y|^q) \leq \mathbb{E}[|Y|^{2q}] \leq C^q\mathbb{E}[|Y|^q] = C^q\sigma^q$, we have

$$\frac{1}{\gamma^{2q}n}\mathrm{Var}(|Y|^q) \leq \frac{1}{n}\frac{C^q\sigma^q}{\max\{n^{-2\alpha}, \sigma^{2q}\}} = C^q \min\left\{\frac{\sigma^q}{n^{1-2\alpha}}, \frac{1}{n\sigma^q}\right\}.$$

and

$$1 - \frac{\sigma^q}{\gamma^q} = 1 - \min\left\{n^\alpha\sigma^q, 1\right\} = \left(1 - n^\alpha\sigma^q\right)_+.$$

Now we provide an upper bound on the $(\gamma_n - \gamma)$ term in the product in inequality (26). By inspection, we have

$$(\gamma_n - \gamma)^2 = \left(\frac{1}{n}\sum_{i=1}^n |Y_i|^q\right)^{\frac{2}{q}} - 2\gamma\gamma_n + \max\left\{n^{-\alpha}, \sigma^q\right\}^{\frac{2}{q}}$$
$$\leq \left(\frac{1}{n}\sum_{i=1}^n |Y_i|^q\right)^{\frac{2}{q}} + \max\left\{n^{-\alpha}, \sigma^q\right\}^{\frac{2}{q}}. \tag{27}$$

We now state a useful intermediate lemma and consequential inequality, deferring its proof to Section B.1.2.

Lemma 7. *Let $q \in [1, 2]$ and $a \in [1, 2]$. Then for any random variable $X \geq 0$,*

$$\mathbb{E}[X^{aq}] \leq \mathbb{E}[X^q]^{2-a}\mathbb{E}[X^{2q}]^{a-1}.$$

As an immediate consequence of Lemma 7, we see that for $q \in [1, 2]$ and non-negative random variables X, we have that if $\mathbb{E}[X^{2q}] \leq C^q\sigma^q$, where $\mathbb{E}[X^q] = \sigma^q$, then

$$\mathbb{E}[X^2] \leq C^{2-q}\sigma^q. \tag{28}$$

To see this, substitute $a = 2/q \in [1, 2]$ in Lemma 7, which yields

$$\mathbb{E}[X^2] = \mathbb{E}[X^{aq}] \leq \mathbb{E}[X^q]^{2-\frac{2}{q}}\mathbb{E}[X^{2q}]^{\frac{2}{q}-1} \leq \sigma^{2q-2}(C^q\sigma^q)^{\frac{2}{q}-1} = C^{2-q}\sigma^q.$$

Returning to our bound on $(\gamma_n - \gamma)$, we find via inequality (28) that

$$\begin{aligned}\mathbb{E}[(\gamma_n - \gamma)^2] &\le \mathbb{E}[|Y|^2] + \max\{n^{-2\alpha/q}, \sigma^2\} \\ &\le \begin{cases} \sigma^2 + \max\{n^{-2\alpha/q}, \sigma^2\} & \text{if } q \ge 2 \\ C^{2-q}\sigma^q + \max\{n^{-2\alpha/q}, \sigma^2\} & \text{if } q < 2 \end{cases} \\ &\le 2 \begin{cases} \max\{n^{-2\alpha/q}, \sigma^2\} & \text{if } q \ge 2 \\ \max\{C^{2-q}\sigma^q, n^{-2\alpha/q}\} & \text{if } q < 2, \end{cases}\end{aligned}$$

where we have used that for $q < 2$ we have

$$\sigma^2 = \mathbb{E}[Y^q]^{2/q} \le \mathbb{E}[Y^2] \le C^{2-q}\sigma^q.$$

In particular, we have by Hölder's inequality that

$$\begin{aligned}\mathbb{E}\left[\left(1 - \frac{\frac{1}{n}\sum_{i=1}^n |Y_i|^q}{\gamma^q}\right)(\gamma_n - \gamma)\right]^2 &\le \mathbb{E}\left[\left(1 - \frac{\frac{1}{n}\sum_{i=1}^n |Y_i|^q}{\gamma^q}\right)^2\right] \mathbb{E}[(\gamma_n - \gamma)^2] \qquad (29) \\ &\le 2\left((1 - n^\alpha\sigma^q)_+^2 + C^q \min\left\{\frac{\sigma^q}{n^{1-2\alpha}}, \frac{1}{n\sigma^q}\right\}\right) \cdot \begin{cases} \max\{n^{-2\alpha/q}, \sigma^2\} & \text{if } q \ge 2 \\ \max\{n^{-2\alpha/q}, C^{2-q}\sigma^q\} & \text{if } q < 2. \end{cases}\end{aligned}$$

We now state a lemma, whose proof we defer to Section B.1.3, which gives us our desired result.

Lemma 8. *For any $\sigma \ge 0$, we have*

$$(1 - n^\alpha\sigma^q)_+^2 \cdot \begin{cases} \max\{n^{-2\alpha/q}, \sigma^2\} & \text{if } q \ge 2 \\ \max\{n^{-2\alpha/q}, C^{2-q}\sigma^q\} & \text{if } q < 2. \end{cases} \le \begin{cases} n^{-2\alpha/q} & \text{if } q \ge 2 \\ C^{2-q}\min\{\sigma^q, n^{-\alpha}\} & \text{if } q < 2. \end{cases} \qquad (30a)$$

and

$$C^q \min\left\{\frac{\sigma^q}{n^{1-2\alpha}}, \frac{1}{n\sigma^q}\right\} \cdot \begin{cases} \max\{n^{-2\alpha/q}, \sigma^2\} & \text{if } q \ge 2 \\ \max\{n^{-2\alpha/q}, C^{2-q}\sigma^q\} & \text{if } q < 2. \end{cases} \le \begin{cases} C^q \frac{1}{n^{1-\alpha+2\alpha/q}} & \text{if } q \ge 2 \\ \max\left\{\frac{C^2}{n}, \frac{C^q}{n^{1-\alpha+2\alpha/q}}\right\} & \text{if } q < 2. \end{cases} \qquad (30b)$$

We now use Lemma 8 to give the remainder of the proof. First, consider the case that $q \ge 2$. Then choosing $\alpha = 1$ we have $\gamma^q = \max\{n^{-1}, \sigma^q\}$, and

$$\left|\mathbb{E}\left[\left(1 - \frac{\frac{1}{n}\sum_{i=1}^n |Y_i|^q}{\gamma^q}\right)(\gamma_n - \gamma)\right]\right|^2 \le 2\left[C^q n^{(1-2/q)\alpha - 1} + n^{-(2/q)\alpha}\right] = \frac{2(1 + C^q)}{n^{2/q}} \le 4\frac{C^q \vee 1}{n^{2/q}}.$$

When $q < 2$, we similarly choose $\alpha = 1$, which yields

$$\left|\mathbb{E}\left[\left(1 - \frac{\frac{1}{n}\sum_{i=1}^n |Y_i|^q}{\gamma^q}\right)(\gamma_n - \gamma)\right]\right|^2 \le 2\max\left\{\frac{C^2}{n}, \frac{C^q}{n^{2/q}}\right\} + 2\frac{C^{2-q}}{n}.$$

(Asymptotically, then, we obtain $4\max\{C^2, C^{2-q}\}/n$.) Referring to inequality (26), we thus have

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n |Y_i|^q\right)^{1/q}\right] \ge \mathbb{E}[|Y|^q]^{1/q} - 2\frac{q-1}{q}\begin{cases} (C^{q/2} \vee 1) \cdot n^{-1/q} & \text{if } q \ge 2 \\ (C \vee C^{1-q/2}) \cdot n^{-1/2} & \text{if } q < 2, \end{cases}$$

which was the desired result.

B.1.2 Proof of Lemma 7

For any random variable X, we know that for $\gamma \in [0,1]$ and any conjugates $p, q \geq 1$, that is, $1/p + 1/q = 1$, we have by Hölder's inequality that

$$\mathbb{E}[X] = \mathbb{E}[X^{\gamma} X^{1-\gamma}] \leq \mathbb{E}[X^{\gamma p}]^{1/p} \mathbb{E}[X^{(1-\gamma)q}]^{1/q}.$$

Now, let $X = Y^{aq}$, and take $1/p = 2 - a$ and $1/q = a - 1$. Then we have for any $\gamma \in [0,1]$ that

$$\mathbb{E}[Y^{aq}] \leq \mathbb{E}[Y^{\frac{\gamma aq}{2-a}}]^{2-a} \mathbb{E}[Y^{\frac{(1-\gamma)aq}{a-1}}]^{a-1}.$$

If we take $\gamma = \frac{2-a}{a} \in [0,1]$, then we obtain

$$\frac{\gamma a}{2-a} = 1 \;\; \text{and} \;\; (1-\gamma)\frac{a}{a-1} = \frac{2(a-1)}{a}\frac{a}{a-1} = 2.$$

This gives the result of the lemma.

B.1.3 Proof of Lemma 8

We begin with inequality (30a). If $\sigma^q \geq n^{-\alpha}$, the result is trivial, as $(1 - n^{\alpha}\sigma^q)_+ = 0$. So we assume that $\sigma^q < n^{-\alpha}$, which implies that $\sigma^2 \geq n^{-2\alpha/q}$, and we know that (for $q < 2$) $C^{2-q}\sigma^q \geq \sigma^2$. Thus, when $q < 2$, we have $\max\{n^{-2\alpha/q}, C^{2-q}\sigma^q\} = C^{2-q}\sigma^q \leq C^{2-q}n^{-\alpha}$. If $q \geq 2$ and $\sigma^q \leq n^{-\alpha}$, then $\sigma^2 \leq n^{-2\alpha/q}$, so that $\max\{\sigma^2, n^{-2\alpha/q}\} = n^{-2\alpha/q}$.

