Learning Models with Uniform Performance
via Distributionally Robust Optimization

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

Departments of Statistics and Electrical Engineering, Stanford University[1]

Decision, Risk, and Operations Division, Columbia Business School[2]

jduchi@stanford.edu, hn2369@columbia.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 providing good performance against perturbations to the data-generating distribution. We give a convex formulation for the problem, providing several convergence guarantees. We prove finite-sample minimax upper and lower bounds, showing that distributional 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 predictive power on certain regions of covariates [70], and high average performance comes at the expense of low performance on minority subpopulations. In this work, we study a procedure that explicitly optimizes performance on tail inputs that suffer high loss.

Modern datasets incorporate heterogeneous (but often latent) subpopulations, and a natural goal is to perform well across all of these [70, 80, 25]. 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 [6]. 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 [48, 52, 22, 83, 96].

In addition to latent heterogeneity in the population, distributional shifts in covariates [89, 12] or unobserved confounding variables (e.g. unmodeled temporal effects [49]) 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 [49, 21, 31, 82, 97] and even when new test data are constructed following identical data construction procedures [76]. Domain

adaptation [89, 12, 13] and multi-task learning methods [28] 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* problem

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

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

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

is the f-divergence [5, 30] 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 $\ell(\theta; X)$, and thus formulation (1) optimizes performance on the tails, as measured by the loss on "hard" examples. In our motivating scenarios of distribution shift or latent subpopulations, as long as the alternative distribution Q remains ρ-close to the data-generating distribution P_0, the model $\theta^\star \in \Theta$ that minimizes the worst-case formulation (1) evidently guarantees that $\mathbb{E}_Q[\ell(\theta^\star; X)] \le \mathcal{R}_f(\theta^\star; P_0)$ and provides the smallest such bound; as we show shortly, this is equivalent to controlling the tail-performance under P_0. In our subsequent discussion, we refer to this behavior as *uniform performance*. 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

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

To build intuition for the worst-case formulation (1), we begin our discussion (in Section 2) by showing that protection against distributional shifts is equivalent to controlling the tail-performance of a model. The modeler's choice of f determines the tail performance she wants to control, and this dual interpretation provides intuition for the appropriate choice of f and ρ. To concretely understand the types of distributional shifts the worst-case formulation (1) protects against, we provide (in Section 2.1) explicit calculations suggesting appropriate choices of f in some situations. Given nontrivial modeling freedom in choosing f and ρ, we begin our study in Section 3 with experiments that substantiate our intuitive explanations. Our experimental and theoretical work demonstrates that the distributionally robust estimator $\widehat{\theta}_n$ trades performance on the tails of the data-generating distribution with average-case performance—which empirical risk minimization optimizes. Empirically, we observe in a number of scenarios that such gains in tail-performance (e.g. hard inputs) come at moderate degradation to the average-case performance, so that the robust estimator (2) achieves fairly low loss uniformly across the input space $\mathcal{X}$. For non worst-case distribution shifts, the worst-case formulation (1) *prima-facie* does not guarantee better performance

2

than empirical risk minimization; the duality between it and tail losses to come suggests that for light-tailed data, distributional robustness comes at little cost to typical-case performance. While work in finance and operations research [17] highlights the benefits of robustness, it is important to investigate the typical shifts one might expect in statistical learning scenarios. To this end, we see in our experiments that the robust estimator (2) sacrifices some average-case performance (which empirical risk minimization optimizes) for lower losses on difficult subpopulations, covariate shift, and other latent confounding.

Although we view a general theoretical characterization of the "right" choice of f and ρ as an important open question, we provide two heuristics for this choice and evaluate their performance on simulation experiments in Section 3. First, as a general approach, we advocate splitting training data non-exchangeably into multiple validation sets, then using these to validate choices f and ρ; we will expand on this later in the paper with concrete examples and experiments. As brief examples, we may group data by its loss or, in supervised learning scenarios with outcome/label Y, by values of Y; when an auxiliary dataset on worse-than-average subpopulations is available, we could use this. The intuition is to use variability within the available data as a proxy for potential departures from the data-generating distribution.

Motivated by our empirical findings in Section 3, 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 which also forms the basis of our above tail-performance interpretation of distributional robustness (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 minimax convergence 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

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

where $k_* = \frac{k}{k-1}$ (Section 4). We provide minimax lower bounds matching these rates in n up to log factors. These results quantify fundamental *statistical costs* for protecting against large distributional shifts (the worst-case region $\{Q : D_f(Q\|P_0) \leq \rho\}$ becomes larger as $k \to 1$, or $k_* \to \infty$).