Now we turn to inequality (30b). First, let us assume that $q \geq 2$. In this case, we have that if $\sigma^q \leq n^{-\alpha}$, then the left-hand expression of (30b) has bound

$$C^q \min\left\{\frac{\sigma^q}{n^{1-2\alpha}}, \frac{1}{n\sigma^q}\right\} n^{-2\alpha/q} = C^q \frac{\sigma^q}{n^{1-2\alpha+2\alpha/q}} \leq C^q \frac{1}{n^{1-\alpha+2\alpha/q}}.$$

On the other hand, for $\sigma^q \geq n^{-\alpha}$, we have

$$C^q \min\left\{\frac{\sigma^q}{n^{1-2\alpha}}, \frac{1}{n\sigma^q}\right\} \sigma^2 = C^q \frac{\sigma^2}{n\sigma^q} = C^q \frac{1}{n\sigma^{q-2}} \leq C^q \frac{1}{n^{1-\alpha+2\alpha/q}},$$

as $q \geq 2$ and $\sigma \geq n^{-\alpha/q}$. In the case that $q < 2$ in inequality (30b), we are left bounding

$$\min\left\{\frac{\sigma^q}{n^{1-2\alpha}}, \frac{1}{n\sigma^q}\right\} \max\{n^{-2\alpha/q}, C^{2-q}\sigma^q\}.$$

Assume first that $n^{-2\alpha/q} \geq C^{2-q}\sigma^q$, or $\sigma^q \leq C^{q-2}n^{-2\alpha/q}$. In this case, the σ maximizing the left minimum is $\sigma^q = \min\{n^{-\alpha}, C^{q-2}n^{-2\alpha/q}\}$, which gives

$$\min\left\{\frac{\sigma^q}{n^{1-2\alpha}}, \frac{1}{n\sigma^q}\right\} \max\{n^{-2\alpha/q}, C^{2-q}\sigma^q\} \leq \frac{1}{n^{1-\alpha+2\alpha/q}}.$$

On the other hand, when $C^{2-q}\sigma^q \geq n^{-2\alpha/q}$, we obtain that we must maximize (over σ) the quantity

$$C^2 \min\left\{\frac{\sigma^{2q}}{n^{1-2\alpha}}, \frac{1}{n}\right\} \leq C^2 \frac{1}{n}.$$

This gives the desired result.

B.2 Proof of Corollary 1

Let $\mathcal{F} := \{\ell(\theta;\cdot) : \mathcal{X} \to \mathbb{R} \quad \text{for} \quad \theta \in \Theta\}$ be our function class. Fix $t > 0$ and let $N = N(\frac{t}{3}, \mathcal{F}, \|\cdot\|_{L^\infty(\mathcal{X})})$ to ease notation, so there exists $\{\theta_1, \cdots, \theta_N\} \subset \Theta$ such that $\{\ell(\theta_1;\cdot), \cdots, \ell(\theta_N;\cdot)\}$ is a $\frac{t}{3}$-cover of $\mathcal{F}$. For any $\theta \in \Theta$, let $i(\theta)$ be such that $\left\|\ell(\theta;\cdot) - \ell(\theta_{i(\theta)};\cdot)\right\|_{L^\infty(\mathcal{X})} \le \frac{t}{3}$. We have

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\mathcal{R}_k(\theta;\widehat{P}_n) - \mathcal{R}_k(\theta;P_0)\right| \\
&\le \sup_{\theta\in\Theta}\left\{\left|\mathcal{R}_k(\theta;\widehat{P}_n) - \mathcal{R}_k(\theta_{i(\theta)};\widehat{P}_n)\right| + \left|\mathcal{R}_k(\theta_{i(\theta)};\widehat{P}_n) - \mathcal{R}_k(\theta_{i(\theta)};P_0)\right| + \left|\mathcal{R}_k(\theta_{i(\theta)};P_0) - \mathcal{R}_k(\theta;P_0)\right|\right\} \\
&\le \max_{i=1,\ldots,N}\left|\mathcal{R}_k(\theta_i;\widehat{P}_n) - \mathcal{R}_k(\theta_i;P_0)\right| + \frac{2t}{3},
\end{aligned}$$

where we have used that $\{\ell(\theta_i;\cdot)\}_{i=1}^N$ is a $t/3$ cover of $\mathcal{F}$. A union bound now implies

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\left|\mathcal{R}_k(\theta;\widehat{P}_n) - \mathcal{R}_k(\theta;P_0)\right|\right) \le N \max_{i=1,\ldots,N}\mathbb{P}\left(\left|\mathcal{R}_k(\theta_i;\widehat{P}_n) - \mathcal{R}_k(\theta_i;P_0)\right| \ge t/3\right).$$

Applying Theorem 2 to each θ_i, we obtain the desired result.

C Proof of Lower Bounds

C.1 Proof of Theorem 3

In this proof and the coming proofs related to Section 5.1, we define

$$\mathfrak{M}_n^{\text{est}} := \inf_{\widehat{R}}\sup_{P\in\mathcal{P}}\mathbb{E}_{P_0}\left[\left|\widehat{R}(Z_1^n) - \mathcal{R}_k(Z)\right|\right]$$

for shorthand, and use it without comment.

Consider the canonical two point hypothesis testing problem between distributions P_0 and P_1: nature first chooses $v \in \{0,1\}$, then conditioned on v draws $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} P_v$. Assuming that $|\mathcal{R}_k(P_0) - \mathcal{R}_k(P_1)| \ge \delta > 0$ for some δ, Le Cam's classical reduction from estimation to testing [53, 87] yields that

$$\mathfrak{M}_n^{\text{est}} \ge \frac{\delta}{2}\left(1 - \|P_0^n - P_1^n\|_{\text{TV}}\right). \tag{31}$$

We use the bound (31) to give the lower bound by choosing P_0 and P_1 so that $\|P_0^n - P_1^n\|_{\text{TV}} \le \frac{1}{2}$ and δ is as large as possible.

First, we show the $O(n^{-\frac{1}{2}})$ lower bound. We begin with a technical

Lemma 9. *Let $c_k = (1 + k(k-1)\rho)^{\frac{1}{k}}$, $p_k = c_k^{-k/(k-1)}$, and $\beta_k := \frac{1}{2}(1 - c_k^{-k})$. For a pair $z_0 \le z_1$, let Z be such that*

$$Z = \begin{cases} z_0 & \text{w.p. } 1-p \\ z_1 & \text{w.p. } p. \end{cases}$$

If $p \ge p_k$, we have $\mathcal{R}_k(Z) = z_1$, and if $p \le p_k$, we have $\mathcal{R}_k(Z) \le c_k p^{\frac{1}{k_*}} z_1 + (1 - c_k p^{\frac{1}{k_*}})z_0$. Further, if $p \le p_k \wedge (1 - (1-\beta)^{1-k_*}p_k)$ for some $\beta \in (0,1)$, then $\mathcal{R}_k(Z) \ge \beta^{\frac{1}{k}} c_k p^{\frac{1}{k_*}} z_1 + (1 - \beta^{\frac{1}{k}} c_k p^{\frac{1}{k_*}})z_0$.

See Section C.1.1 for a proof.

Now, consider the two distributions $Z_1 \sim P_1$, $Z_2 \sim P_2$

$$Z_1 = \begin{cases} 0 & \text{w.p. } 1 - p_k - \delta \\ M & \text{w.p. } p_k + \delta \end{cases}, \qquad Z_2 = \begin{cases} 0 & \text{w.p. } 1 - p_k + \delta \\ M & \text{w.p. } p_k - \delta \end{cases}$$

for some $0 < \delta \le p_k \wedge (1 - p_k)$ to be chosen later. Note that $p_k = c_k^{-k_*} < 1$ as $c_k > 1$. We use the version of Z_1 and Z_2 such that $Z_1(\cdot)$ and $Z_2(\cdot)$ are upper semi-continuous.

From Lemma 9, we have that $\mathcal{R}_k(Z_1) = M$ and $\mathcal{R}_k(Z_2) \le M c_k (p_k - \delta)^{\frac{1}{k_*}}$. Consequently, P_1 and P_2 are separated in the robust objective

$$|\mathcal{R}_k(Z_1) - \mathcal{R}_k(Z_2)| \ge M(1 - c_k(p_k - \delta)^{\frac{1}{k_*}}) \ge \frac{c_k^{k_*}}{k_*} M\delta$$

where we used Taylor's theorem

$$c_k(p_k - \delta)^{\frac{1}{k_*}} = c_k(c_k^{-k_*} - \delta)^{\frac{1}{k_*}} \le c_k \left(c_k^{-1} - \frac{1}{k_*} c_k^{\frac{k_*}{k}} \delta \right) = 1 - \frac{1}{k_*} c_k^{k_*} \delta.$$

It suffices to show that $\|P_1^n - P_2^n\|_{\mathrm{TV}} \le \frac{1}{2}$ for $\delta = \sqrt{\frac{p_k(1-p_k)}{8n}} \wedge \frac{1}{2}(1 - p_k) \wedge p_k$. By Pinsker's inequality, we have $\|P_1^n - P_2^n\|_{\mathrm{TV}}^2 \le \frac{n}{2} D_{\mathrm{kl}}(P_2 \| P_1)$ so it is enough to show $D_{\mathrm{kl}}(P_2 \| P_1) \le \frac{1}{n}$ for the given value of δ. To this end, we note that for $\delta \le \frac{1}{2}(1 - p_k)$,

$$D_{\mathrm{kl}}(P_2 \| P_1) = (1 - p_k + \delta) \log \frac{1 - p_k + \delta}{1 - p_k - \delta} + (p_k - \delta) \log \frac{p_k - \delta}{p_k + \delta} \le \frac{8\delta^2}{p_k(1 - p_k)}.$$

Setting $\delta = \sqrt{\frac{p_k(1-p_k)}{8n}} \wedge \frac{1}{2}(1 - p_k) \wedge p_k$, we then have that $D_{\mathrm{kl}}(P_2 \| P_1) \le \frac{1}{n}$.

For the second $O(n^{-\frac{1}{k_*}})$ bound, consider the random variables $Z_1 \sim P_1$ and $Z_2 \sim P_2$ with

$$Z_1 \equiv 0, \qquad Z_2 = \begin{cases} 0 & \text{w.p. } 1 - \delta \\ M & \text{w.p. } \delta \end{cases}$$

for some $\delta > 0$ to be choosen later. We have $\mathcal{R}_k(Z_1) = 0$ trivially, and since $1 - (1 - \beta)^{1-k_*} p_k > 0 \equiv 1 - c_k^{-k} > \beta$ holds for $\beta_k = \frac{1}{2}(1 - c_k^{-k})$, we have

$$\mathcal{R}_k(Z_2) \ge M \beta_k^{\frac{1}{k}} c_k \delta^{\frac{1}{k_*}}$$

for $0 < \delta \le p_k \wedge (1 - (1 - \beta_k)^{1-k_*} p_k)$ by Lemma 9. This gives the the separation $|\mathcal{R}_k(P_1) - \mathcal{R}_k(P_2)| \ge M \beta_k^{\frac{1}{k}} c_k \delta^{\frac{1}{k_*}}$.

Noting that

$$D_{\mathrm{kl}}(P_1 \| P_2) = -\log(1 - \delta) \le \frac{\delta}{1 - \delta} \le 2\delta$$

for $\delta \le \frac{1}{2}$, we obtain

$$\mathfrak{M}_n^{\mathrm{est}} \ge \frac{1}{4} M \beta_k^{\frac{1}{k}} c_k \delta^{\frac{1}{k_*}} \left(1 - \sqrt{\frac{n}{2} D_{\mathrm{kl}}(P_1 \| P_2)} \right) \ge \frac{1}{8} M c_k \beta_k^{\frac{1}{k}} \delta^{\frac{1}{k_*}}$$

where in the first inequality we used the reduction (31) and Pinsker's inequality as before. The desired result follows by setting $\delta = \frac{1}{4n} \wedge p_k \wedge (1 - (1 - \beta_k)^{1-k_*} p_k)$.

C.1.1 Proof of Lemma 9

Define the objective function in the dual representation (8) as

$$g(\eta) := c_k \left((1-p) (z_0 - \eta)_+^{k_*} + p (z_1 - \eta)_+^{k_*} \right)^{\frac{1}{k_*}} + \eta.$$

Taking subgradients, we obtain

$$\partial g(\eta) = \begin{cases} 1 & \text{if } \eta > z_1 \\ [1 - c_k p^{\frac{1}{k_*}}, 1] & \text{if } \eta = z_1 \\ 1 - c_k p^{\frac{1}{k_*}} & \text{if } z_0 \le \eta < z_1 \\ 1 - c_k \frac{(1-p)(z_0-\eta)^{\frac{1}{k-1}} + p(z_1-\eta)^{\frac{1}{k-1}}}{\left((1-p)(z_0-\eta)^{k_*} + p(z_1-\eta)^{k_*}\right)^{\frac{1}{k}}} & \text{if } \eta < z_0. \end{cases}$$

If $c_k p^{\frac{1}{k_*}} \ge 1$ then $\eta^* = \operatorname{argmin}_\eta g(\eta)$ is attained at z_1 by convexity, and $R(P) = g(\eta^*) = z_1$. If $c_k p^{\frac{1}{k_*}} < 1$, we have $\eta^* \le z_0$ so that

$$g(\eta^*) \le g(z_0) = c_k p^{\frac{1}{k_*}} z_1 + (1 - c_k p^{\frac{1}{k_*}}) z_0,$$

which gives the second claim.

For the second inequality, noting that

$$\mathcal{R}_k(Z) = z_0 + (z_1 - z_0) \sup \left\{ q \in [0,1] : (1-p)^{1-k}(1-q)^k + p^{1-k} q^k \le c_k^k \right\},$$

it suffices to show that $q = \beta^{\frac{1}{k}} c_k p^{\frac{1}{k_*}}$ is feasible when $p \le 1 - (1-\beta)^{1-k_*} p_k$. Indeed, we have

$$(1-p)^{1-k}(1 - \beta^{\frac{1}{k}} c_k p^{\frac{1}{k_*}})^k + p^{1-k}(\beta^{\frac{1}{k}} c_k p^{\frac{1}{k_*}})^k \le (1-p)^{1-k} + \beta c_k^k \le c_k^k$$

where we used $(1-p)^{1-k} \le (1-\beta) c_k^k$ in the last inequality.

C.2 Proof of Proposition 4

We proceed by LeCam's method as in Theorem 3. Let $Z_1 \sim P_1$, $Z_2 \sim P_2$ have distribution

$$Z_1 = \begin{cases} 0 & \text{w.p. } 1-p \\ M & \text{w.p. } p, \end{cases} \qquad Z_2 = \begin{cases} 0 & \text{w.p. } 1-p-\delta \\ M & \text{w.p. } p+\delta \end{cases}$$

for some $\delta \in (0,1)$ to be chosen later. As before, we show that $\mathcal{R}_f(Z_1)$ and $\mathcal{R}_f(Z_2)$ are well-separated but P_1 and P_2 are close in total variation distance.

By definition, we have

$$\mathcal{R}_f(Z_1) = \sup \{ qM : h_f(q;p) \le \rho, q \in [0,1] \} = M q(p)$$

and similarly, $\mathcal{R}_f(Z_2) = M q(p+\delta)$. For δ small enough, the implicit function theorem applies to $h_f(q(p), p) = 0$ by our hypothesis. Consequently, we $q(\cdot)$ is continuously differentiable on a neighbhorhood of p with

$$q'(p) = -\frac{\partial_p h_f(q(p);p)}{\partial_q h_f(q(p);p)} > 0,$$

where strict positivity follows by the strict convexity the we assume in the proposition. Taylor's theorem implies

$$\mathcal{R}_f(Z_2) - \mathcal{R}_f(Z_1) = q(p+\delta) - q(p) = q'(p)\delta + o(\delta)$$

as $\delta \to 0$.

We now pick δ such that $\|P_1^n - P_2^n\|_{\mathrm{TV}} \le \frac{1}{2}$. By Pinsker's inequality and standard KL vs. χ^2-divergence inequalities [79, Lemmas 2.5–2.7], we have $\|P_1^n - P_2^n\|_{\mathrm{TV}}^2 \le \frac{n}{2} D_{\mathrm{kl}}(P_1 \| P_2)$; we will choose δ such that $D_{\mathrm{kl}}(P_1 \| P_2) \le \frac{1}{n}$. For $\delta \in [0, p]$, Lemma 2.7 of [79] yields

$$D_{\mathrm{kl}}(P_1 \| P_2) \le \frac{\delta^2}{p} + \frac{\delta^2}{1-p} = \frac{\delta^2}{p(1-p)}.$$

Setting $\delta_n = \sqrt{\frac{p(1-p)}{n}}$, we obtain from the reduction from estimation to hypothesis testing (31) that

$$\mathfrak{M}_n^{\mathrm{est}} \ge \frac{M}{8} q'(p) \sqrt{\frac{p(1-p)}{n}} + o\left(\frac{1}{\sqrt{n}}\right),$$

which gives the result.

C.3 Proof of Proposition 5

We use LeCam's method and proceed similarly as in the second part of Section C.1. Consider the two distributions $Z_1 \sim P_1$, $Z_2 \sim P_2$ with

$$Z_1 \equiv 0, \qquad Z_2 = \begin{cases} 0 & \text{w.p. } 1-\delta \\ M & \text{w.p. } \delta, \end{cases}$$

where we set $\delta = \frac{1}{2(n \vee C_{f,\rho,m})}$. Then $\mathcal{R}_f(Z_1) = 0$, and to show separation of $\mathcal{R}_f(Z_2)$, we require a bit of work, beginning with the following lemma.

Lemma 10. *For $\delta = \frac{1}{2(n \vee C_{f,\rho,m})}$, define Q by $Q(Z = M) = \left(\frac{\rho}{2m}\right)^{\frac{1}{k}} \delta^{\frac{1}{k_*}}$ and $Q(Z = 0) = 1 - Q(Z = M)$. Then $D_f(Q \| P_2) \le \rho$.*

Proof We have

$$\begin{aligned} &\delta f\left(\frac{\left(\frac{\rho}{2m}\right)^{\frac{1}{k}} \delta^{\frac{1}{k_*}}}{\delta}\right) + (1-\delta) f\left(\frac{1 - \left(\frac{\rho}{2m}\right)^{\frac{1}{k}} \delta^{\frac{1}{k_*}}}{1-\delta}\right) \\ &\overset{(a)}{\le} \delta f\left(\left(\frac{\rho}{2m}\right)^{\frac{1}{k}} \delta^{-\frac{1}{k}}\right) + (1-\delta) f\left(1 - \left(\frac{\rho}{2m}\right)^{\frac{1}{k}} \delta^{\frac{1}{k_*}}\right) \overset{(b)}{\le} \delta f\left(\left(\frac{\rho}{2m}\right)^{\frac{1}{k}} \delta^{-\frac{1}{k}}\right) + \frac{\rho}{2} \end{aligned}$$

where in step (a), we used that f is non-increasing on $(0,1)$ along with $\frac{1 - \left(\frac{\rho}{2m}\right)^{1/k} \delta^{1/k_*}}{1-\delta} \in (0,1)$, and in step (b), we used the definition of $f^{-1}(s) = \inf\{t \in [0,1] : f(t) \le s\}$.

Next, note that since $\left(\frac{\rho}{2m}\right)^{\frac{1}{k}} \delta^{-\frac{1}{k}} \ge \left\{(n \vee C_{f,\rho,m}) \rho m^{-1}\right\}^{\frac{1}{k}}$ for the given range of δ, we have $f((\frac{\rho}{2m})^{1/k} \delta^{-1/k}) \le \frac{\rho}{2\delta}$ by hypothesis. We conclude that $D_f(Q \| P_2) \le \rho$. $\square$

As a consequence of Lemma 10, we have $\mathcal{R}_f(Z_2) \ge M(\frac{\rho}{2m})^{1/k} \delta^{1/k_*}$. As $\mathcal{R}_f(Z_1) = 0$, we have $|\mathcal{R}_f(Z_1) - \mathcal{R}_f(Z_2)| \ge M(\frac{\rho}{2m})^{1/k} \delta^{1/k_*}$. Proceeding similarly as in the last paragraph of Section C.1 we obtain the result.

C.4 Proof of Theorem 6

Define the optimization distance between two distributions P_0 and P_1 (cf. [1, 31]) by

$$\mathrm{d}_{\mathrm{opt}}(P_0, P_1; f) := \sup \left\{ \delta \geq 0 : \begin{array}{l} \mathcal{R}_f(\theta; P_0) \leq \mathcal{R}_f(\theta_0^*; P_0) + \delta \quad \text{implies } \mathcal{R}_f(\theta; P_1) \geq \mathcal{R}_f(\theta_1^*; P_1) + \delta \\ \mathcal{R}_f(\theta; P_1) \leq \mathcal{R}_f(\theta_1^*; P_1) + \delta \text{ implies } \mathcal{R}_f(\theta; P_0) \geq \mathcal{R}_f(\theta_0^*; P_0) + \delta \end{array} \right\}$$

where $\theta_v \in \operatorname{argmin}_{\theta \in \Theta} \mathcal{R}_f(\theta; P_v)$. With this result, we have the following standard lemma, which is a reduction of optimization to testing.

We have the following reduction from distributionally robust optimization to hypothesis testing, which is based on Le Cam's two-point hypothesis testing reduction.