Since these minimax guarantees do not necessarily reflect the typical behavior of the estimators, 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 arise in many guises across statistics, machine learning, applied probability, simulation, and optimization; we give a necessarily abridged survey of the many strains of work and their respective 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} [89, 53, 18, 92, 93, 98]. 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 [49], and time-varying-coefficient models are effective when time indices are available [40, 26]. When one believes there may be latent subpopulations, mixture model approaches can model latent membership directly [4, 41, 69, 27]. 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, Bühlmann, Meinshausen, and colleagues connect methods that achieve good performance on all subpopulations with causal interventions. In this vein, they study maximin effects on heterogeneous datasets and learn linear models that maximize relative performance over the worst (observed) subgroup [70], which connects to minimax regret in linear models [38, 14, 80, 25, 81]. Without access to information about particular subpopulations, the worst-case formulation (1) is more conservative than their approaches, but can still achieve good performance, as we see in our experimental evaluation.

The idea to build predictors robust to perturbation of an underlying data-generating distribution has a long history across multiple fields. In dynamical systems and control, Petersen et al. [74] build worst-case optimal controllers for systems whose uncertain dynamics are described by Kullback-Leibler (KL) divergence balls. In econometrics, Hansen and Sargent [50] study systems in which rational agents dynamically make decisions assuming worst-case (dynamics) model misspecification, where the misspecification is bounded by an evolving KL-divergence quantity. There is also substantial work in characterizing worst-case sensitivity of risk measures to distributional misspecification [46, 10, 62, 37, 63, 45]. A common goal in such sensitivity calculations is an asymptotic expansion of a risk measure as the radius ρ of the region of misspecification decreases to 0. In contrast, we study statistical properties of the worst-case formulation (1) given observations drawn from the data generating distribution P_0, so that we must both address statistical uncertainty and challenges of robustness.

In the optimization literature, a body of work studies distributionally robust optimization problems. Several authors investigate worst-case regions arising out of moment conditions on the data vector X [33, 57, 17]. Other work [15, 17, 36, 72, 62, 64] 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. [36] and Lam and Zhou [64] show how such approximations correspond to generalized empirical likelihood [73] confidence bounds on $\mathbb{E}_{P_0}[\ell(\theta; X)]$. These procedures are identical to the plug-in (2) except that the radius decreases as ρ/n. Thus, the magnitude of this radius depends on whether the modeler's goal is good performance with respect to $\mathbb{E}_{P_0}[\ell(\theta; X)]$ (radius shrinks as ρ/n), or—as is the case here—robustness under distributional shifts (radius ρ is fixed).

An alternative to our f-divergence based sets $\{Q : D_f(Q\|P_0) \leq \rho\}$ are Wasserstein balls [105, 75, 106, 84, 19, 20, 42, 39, 91, 68]. Such approaches are satisfying, as Wasserstein balls 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 [84, 39, 91], and they remain computationally challenging. Furthermore, most guarantees [20, 39, 84] for these problems also consider approximation only of the canonical (population) loss $\mathbb{E}_{P_0}[\ell(\theta; X)]$ using

4

shrinking radius $\rho_n \to 0$. In comparison, our f-divergence formulation is computationally efficient to solve, even in large-scale learning scenarios [71, 72].

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, and $Z_n \overset{p}{\to} Z$ for convergence in probability. We let $\ell^\infty(\mathcal{Z})$ the space of bounded real-valued functions on $\mathcal{Z}$ equipped with the supremum norm. We let $D_{\chi^2}(P\|Q) = \frac{1}{2} \int (dP/dQ - 1)^2 \, dQ$ be the χ^2-divergence. For $Z \sim P$, $\operatorname{ess\,sup}_P Z$ is its essential supremum. We make the dependence on the underlying measure explicit when we write expectations (e.g. $\mathbb{E}_P[X]$), except for when $P = P_0$. For $k \in (1, \infty)$, we let $k_* := k/(k-1)$. By $\nabla \ell(\theta; X)$ we 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 $\mathcal{R}_f(\theta; P_0)$, deferring formulation in terms of worst subpopulations to Example 3 to come. The dual form gives a single convex minimization problem for computing the empirical plug-in estimator (2) in place of the minimax formulation, and it makes explicit the role that $t \mapsto f(t)$ plays in defining such a *risk-averse* version of the usual average loss $\mathbb{E}_{P_0}[\ell(\theta; X)]$. This provides an equivalence between distributional robustness and tail-performance, which we draw on subsequently both statistical and computational reasons. Defining the *uncertainty region*

$$\mathcal{U}_P := \{Q : D_f(Q\|P) \leq \rho\},$$

we may use the likelihood ratio $L(x) := dQ(x)/dP_0(x)$ to reformulate our distributionally robust problem (1) via

$$\mathcal{R}_f(\theta; P_0) = \sup_P \{\mathbb{E}_P[\ell(\theta; X)] : P \in \mathcal{U}_{P_0}\}$$
$$= \sup_{L \geq 0} \{\mathbb{E}_{P_0}[L(X)\ell(\theta; X)] \mid \mathbb{E}_{P_0}[f(L(X))] \leq \rho, \mathbb{E}_{P_0}[L(X)] = 1\}, \tag{3}$$

where the supremum is over measurable functions. We now recall Ben-Tal et al. [15] and Shapiro's dual reformulation of the quantity (3), where $f^*(s) := \sup_t\{st - f(t)\}$ is the usual Fenchel conjugate.