Lemma 11 (Chs. 5.1–5.2[31])**.** *Assume that $P_0, P_1 \in \mathcal{P}$ are such that $d_{\mathrm{opt}}(P_0, P_1; f) \geq \delta$. Then*

$$\mathfrak{M}_n(\mathcal{P}, f, \ell) \geq \frac{\delta}{2} \left(1 - \|P_0^n - P_1^n\|_{\mathrm{TV}}\right).$$

With this inequality in hand, we proceed by We first show the $\Omega(n^{-\frac{1}{2}})$ lower bound. Consider the two distributions $X_1 \sim P_1$, $X_2 \sim P_2$ with

$$X_1 = \begin{cases} -1 & \text{w.p. } 1 - p_k - \delta \\ \epsilon & \text{w.p. } p_k + \delta, \end{cases} \qquad X_2 = \begin{cases} -1 & \text{w.p. } 1 - p_k + \delta \\ \epsilon & \text{w.p. } p_k - \delta \end{cases}$$

where $\epsilon = \frac{\delta}{2k_* p_k}$ for some $0 < \delta \leq p_k \wedge (1 - p_k)$ to be choosen later. Note that

$$\mathcal{R}_k(\theta; P) = \begin{cases} \theta \sup_{Q \ll P} \left\{ \mathbb{E}_Q[X] : D_f\left(Q \| P\right) \leq \rho \right\} & \text{if } \theta \geq 0 \\ \theta \inf_{Q \ll P} \left\{ \mathbb{E}_Q[X] : D_f\left(Q \| P\right) \leq \rho \right\} & \text{if } \theta < 0. \end{cases}$$

For $\delta \leq 1 - 2p_k$, we from Lemma 9 that $\mathcal{R}_k(\theta; P_1) = -\theta \mathbf{1}\left\{\theta < 0\right\} + \epsilon\theta \mathbf{1}\left\{\theta \geq 0\right\}$ and $\mathcal{R}_k(\theta; P_2) = -\theta$ when $\theta < 0$. Now, we have $\mathcal{R}_k(\theta; P_2) \leq -\epsilon\theta$ when $\theta \geq 0$ since

$$\begin{aligned} \sup_{Q \ll P_2} \left\{ \mathbb{E}_Q[X_2] : D_f\left(Q \| P_2\right) \leq \rho \right\} &\leq \epsilon c_k (p_k - \delta)^{\frac{1}{k_*}} + (c_k (p_k - \delta)^{\frac{1}{k_*}} - 1) \\ &\leq \epsilon c_k (p_k - \delta)^{\frac{1}{k_*}} - \frac{\delta}{k_* p_k} \leq \epsilon - \frac{\delta}{k_* p_k} = -\epsilon. \end{aligned} \tag{32}$$

Here, we used Taylor's theorem

$$c_k (p_k - \delta)^{\frac{1}{k_*}} = c_k (c_k^{-k_*} - \delta)^{\frac{1}{k_*}} \leq c_k \left(c_k^{-1} - \frac{1}{k_*} c_k^{\frac{k_*}{k}} \delta \right) = 1 - \frac{1}{k_*} c_k^{k_*} \delta.$$

If we let $\theta_i^\star := \operatorname{argmin}_{\theta \in \Theta} \mathcal{R}_k(\theta; P_i)$ for $i = 1, 2$, we have $\theta_1^\star = 0$, $\theta_2^\star = M$ and $\mathcal{R}_k(\theta_1^\star; P_1) = 0$, $\mathcal{R}_k(\theta_2^\star; P_2) \leq -M\epsilon$. We then have the following lemma.

Lemma 12. *Let the above conditions hold. Then $d_{\mathrm{opt}}(P_1, P_2; f_k) \geq \frac{\epsilon}{2} M$.*

Proof Let $\theta \in [-M, M]$ be such that $\mathcal{R}_k(\theta; P_1) \leq \mathcal{R}_k(\theta_1^\star; P_1) + M\kappa$ for some $\kappa \in [0, \frac{\epsilon}{2}]$. From $\mathcal{R}_k(\theta; P_1) - \mathcal{R}_k(\theta_1^\star; P_1) = \mathcal{R}_k(\theta; P_1) = -\theta \mathbf{1}\left\{\theta < 0\right\} + \epsilon\theta \mathbf{1}\left\{\theta > 0\right\} \leq M\kappa$, we have $-\kappa \leq \frac{\theta}{M} \leq \frac{\kappa}{\epsilon}$.

Applying this bound, we obtain

$$\begin{aligned}\mathcal{R}_k(\theta;P_2)-\mathcal{R}_k(\theta_2^\star;P_2) &= \begin{cases}(\theta-M)\sup_{Q\ll P_2}\{\mathbb{E}_Q[X_2]: D_f(Q\|P_2)\le\rho\} & \text{if } \theta\ge 0\\ -\theta - M\sup_{Q\ll P_2}\{\mathbb{E}_Q[X_2]: D_f(Q\|P_2)\le\rho\} & \text{if } \theta<0\end{cases}\\ &\ge -\theta\mathbf{1}\{\theta<0\}-\epsilon\theta\mathbf{1}\{\theta\ge 0\}+M\epsilon\\ &\ge -\theta\mathbf{1}\{\theta<0\}-M\kappa\mathbf{1}\{\theta\ge 0\}+M\epsilon\ge\frac{M\epsilon}{2}\ge M\kappa\end{aligned}$$

where we used the bound (32) to get the second inequality.

On the other hand, assume $\mathcal{R}_k(\theta;P_2)\le\mathcal{R}_k(\theta_2^\star;P_2)+M\kappa$. In this case, we claim that $\theta\ge 0$ necessarily. Indeed, if $\theta<0$, then using the bound (32),

$$\mathcal{R}_k(\theta;P_2)=-\theta\le\mathcal{R}_k(\theta_2^\star;P_2)+M\kappa\le -M\epsilon+M\kappa=-M(\epsilon-\kappa)<0$$

which yields a contradiction. Now, from $\theta\ge 0$ and $\mathcal{R}_k(\theta;P_2)\le\mathcal{R}_k(\theta_2^\star;P_2)+M\kappa$, we again obtain from the bound (32)

$$M\kappa\ge(\theta-\theta_2^\star)\sup_{Q\ll P_2}\{\mathbb{E}_Q[X]: D_f(Q\|P_2)\le\rho\}\ge\epsilon(M-\theta).$$

Hence, we have $\theta\ge M\left(1-\frac{\kappa}{\epsilon}\right)$, and

$$\mathcal{R}_k(\theta;P_1)=\epsilon\theta\ge\epsilon M\left(1-\frac{\kappa}{\epsilon}\right)=M(\epsilon-\kappa)\ge\frac{M\epsilon}{2}\ge\mathcal{R}_k(\theta_1^\star;P_1)+M\kappa$$

for $\kappa\in[0,\frac{\epsilon}{2}]$. We conclude that the claimed separation in d_{opt} holds. □

Now, we argue as in the proof of Theorem 3. Noting that $D_{\mathrm{kl}}(P_1\|P_2)\le\frac{\delta^2}{p_k(1-p_k)}$ (e.g. [79, Lemma 2.7]) for $0\le\delta\le(1-p_k)$, let $\delta=\sqrt{\frac{p_k(1-p_k)}{2n}}\wedge\frac{1}{2}(1-p_k)\wedge(1-2p_k)\wedge p_k$. Then Lemma 11 yields

$$\mathfrak{M}_n(\mathcal{P},f_k,\ell)\ge\frac{M\epsilon}{4}\left(1-\sqrt{\frac{n}{2}D_{\mathrm{kl}}(P_1'\|P_2')}\right)\ge\frac{M\delta}{8k_*p_k},$$

which gives the first result of the theorem.

Next, we show the second $\Omega(n^{-\frac{1}{k_*}})$ lower bound. Consider the distributions $X_1\sim P_1$, $X_2\sim P_2$

$$X_1\equiv-\epsilon,\qquad X_2=\begin{cases}-\epsilon & \text{w.p. } 1-\delta\\ 1 & \text{w.p. } \delta\end{cases}$$

where $\epsilon:=\frac{1}{2}\beta_k^{\frac{1}{k}}c_k\delta^{\frac{1}{k_*}}$ for some $0<\delta\le p_k\wedge(1-p_k)\wedge(1-(1-\beta_k)^{1-k_*}p_k)$ to be choosen later. Now, we again show that $\mathrm{d}_{\mathrm{opt}}(P_1,P_2;f_k)\ge\frac{\epsilon}{2}$. To this end, first observe that $\mathcal{R}_k(\theta;P_1)=-\epsilon\theta$. From the first part of Lemma 9, we have $\mathcal{R}_k(\theta;P_2)=-\epsilon\theta\ge 0$ when $\theta<0$. For $\theta\ge 0$, the last inequality in Lemma 9 gives

$$\mathcal{R}_k(\theta;P_2)\ge\beta_k^{\frac{1}{k}}c_k\delta^{\frac{1}{k_*}}\theta-(1-\beta_k^{\frac{1}{k}}c_k\delta^{\frac{1}{k_*}})\epsilon\theta=\left((1+\epsilon)\beta_k^{\frac{1}{k}}c_k\delta^{\frac{1}{k_*}}-\epsilon\right)\theta\ge\epsilon\theta$$

since $\epsilon=\frac{1}{2}\beta_k^{\frac{1}{k}}c_k\delta^{\frac{1}{k_*}}$. Denoting $\theta_i^\star:=\mathrm{argmin}_{\theta\in\Theta}\mathcal{R}_k(\theta;P_i)$ again, we consequently obtain $\theta_1^\star=M$, $\theta_2^\star=0$ with $\mathcal{R}_k(\theta_1^\star;P_1)=-M\epsilon$, $\mathcal{R}_k(\theta_2^\star;P_2)=0$.

Next, we show $\mathrm{d}_{\mathrm{opt}}(P_1, P_2; f_k) \geq \frac{M\epsilon}{2}$. Assume that $\theta \in [-M, M]$ satisfies $\mathcal{R}_k(\theta; P_1) \leq \mathcal{R}_k(\theta_1^\star; P_1) + M\kappa = -M\epsilon + M\kappa \;\equiv\; \theta \geq M\left(1 - \frac{\kappa}{\epsilon}\right)$ for some $\kappa \in [0, \frac{\epsilon}{2}]$. This implies

$$\mathcal{R}_k(\theta; P_2) \geq M\epsilon\left(1 - \frac{\kappa}{\epsilon}\right) \geq M\frac{\epsilon}{2} = \mathcal{R}_k(\theta_2^\star; P_2) + \frac{M\epsilon}{2} \geq \mathcal{R}_k(\theta_2^\star; P_2) + M\kappa.$$

On the other hand, if $\mathcal{R}_k(\theta; P_2) \leq \mathcal{R}_k(\theta_2^\star; P_2) + M\kappa$ then $\epsilon|\theta| \leq \mathcal{R}_k(\theta; P_2) \leq M\kappa$ so that $|\theta| \leq \frac{M\kappa}{\epsilon}$. Consequently, we have

$$\begin{aligned}\mathcal{R}_k(\theta; P_1) &= -\epsilon\theta \geq -M\kappa = M(-\epsilon + \epsilon - \kappa) \\ &\geq M\left(-\epsilon + \frac{\epsilon}{2}\right) = \mathcal{R}_k(\theta_1^\star; P_1) + \frac{M\epsilon}{2} \geq \mathcal{R}_k(\theta_1^\star; P_1) + M\kappa\end{aligned}$$

and we conclude $\mathrm{d}_{\mathrm{opt}}(P_1, P_2; f_k) \geq \frac{M\epsilon}{2}$.

Proceeding as in the proof of the second part of Theorem 3, we note that $D_{\mathrm{kl}}(P_1 \| P_2) \leq 2\delta$ when $\delta \leq \frac{1}{2}$. Setting $\delta = \frac{1}{4n} \wedge p_k \wedge (1 - (1 - \beta_k)^{1-k_*} p_k)$, we conclude

$$\mathcal{M}_n^{\mathrm{opt}} \geq \frac{M}{16}\beta_k^{\frac{1}{k}} c_k \delta^{\frac{1}{k_*}} = \frac{M}{16}\beta_k^{\frac{1}{k}} c_k \left(\frac{1}{4n} \wedge p_k \wedge (1 - (1 - \beta_k)^{1-k_*} p_k)\right)^{\frac{1}{k_*}}.$$

C.5 Proof of Proposition 7

For p given by hypothesis, recall the definition (15) of $q(p)$. Following the same logic as in the proof of Proposition 4, the implicit function theorem implies that $q(\cdot)$ is continuously differentiable near p with

$$q'(p) = \frac{-\partial_p h_f(q(p); p)}{\partial_q h_f(q(p); p)} > 0,$$

where $h_f(q; p) = pf(\frac{q}{p}) + (1 - p)f(\frac{1-q}{1-p})$ as before. From Taylor's theorem, we then have

$$q(p + \delta) = q(p) + q'(p)\delta + r(\delta)$$

for a remainder $r(\delta) = o(\delta)$ as $\delta \to 0$. For small $\delta > 0$, define

$$\epsilon_\delta := \left(q(p) + \frac{1}{2}\left(q'(p)\delta + r(\delta)\right)\right)^{-1} - 1 > 0.$$

We use the reduction from robust optimization to testing of Lemma 11. For some $\delta \in (0, q(p) - p)$ to be choosen later, consider the two distributions $X_1 \sim P_1$, $X_2 \sim P_2$ with

$$X_1 = \begin{cases} -1 & \text{w.p. } 1 - p \\ \epsilon_\delta & \text{w.p. } p, \end{cases} \qquad X_2 = \begin{cases} -1 & \text{w.p. } 1 - p - \delta \\ \epsilon_\delta & \text{w.p. } p + \delta. \end{cases}$$

For $\ell(\theta; X) = \theta X$, we show that $\theta \mapsto \mathcal{R}_f(\theta; P_1)$ and $\theta \mapsto \mathcal{R}_f(\theta; P_2)$ are well-separated in the distance $\mathrm{d}_{\mathrm{opt}}(\cdot, \cdot)$, but P_1 and P_2 are close in total variation distance. By definition

$$\mathcal{R}_f(\theta; P_1) = \begin{cases} -\theta(1 - (1 + \epsilon_\delta)q(p)) & \text{if } \theta \geq 0 \\ -\theta(-\epsilon_\delta + (1 + \epsilon_\delta)q(1 - p)) & \text{otherwise,} \end{cases}$$

and similarly,

$$\mathcal{R}_f(\theta; P_2) = \begin{cases} -\theta(1 - (1 + \epsilon_\delta)q(p + \delta)) & \text{if } \theta \geq 0 \\ -\theta(-\epsilon_\delta + (1 + \epsilon_\delta)q(1 - p - \delta)) & \text{otherwise.} \end{cases}$$

By our choice of ϵ_δ, observe

$$1 - (1+\epsilon_\delta)q(p) > 0, \qquad \text{but} \qquad 1 - (1+\epsilon_\delta)q(p+\delta) \le 0,$$

and $q(p) > p$ so that $q(p) > p + \delta$ for small δ, and similarly $q(1-p-\delta) + \delta > 1 - p$. Consequently, $1 + \epsilon_\delta < \frac{1}{q(p)} < \frac{1}{p+\delta} < \frac{1}{1-q(1-p-\delta)}$, and so

$$-\epsilon_\delta + (1+\epsilon_\delta)q(1-p) \ge -\epsilon_\delta + (1+\epsilon_\delta)q(1-p-\delta) > 0.$$

Thus, we have $\mathcal{R}'_f(\theta; P_1) < 0$ for all θ, while $\mathcal{R}'_f(\theta; P_2) > 0$ for $\theta > 0$ and $\mathcal{R}'_f(\theta; P_2) < 0$ for $\theta < 0$. We conclude that $\theta_i^\star := \operatorname{argmin}_{\theta \in [-M, M]} \mathcal{R}_f(\theta; P_i)$ satisfies $\theta_1^\star = M$ and $\theta_2^\star = 0$.

We now show $\mathrm{d}_{\mathrm{opt}}(P_1, P_2) \ge M\Delta_\delta$, where

$$\Delta_\delta := \frac{q'(p)\delta + r(\delta)}{4\left(q(p) + \frac{1}{2}(q'(p)\delta + r(\delta))\right)} = \frac{1}{4}(1+\epsilon_\delta)\left(q'(p)\delta + r(\delta)\right).$$

In the sequel, we use the following identities to simplify computation:

$$2\Delta_\delta = 1 - (1+\epsilon_\delta)q(p), \quad \text{and} \quad -2\Delta_\delta = 1 - (1+\epsilon_\delta)q(p+\delta).$$

First, for any $\kappa \in [0, \Delta_\delta]$, consider θ such that

$$\mathcal{R}_f(\theta; P_1) \le \mathcal{R}_f(\theta_1^\star; P_1) + M\kappa.$$

Assume for contradiction that $\theta < 0$: the above bound implies

$$\begin{aligned}\theta &\ge \frac{M}{-\epsilon_\delta + (1+\epsilon_\delta)q(1-p)}\left(1 - (1+\epsilon_\delta)q(p) - \kappa\right) \\ &= \frac{M}{-\epsilon_\delta + (1+\epsilon_\delta)q(1-p)}\left(2\Delta_\delta - \kappa\right) \ge 0.\end{aligned}$$

For $\theta \ge 0$, the optimality bound implies

$$\theta \ge M\left(1 - \frac{\kappa}{1 - (1+\epsilon_\delta)q(p)}\right) = M\left(1 - \frac{\kappa}{2\Delta_\delta}\right).$$

Using this bound, we obtain

$$\begin{aligned}\mathcal{R}_f(\theta; P_2) - \mathcal{R}_f(\theta_2^\star; P_2) = \mathcal{R}_f(\theta; P_2) &= -(1 - (1+\epsilon_\delta)q(p+\delta))\theta = 2M\Delta_\delta\left(1 - \frac{\kappa}{2\Delta_\delta}\right) \\ &\ge M\left(2\Delta_\delta - \kappa\right) \ge M\kappa.\end{aligned}$$

Next, for any $\kappa \in [0, \Delta_\delta]$, consider θ such that

$$\mathcal{R}_f(\theta; P_2) \le \mathcal{R}_f(\theta_2^\star; P_2) + M\kappa = M\kappa,$$

which implies $\theta \le -\frac{M\kappa}{1-(1+\epsilon_\delta)q(p+\delta)}$ if $\theta \ge 0$, and $\theta \ge \frac{M\kappa}{-\epsilon_\delta + (1+\epsilon_\delta)q(1-p-\delta)}$ if $\theta < 0$. When $\theta \ge 0$, we then obtain

$$\mathcal{R}_f(\theta; P_1) - \mathcal{R}_f(\theta_1^\star; P_1) = (1 - (1+\epsilon_\delta)q(p))(M - \theta) \ge 2M\Delta_\delta\left(1 + \frac{\kappa}{2\Delta_\delta}\right) \ge M\kappa.$$

When $\theta < 0$, we get

$$\mathcal{R}_f(\theta; P_1) - \mathcal{R}_f(\theta_1^\star; P_1) \ge M\kappa\left(\frac{-\epsilon_\delta + (1+\epsilon_\delta)q(1-p)}{-\epsilon_\delta + (1+\epsilon_\delta)q(1-p-\delta)} + 2\right) \ge M\kappa.$$

We thus conclude that $\mathrm{d}_{\mathrm{opt}}(P_1, P_2) \geq M\Delta_\delta$ as claimed.

We now pick δ such that $\|P_1^n - P_2^n\|_{\mathrm{TV}} \leq \frac{1}{2}$. By Pinsker's inequality, we have $\|P_1^n - P_2^n\|_{\mathrm{TV}}^2 \leq \frac{n}{2} D_{\mathrm{kl}}\left(P_1 \| P_2\right)$, and letting $\delta_n = \sqrt{\frac{p(1-p)}{n}}$, we get $D_{\mathrm{kl}}\left(P_1 \| P_2\right) \leq \frac{1}{n}$ as for $\delta \in [0, p]$, we have as usual that $D_{\mathrm{kl}}\left(P_1 \| P_2\right) \leq \frac{\delta^2}{p(1-p)}$. From the reduction from distributionally robust optimization to hypothesis testing (Lemma 11), we conclude

$$\mathcal{M}_n^{\mathrm{opt}} \geq \frac{M}{4} \Delta_{\delta_n}.$$

Multiplying both sides by $\sqrt{n}$ and taking $n \to \infty$, we obtain the result.

C.6 Proof of Proposition 8

We proceed as in the second part of Section C.4. We use Lemma 11 on the distributions $X_1 \sim P_1$, $X_2 \sim P_2$

$$X_1 \equiv -\epsilon, \qquad X_2 = \begin{cases} -\epsilon & \text{w.p. } 1-\delta \\ 1 & \text{w.p. } \delta \end{cases}$$

where $\epsilon := \left(\frac{\rho}{2m}\right)^{\frac{1}{k}} \delta^{\frac{1}{k_*}}$ for some

$$0 < \delta \leq \frac{1}{2C_{f,\rho,m}} \wedge \frac{\rho}{2m} \left(\left(\frac{2}{3}\right)^k \wedge \frac{1}{2} \left(\frac{\rho}{2m}\right)^{-k_*} \right)$$

to be choosen later. Now, we again show that $\mathrm{d}_{\mathrm{opt}}(P_1, P_2; f) \geq \frac{\epsilon}{2}$. To this end, first observe that $\mathcal{R}_f(\theta; P_1) = -\epsilon\theta$. When $\theta < 0$, we have $\mathcal{R}_f(\theta; P_2) \geq \theta \mathbb{E}[X_2] \geq -\theta(\epsilon(1-\delta) - \delta) \geq 0$ as $\epsilon \leq \frac{1}{2}$ in the given range of δ. When $\theta \geq 0$, recall that Q such that $Q(Z = M) = \left(\frac{\rho}{2m}\right)^{\frac{1}{k}} \delta^{\frac{1}{k_*}}$ and $Q(Z = 0) = 1 - Q(Z = M)$, satisfies $D_f\left(Q \| P_2\right) \leq \rho$ by Lemma 10. Hence, we have for $\theta \geq 0$

$$\mathcal{R}_f(\theta; P_2) \geq \epsilon\theta.$$

Denoting $\theta_i^\star := \operatorname{argmin}_{\theta \in \Theta} \mathcal{R}_f(\theta; P_i)$ again, we consequently obtain $\theta_1^\star = M$, $\theta_2^\star = 0$ with $\mathcal{R}_f(\theta_1^\star; P_1) = -M\epsilon$, $\mathcal{R}_f(\theta_1^\star; P_2) = 0$.

Using an identical argument as in the second part of Section C.4, we can show $\mathrm{d}_{\mathrm{opt}}(P_1, P_2; f) \geq \frac{M\epsilon}{2}$. Setting

$$\delta = \frac{1}{2(n \vee C_{f,\rho,m})} \wedge \frac{\rho}{2m} \left(\left(\frac{2}{3}\right)^k \wedge \frac{1}{2} \left(\frac{\rho}{2m}\right)^{-k_*} \right)$$

and using the same argument as in Section C.4, we obtain the result.

D Proofs of Consistency

We begin this section with a brief review of the theory of epi-convergence [49, 64], which governs convergence of solutions to optimization problems, so we consequently use its tools to develop our consistency results.

We begin with some necessary set-valued analysis.

Definition 1. *Let $\{A_n\}$ be a sequence of subsets of $\mathbb{R}^d$. The* limit supremum *(or* limit exterior *or* outer limit*) and* limit infimum *(*limit interior *or* inner limit*) of the sequence $\{A_n\}$ are*

$$\limsup_n A_n := \left\{ v \in \mathbb{R}^d \mid \liminf_{n\to\infty} \operatorname{dist}(v, A_n) = 0 \right\} \quad \text{and}$$
$$\liminf_n A_n := \left\{ v \in \mathbb{R}^d \mid \limsup_{n\to\infty} \operatorname{dist}(v, A_n) = 0 \right\}.$$

Recall that the epigraph of a function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is

$$\operatorname{epi} h := \{(x,t) \in \mathbb{R}^d \times \mathbb{R} \mid h(x) \le t\}.$$

Based on Definition 1 of limits of sets, we say that $\lim_n A = A_\infty$ if $\limsup_n A_n = \liminf_n A_n = A_\infty \subset \mathbb{R}^d$, and we have the following notion of convergence of functions in terms of their epigraphs.

Definition 2. *A sequence of functions h_n* epi-converges *to a function h, denoted $h_n \xrightarrow{\text{epi}} h$, if*

$$\operatorname{epi} h = \liminf_{n\to\infty} \operatorname{epi} h_n = \limsup_{n\to\infty} \operatorname{epi} h_n. \tag{33}$$

If $\operatorname{dom} h \neq \varnothing$, meaning that h is proper, epigraphical convergence (33) for closed convex functions has the following equivalent characterizations.

Lemma 13 (Theorem 7.17, Rockafellar and Wets [64])**.** *Let $h_n : \mathbb{R}^d \to \overline{\mathbb{R}}, h : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed convex and proper. Then $h_n \xrightarrow{\text{epi}} h$ is equivalent to either of the following two conditions.*

(i) There exists a dense set $A \subset \mathbb{R}^d$ such that $h_n(v) \to h(v)$ for all $v \in A$.

(ii) For all compact $C \subset \operatorname{dom} h$ not containing a boundary point of $\operatorname{dom} h$*,*

$$\lim_{n\to\infty} \sup_{v\in C} |h_n(v) - h(v)| = 0.$$

Importantly for our development, epigraphical convergence implies the infimal value convergence, and under additional conditions, convergence of solution sets.

Lemma 14 (Theorem 7.31, Rockafellar and Wets [64])**.** *Let $h_n : \mathbb{R}^d \to \overline{\mathbb{R}}, h : \mathbb{R}^d \to \overline{\mathbb{R}}$ satisfy $h_n \xrightarrow{\text{epi}} h$ and $-\infty < \inf h < \infty$. Let $S_n(\varepsilon) = \{\theta \mid h_n(\theta) \le \inf h_n + \varepsilon\}$ and $S(\varepsilon) = \{\theta \mid h(\theta) \le \inf h + \varepsilon\}$. Then $\limsup_n S_n(\varepsilon) \subset S(\varepsilon)$ for all $\varepsilon \ge 0$, and $\limsup_n S_n(\varepsilon_n) \subset S(0)$ whenever $\varepsilon_n \downarrow 0$.*

Lemma 15 (Proposition 7.33, Rockafellar and Wets [64])**.** *Let $h_n : \mathbb{R}^d \to \overline{\mathbb{R}}, h : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed and proper. If h_n has bounded sublevel sets and $h_n \xrightarrow{\text{epi}} h$, then $\inf_v h_n(v) \to \inf_v h(v)$.*

D.1 Proof of Proposition 9

To ease notation, we fix $\theta \in \Theta$ and denote $Z(x) := \ell(\theta; x)$, and we typically omit the dependence of $\mathcal{R}$ on θ (as it is fixed), writing $\mathcal{R}_f(P)$ and $\mathcal{R}_k(P)$. The proof builds out of the epi-convergence theory we outline in the beginning of Section D.

By Proposition 1, strong duality (5) holds for both $P = P_0$ and $P = \widehat{P}_n$. For a probability measure P, define the dual objective

$$g_{f,P}(\lambda, \eta) := \begin{cases} \mathbb{E}_P\left[\lambda f^*\left(\frac{Z-\eta}{\lambda}\right)\right] + \rho\lambda + \eta & \text{if } \lambda \ge 0 \\ \infty & \text{otherwise,} \end{cases}$$

where we recall our convention (4) with the perspective. Using that $f^*(s) \geq 0$ for $s \geq 0$ and our assumption that $\mathbb{E}[f^*(|Z|)] < \infty$, the strong law of large numbers implies that

$$\mathcal{E} := \left\{ \lim_{n\to\infty} g_{f,\widehat{P}_n}(\lambda, \eta) = g_{f,P_0}(\lambda, \eta) \;\; \text{for all} \;\; \lambda \in \mathbb{Q}, \eta \in \mathbb{Q} \right\}$$

has P_0-measure 1. We now show that the functions g_f are both closed. To that end, note that standard conjugacy calculations [41, Prop. I.6.1.2] imply $1 \in \partial f^*(0) = \operatorname{argmax}_t\{-f(t)\}$, as $f(1) = 0$, $t = 1$ minimizes f, and $f^*(0) = 0$. Thus we have $f^*(s) \geq f^*(0) + s$ for all s, so that

$$\lambda f^*\left(\frac{z-\eta}{\lambda}\right) - (z - \eta) \geq 0.$$

Fatou's lemma then implies that for $v = (\eta, \lambda)$ and $v_0 = (\eta_0, \lambda_0)$ we have

$$\begin{aligned}
&\liminf_{v\to v_0} \left\{ \mathbb{E}_P\left[\lambda f^*\left(\frac{Z-\eta}{\lambda}\right) - (Z-\eta)\right] + \rho\lambda + \eta \right\} \\
&\geq \mathbb{E}_P\left[\liminf_{v\to v_0} \left\{ \lambda f^*\left(\frac{Z-\eta}{\lambda}\right) - (Z-\eta)\right\}\right] + \rho\lambda_0 + \eta_0 \\
&\geq \mathbb{E}_P\left[\lambda_0 f^*\left(\frac{Z-\eta_0}{\lambda_0}\right) - (Z-\eta_0)\right] + \rho\lambda_0 + \eta_0,
\end{aligned}$$

where the last inequality follows by the lower semicontinuity of the perspective (4). Using Lebesgue's dominated convergence theorem on $(Z - \eta)$, using the dominating function $|Z| + |\eta|$, we have thus shown that both $g_{f,\widehat{P}_n}$ and g_{f,P_0} are lower semicontinuous. Lemma 13 implies that $g_{f,\widehat{P}_n} \stackrel{\text{epi}}{\to} g_{f,P_0}$ with probability 1.

Finally, we would like to apply Lemma 15; to do so, we must show that $g_{f,\widehat{P}_n}$ is (eventually) coercive. For this, we note that $\lambda f^*(\frac{Z-\eta}{\lambda}) - Z + \eta \geq 0$ as above, so that $g_{f,P}(\eta, \lambda) \geq \rho\lambda + \mathbb{E}_P[Z]$, and thus for any P for which $\mathbb{E}_P[Z]$ exists, $\lim_{\lambda\to\infty} \inf_\eta g_{f,P}(\eta, \lambda) = \infty$. To show coercivity of $g_{f,P}$ as $\|(\eta, \lambda)\| \to \infty$, we thus need only consider limits taken as λ remains bounded. Now, we claim that under the conditions of the lemma,

$$\limsup_{s\to-\infty} \frac{f^*(s)}{s} = \epsilon < 1 \;\; \text{and} \;\; \liminf_{s\to\infty} \frac{f^*(s)}{s} = \infty. \tag{34}$$

Deferring the proof of the claims (34), let us show how they imply that g_{f,P_0} is coercive. Assume that $0 \leq \lambda \leq \Lambda < \infty$. For any constant $K < \infty$, $K > \Lambda$, there exist $b, c < \infty$ such that $|z| \leq b$ and $\eta < -c$ imply that $f^*(\frac{z-\eta}{\lambda}) \geq K|\eta|/\Lambda$, and similarly, $\eta > c$ implies $\lambda f^*(\frac{z-\eta}{\lambda}) \geq -\frac{1+\epsilon}{2}\eta$. For $\eta < -c$, then, we have

$$g_{f,P}(\eta, \lambda) \geq P(|Z| \leq b)\left[\frac{K|\eta|}{\Lambda} + \rho\lambda + \eta\right] + P(|Z| > b)\rho\lambda + \mathbb{E}_P[\mathbf{1}\left\{|Z| > b\right\} Z],$$

and for $\eta > c$ we similarly have

$$g_{f,P}(\eta, \lambda) \geq P(|Z| \leq b)\left[\rho\lambda + \frac{\epsilon\eta}{2}\right] + P(|Z| > b)\rho\lambda + \mathbb{E}_P[\mathbf{1}\left\{|Z| > b\right\} Z].$$

Whenever $\mathbb{E}_P[|Z|] < \infty$, we see that $\lim_{|\eta|\to\infty} \inf_{\lambda\in[0,\Lambda]} g_{f,P}(\eta, \lambda) = \infty$, so that $g_{f,P}$ is coercive. Consequently, the claim (34), coupled with our assumption that $\mathbb{E}_{P_0}[|Z|] < \infty$, implies that g_{f,P_0} is coercive. Because $g_{f,\widehat{P}_n} \stackrel{\text{epi}}{\to} g_{f,P_0}$, we have uniform convergence of $g_{f,\widehat{P}_n}$ to g_{f,P_0} on compacta (Lemma 13), and thus $g_{f,\widehat{P}_n}$ is eventually coercive. Lemma 15 thus implies the result.

Finally, we return to the claim (34). For the first claim, we have for $s < 0$ that

$$\frac{1}{s} \sup_{t \geq 0} \{st - f(t)\} = \inf_{t \geq 0} \left\{ t + \frac{f(t)}{|s|} \right\},$$

which is decreasing as $s \downarrow -\infty$, and letting $t_0 < 1$ be any value for which $f(t_0) < \infty$ (as f is finite near $t = 1$), we have $\limsup_{s \to -\infty} \frac{1}{s} f^*(s) \leq t_0 < 1$ as desired. For the second claim of inequalities (34), use that $f(t) < \infty$ for all $t \geq 1$; for each $n \in \mathbb{N}$, then, there exists $s < \infty$ such that $f(n)/s \leq 2$, so that $\frac{1}{s} f^*(s) = \sup_{t \geq 0} \{t - f(t)/s\} \geq n - 2$. Taking $n \to \infty$ gives the claim.

D.2 Proof of Proposition 10

The epi-convergence theory of the beginning of Section D, combined with Proposition 9, gives most of the results. First, we know that $\mathcal{R}_f(\theta; \widehat{P}_n)$ and $\mathcal{R}_f(\theta; P_0)$ are lower semicontinuous in θ, as each is the supremum of closed convex functions $\theta \mapsto \int \ell(\theta; x) dP(x)$. Combined with Proposition 9, we have that $\mathcal{R}_f(\cdot; \widehat{P}_n) \stackrel{\text{epi}}{\to} \mathcal{R}_f(\cdot; P_0)$ with P_0-probability 1. Using the coercivity of $\mathcal{R}_f(\cdot; P_0)$ and that $\mathcal{R}_f(\theta; P_0) < \infty$ on an open set containing $S_{P_0}(\Theta, 0)$, we take any compact set $C \subset \mathbb{R}^d$ containing $S_{P_0}(\Theta, 0)$ with $\mathcal{R}_f(\theta; P_0) < \infty$ on C, and we obtain $\sup_{\theta \in C} |\mathcal{R}_f(\theta; P_0) - \mathcal{R}_f(\theta; \widehat{P}_n)| \stackrel{a.s.}{\to} 0$ by Lemma 13. The convexity of $\mathcal{R}_f(\cdot; \widehat{P}_n)$ then implies that $\mathcal{R}_f(\cdot; \widehat{P}_n)$ is coercive eventually, so that it has bounded sublevel sets, and Lemma 15 implies that $\inf_{\theta \in \Theta} \mathcal{R}_f(\theta; \widehat{P}_n) \stackrel{a.s.}{\to} \inf_{\theta \in \Theta} \mathcal{R}_f(\theta; P_0)$.

For the second result, we use that for any sequence $\varepsilon_n \geq 0$, eventually the set $S_{\widehat{P}_n}(\Theta, \varepsilon_n)$ is non-empty by coercivity, and then Lemma 14 implies that

$$\limsup_n S_{\widehat{P}_n}(\Theta, \varepsilon_n) \subset S_{P_0}(\Theta, 0).$$

In turn, this yields that $\lim_n d_{\subset}(S_{\widehat{P}_n}(\Theta, \varepsilon_n)) = 0$ as $S_{P_0}(\Theta, 0)$ is compact by the coercivity assumption.

E Proof of Limit Theorems

E.1 Proof of Lemma 2

To ease notation, let $Z = \ell(\theta_0; X)$, and recall from Lemma 1 (and its proof in Section A.2) that we may rewrite the dual as

$$g_P(\theta, \lambda, \eta) = \frac{1}{\lambda^{k_* - 1}} \frac{(k-1)^{k_*}}{k} \mathbb{E}_P \left[(Z - \eta)_+^{k_*} \right] + \left(\rho + \frac{1}{k(k-1)} \right) \lambda + \eta.$$

In this case, it is clear that the minimizing λ is unique as in Eq. (21), with

$$g(\eta) := \inf_{\lambda \geq 0} g_P(\theta, \lambda, \eta) = c_k(\rho) \mathbb{E}_P \left[(Z - \eta)_+^{k_*} \right]^{1/k_*} + \eta,$$

where $c_k(\rho) = (k(k-1)\rho + 1)^{1/k} > 1$. It is evident that g is convex and coercive in η. Now, for all $\eta \geq \operatorname{ess\,sup} Z$ we have $g(\eta) = \eta$, so that g is strictly increasing in $\eta \geq \operatorname{ess\,sup} Z$. On the set $(-\infty, \operatorname{ess\,sup} Z)$, we claim that g is strictly convex. Indeed, for $\eta_1 \neq \eta_2 \in (-\infty, \operatorname{ess\,sup} Z)$ and

$\alpha \in (0,1)$, we have

$$g(\alpha\eta_1 + (1-\alpha)\eta_2) \le \left\|\alpha\left(Z-\eta_1\right)_+ + (1-\alpha)\left(Z-\eta_2\right)_+\right\|_{k_*,P} + \alpha\eta_1 + (1-\alpha)\eta_2$$
$$\overset{(*)}{<} \alpha\left\|\left(Z-\eta_1\right)_+\right\|_{k_*,P} + (1-\alpha)\left\|\left(Z-\eta_2\right)_+\right\|_{k_*,P} + \alpha\eta_1 + (1-\alpha)\eta_2$$

where step $(*)$ follows as equality in the triangle inequality $\|Y_1 + Y_2\| \le \|Y_1\| + \|Y_2\|$ if and only if there exists $c \in \mathbb{R}$ such that $Y_1 = cY_2$ with probability one.

E.2 Proof of Theorem 11

We use a powerful result on asymptotic normality that we show applies in our setting. To state the result, we require a bit of (temporary) notation. First, recall the definition of bracketing numbers for a collection of functions.

Definition 3. *Let $\|\cdot\|$ be a (semi-)norm on $\mathcal{H}$. For functions $l, u : \mathcal{X} \to \mathbb{R}$ with $l \le u$, the* bracket *$[l,u]$ is the set of functions $h : \mathcal{X} \to \mathbb{R}$ such that $l \le h \le u$, and $[l,u]$ is an ϵ-*bracket *if $\|l - u\| \le \epsilon$. Brackets $\{[l_i, u_i]\}_{i=1}^m$* cover *$\mathcal{H}$ if for all $h \in \mathcal{H}$, there is some bracket i such that $h \in [l_i, u_i]$. The* bracketing number *$N_{[]}(\epsilon, \mathcal{H}, \|\cdot\|)$ is the minimum number of ϵ-brackets needed to cover $\mathcal{H}$.*

Now, let $\mathcal{V} \subset \mathbb{R}^d$ be a convex set and $H : \mathcal{V} \times \mathcal{X} \to \mathbb{R}$ be a collection of criterion functions, where $\widehat{v}_n = \operatorname{argmin}_{v\in\mathcal{V}} \mathbb{E}_{\widehat{P}_n}[H(v;X)]$. Assume that $v^\star = \operatorname{argmin}_{v\in\mathcal{V}} \mathbb{E}_{P_0}[H(v;X)]$ exists and is unique, and for $\epsilon > 0$, define the localized function classes

$$\mathcal{H}_\epsilon := \left\{x \mapsto H(v;x) - H(v^\star;x) : \|v - v^\star\| \le \epsilon\right\}.$$

We say that $M_\epsilon : \mathcal{X} \to \mathbb{R}_+$ is an envelope for $\mathcal{H}_\epsilon$ if $h \in \mathcal{H}_\epsilon$ implies $|h(x)| \le M_\epsilon(x)$; without further mention we take $M_\epsilon(x) := \sup_{\|v-v^\star\|\le\epsilon} |H(v;x) - H(v^\star;x)|$. With these definitions, we have the following result.

Lemma 16 ([83, Theorem 3.2.10])**.** *Let the conditions above hold, and assume that $\mathcal{H}_\epsilon$ has envelope M_ϵ with $\mathbb{E}[M_\epsilon^2] < \infty$. Assume additionally that*

(i) The function $v \mapsto R(v) := \mathbb{E}[H(v;X)]$ is $\mathcal{C}^2$ near $v^\star$ and $\nabla^2 R(v^\star) \succ 0$.

(ii) The bracketing integral of $\mathcal{H}_\epsilon$ is uniformly bounded as $\epsilon \to 0$: for some $\epsilon_0 > 0$,

$$\int_0^\infty \sup_{\epsilon<\epsilon_0} \sqrt{\log N_{[]}\left(\delta \left\|M_\epsilon\right\|_{P_0,2}, \mathcal{H}_\epsilon, L_2(P_0)\right)}\, d\delta < \infty. \tag{35}$$

(iii) There exists $C < \infty$ such that $\mathbb{E}[M_\epsilon(X)^2] \le C\epsilon^2$ for all small ϵ.

(iv) There exists a centered Gaussian process G on $\mathbb{R}^d$ where $G(v) = G(v')$ P_0-almost surely only if $v = v'$ such that for every $c, K > 0$,

$$\lim_{\epsilon\to 0} \epsilon^{-2}\mathbb{E}[M_\epsilon(X)^2 \mathbf{1}\left\{M_\epsilon(X) > c\right\}] = 0, \tag{36a}$$

$$\lim_{\epsilon\to 0} \limsup_{\delta\to 0} \sup_{\|u_1-u_2\|<\epsilon, \|u_1\|\vee\|u_2\|\le K} \delta^{-2}\mathbb{E}[(H(v^\star+\delta u_1;X) - H(v^\star+\delta u_2;X))^2] = 0 \tag{36b}$$

$$\lim_{\delta\to 0} \delta^{-2}\mathbb{E}[(H(v^\star+\delta u_1;X) - H(v^\star+\delta u_2;X))^2] = \mathbb{E}[(G(u_1) - G(u_2))^2]. \tag{36c}$$

Then, there exists a version of G with bounded, uniformly continuous sample paths on compacta. Further, if $\widehat{v}_n \in \mathcal{V}$ satisfies $\mathbb{E}_{\widehat{P}_n}[H(\widehat{v}_n; X)] \le \inf_{v \in \mathcal{V}} \mathbb{E}_{\widehat{P}_n}[H(v; X)] + O_P(1/n)$ and $\widehat{v}_n \xrightarrow{a.s.} v^\star$, then $\sqrt{n}(\widehat{v}_n - v^\star)$ converges in distribution to the unique maximizer of the process

$$u \mapsto G(u) + \frac{1}{2} u^T \nabla^2 R(v^\star) u.$$

We now show how under the conditions specified in Theorem 11, our problem satisfies the conditions of Lemma 16. We first provide notation and a few additional definitions for shorthand. Define

$$H(\theta, \lambda, \eta; X) := \lambda f^* \left(\frac{\ell(\theta; X) - \eta}{\lambda} \right) + \rho\lambda + \eta,$$

so that $g_P(\theta, \lambda, \eta) = \mathbb{E}_P[H(\theta, \lambda, \eta; X)]$. Let $(\widehat{\theta}_n, \widehat{\lambda}_n, \widehat{\eta}_n)$ be the empirical minimizer

$$(\widehat{\theta}_n, \widehat{\lambda}_n, \widehat{\eta}_n) \in \operatorname*{argmin}_{\theta, \lambda \ge 0, \eta} \mathbb{E}_{\widehat{P}_n}[H(\theta, \lambda, \eta; X)].$$

For $\epsilon > 0$, define the collection

$$\mathcal{H}_\epsilon := \left\{ x \mapsto H(\theta, \lambda, \eta; x) - H(\theta^\star, \lambda^\star, \eta^\star; x) : \|\theta - \theta^\star\| + |\lambda - \lambda^\star| + |\eta - \eta^\star| \le \epsilon \right\}. \tag{37}$$

We claim that the envelope M_ϵ exists for the set (37). First, we note that ∇H exists with probability 1: by our Assumption C that g_{P_0} is $\mathcal{C}^2$ near $(\theta^\star, \lambda^\star, \eta^\star)$, we know that g_{P_0} is continuously differentiable. Then For $h(t, x)$ an arbitrary function, convex in t, $\int h(t, x) dP(x)$ is differentiable at some t_0 if and only if $t \mapsto h(t, x)$ is differentiable at t_0 for P-almost all x [12]. Consequently, for P_0-almost all x we have $\nabla H(\cdot; x)$ exists in a neighborhood of $(\theta^\star, \lambda^\star, \eta^\star)$, and

$$\nabla H(\theta, \lambda, \eta; x) = \begin{bmatrix} f^{*\prime}\left(\frac{\ell(\theta;x)-\eta}{\lambda}\right) \nabla \ell(\theta; x) \\ -f^{*\prime}\left(\frac{\ell(\theta;x)-\eta}{\lambda}\right) + 1 \\ f^*\left(\frac{\ell(\theta;x)-\eta}{\lambda}\right) - \frac{1}{\lambda} f^{*\prime}\left(\frac{\ell(\theta;x)-\eta}{\lambda}\right)(\ell(\theta; x) - \eta) + \rho \end{bmatrix} \tag{38}$$

for (θ, λ, η) near $(\theta^\star, \lambda^\star, \eta^\star)$. We begin with a simple technical lemma.

Lemma 17. *Let f satisfy the conditions of Theorem 11 and $k_* = \frac{k}{k-1}$. Then $\limsup_{s \to \infty} f^*(s)/s^{k_*} < \infty$, and for any $t(s) \in \partial f^*(s)$, $t(s) \ge 0$ and $\limsup_{s \to \infty} t(s)/s^{\frac{1}{k-1}} < \infty$.*

Proof We begin with the first claim, recalling the assumption that $\liminf_{t \to \infty} f(t)/t^k > 0$, so that for some $t_0 < \infty$ there exists $c > 0$ such that $f(t) \ge ct^k$ for all $t \ge t_0$. Thus for $s \ge 0$, we have

$$f^*(s) = \sup_{t \ge 0} \{st - f(t)\} \le \sup_{t \in [0, t_0]} \{st - f(t)\} \vee \sup_{t \ge t_0} \{st - f(t)\} \le st_0 \vee \sup_{t \ge t_0} \{st - ct^k\} \le st_0 \vee Cs^{k_*}.$$

Now we show the second claim. To see this, recall the standard conjugacy result [41] that $t(s) \in \operatorname{argmax}\{st - f(t)\}$, so that $t(s) \ge 0$ always, and let $\hat{t} = (s/kc)^{\frac{1}{k-1}}$. Assume that s is large enough that $f(t) \ge ct^k$ for $t > \hat{t}$. Then for $t > \hat{t}$, we have

$$st - f(t) \le st - ct^k < s\hat{t} - c\hat{t}^k,$$

as $\hat{t}$ uniquely maximizes $st - ct^k$. Thus t cannot belong to $\partial f^*(s)$, giving the result. □

With Lemma 17 in hand, the next lemma follows.

Lemma 18. *There exists a constant $C < \infty$ and a neighborhood U of $(\theta^\star, \lambda^\star, \eta^\star)$ such that $M(x) := \sup_{(\theta,\lambda,\eta)\in U} \|\nabla H(\theta, \lambda, \eta; x)\|$ satisfies*

$$M(x) \le C \left[\frac{|\ell(\theta^\star; x)|^{k_*} + |\eta^\star|^{k_*}}{\lambda^{\star k_*}} + L(x)^{k_*} \right],$$

and $M_\epsilon(x) := M(x) \cdot \epsilon$ is an envelope for $\mathcal{H}_\epsilon$.

Proof The result is a standard algebraic exercise, coupled with the fact that a convex function h is Lipschitz in an ϵ-neighborhood of a point t_0 with constant $\sup_t \{\|\partial h(t)\|_2 \mid \|t - t_0\| \}$ (cf. [41]). Thus, we bound the components of ∇H from Eq. (38); we only bound $\nabla_\theta H$ as the others are completely similar. For (θ, λ, η) in a neighborhood U of $(\theta^\star, \lambda^\star, \eta^\star)$, we have for constants $C < \infty$ that may change from line to line

$$\begin{aligned}
\|\nabla_\theta H(\theta, \lambda, \eta; x)\| &= f^{*\prime}\left(\frac{\ell(\theta; x) - \eta}{\lambda}\right) \|\nabla \ell(\theta; x)\| \\
&\stackrel{(i)}{\le} C \left|\frac{\ell(\theta; x) - \eta}{\lambda}\right|^{\frac{1}{k-1}} \|\nabla \ell(\theta; x)\| \\
&\stackrel{(ii)}{\le} C \left|\frac{\ell(\theta; x) - \eta}{\lambda}\right|^{\frac{k}{k-1}} + C \|\nabla \ell(\theta; x)\|^{k_*} \\
&\stackrel{(iii)}{\le} C\frac{|\eta|^{k_*}}{\lambda^{k_*}} + C\frac{|\ell(\theta^\star; x)|^{k_*}}{\lambda^{k_*}} + CL(x)^{k_*},
\end{aligned}$$

where inequality (i) follows from Lemma 17, (ii) follows by the Fenchel-Young inequality that $ab \le (1/k)|a|^k + (1/k_*)|b|^{k_*}$, while inequality (iii) is a consequence of Assumption B.1. The remainder of the derivation follows from straightforward algebra once we note that $\lambda/\lambda^\star$ is bounded for λ near $\lambda^\star$. □

Finally, we show that each of the conditions of Lemma 16 holds for our problem. That $\mathbb{E}[M_\epsilon(X)^2] < \infty$ is immediate by Assumption B on the moments of ℓ and $\nabla\ell$. For condition (i), we have Assumption C. For the bracketing integral condition (35), From a standard bound on bracketing numbers for Lipschitz functions [83, Theorem 2.7.11], we have

$$\log N_{[]}\left(\delta \|M_\epsilon\|, \mathcal{H}_\epsilon, L_2(P_0)\right) \le (d+2) \log\left(1 + \frac{2}{\delta}\right)$$

for ϵ small enough, so that the bracketing integral is bounded. Each of the quantities (36) follows by Lebesgue's dominated convergence theorem. For condition (36a), we have $M_\epsilon(x)^2 \mathbf{1}\{M_\epsilon(x) > c\}/\epsilon^2 = M(x)^2 \mathbf{1}\{M(x) > c/\epsilon\} \to 0$ as $\epsilon \to 0$, and it is dominated by $M(x)$. For condition (36b), we have for $v^\star = (\theta^\star, \lambda^\star, \eta^\star)$ that

$$|H(v^\star + \delta u_1; x) - H(v^\star + \delta u_2; x)| \le \sup_{v \text{ near } v^\star} \|\nabla H(v; x)\| \, \delta \|u_1 - u_2\| \le M(x)\delta \|u_1 - u_2\|$$

by Lemma 18. Thus the dominated convergence theorem again implies the convergence (36b). For the covariance condition (36c), we use the differentiability of H as in Eq. (38) to see that with $v^\star$ as above, $\frac{1}{\delta}(H(v^\star + \delta u_1; x) - H(v^\star + \delta u_2; x)) \to \langle \nabla H(v^\star; x), u_1 - u_2 \rangle$ and it is dominated by $M(x)\|u_1 - u_2\|$. Thus, we may take

$$G(u) := \langle W, u \rangle \quad \text{for} \quad W \sim \mathsf{N}\left(0, \mathrm{Cov}(\nabla H(\theta^\star, \lambda^\star, \eta^\star; X))\right)$$

as our Gaussian process. The theorem is then an immediate consequence of Lemma 16.