Proposition 1 (Shapiro [87, Section 3.2]). *Let P be a probability measure on $(\mathcal{X}, \mathcal{A})$ and $\rho > 0$. Then*

$$\mathcal{R}_f(\theta; P) = \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \left\{ \mathbb{E}_P\left[\lambda f^*\left(\frac{\ell(\theta; X) - \eta}{\lambda}\right)\right] + \lambda\rho + \eta \right\} \tag{4}$$

for all θ. Moreover, if the supremum on the left hand side is finite, there are finite $\lambda(\theta) \geq 0$ 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 (4) is jointly convex in (θ, η, λ). While interior point methods [24] 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 [71, 72].

5

Divergence families Much of our development centers on two families of divergences. The *Rényi α-divergence* [104] between distributions P and Q is

$$D_\alpha(P\|Q) := \frac{1}{\alpha - 1} \log \int \left(\frac{dP}{dQ}\right)^\alpha dQ, \tag{5}$$

where the limit as $\alpha \to 1$ satisfies $D_1(P\|Q) = D_{\mathrm{kl}}(P\|Q)$. For analytical reasons, we use the equivalent Cressie-Read family of f-divergences [29]. These are parameterized by $k \in (-\infty, \infty) \setminus \{0, 1\}$, $k_* = \frac{k}{k-1}$, with

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

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 $f_1(t) = t \log t - t + 1$, and we frequently use the shorthand

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

While most of our results generalize to other values of k, we focus temporarily on $k \in (1, \infty)$ for ease of exposition (only our finite-sample guarantees in Section 4 require $k \in (1, \infty)$). By minimizing out $\lambda \geq 0$ in the dual form (4), we obtain a simplified formulation for the Cressie-Read family (6).

Lemma 1. *For any probability P on $(\mathcal{X}, \mathcal{A})$, $k \in (1, \infty)$, $k_* = k/(k-1)$, any $\rho > 0$, and $c_k(\rho) := (1 + k(k-1)\rho)^{\frac{1}{k}}$, we have for all $\theta \in \Theta$*

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

See Section A.1 for the proof. The simplified dual form (8) shows that protecting against worst-case distributional shifts is equivalent to optimizing the tail-performance of a model; the worst-case objective $\mathcal{R}_k(\theta; P)$ only penalizes losses above the optimal dual variable $\eta^\star(\theta)$. The $L^{k_*}(P)$-norm upweights these tail values of $\ell(\theta; x)$, giving a worst-case objective that focuses on "hard" regions of $\mathcal{X}$. Eq. (8) also makes explicit the relationship between the growth f_k and the worst-case objective $\mathcal{R}_k(\theta; P)$: as growth of $f_k(t)$ for large t becomes steeper ($k \uparrow \infty$), the f-divergence ball $\{Q : D_{f_k}(Q\|P) \leq \rho\}$ shrinks, and the risk measure $\mathcal{R}_k(\theta; P)$ becomes less conservative (smaller). Since the dual form (8) quantifies this with the $L^{k_*}(P)$-norm of the loss above the quantile η, we see that f_k with $k \in (1, \infty)$ is a possible choice if the loss has finite k_*-moments under the nominal distribution P_0. In contrast, the worst-case formulation (1) corresponding to the KL-divergence ($k = 1$) is finite only when the moment generating function of the loss exists [3].[1]

An extensive literature on coherent risk measures defines utility functions that exhibit "sensible" tail risk preference [8, 78, 60, 88]; there is a duality between distributionally robust optimization and coherent risk measures [e.g. 88, Thm. 6.4]. 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 [60] proposes the dual form (8) as a higher order generalization of the classical conditional value-at-risk [78], which corresponds to $\mathcal{R}_k(\theta; P)$ defined with $k = \infty$ (or $k_* = 1$) in our notation.

---

[1]This correspondence between higher moments and divergences holds in more generality in that if $f(t)$ grows asymptotically as t^k as $t \to \infty$, then the dual exhibits similar k_*th moment behavior; see supplementary Appendix A.2.

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 and to connect to our worst-case subpopulation considerations (Examples 3–5). We begin with a generic description and specialize subsequently, deferring heuristic procedures for choosing f and ρ (and empirical efficacy evaluations) to the next section.

Example 1 (Generic distributional shift): Consider data in pairs (X, Y), where X is a feature (covariate) vector and Y is 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 marginal distribution μ on U, let $P_\mu((X, Y) \in A) := \int P_0((X, Y) \in A \mid U = u) d\mu(u)$. We have the essentially tautological correspondence

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

The robustness set is a family of distributional interventions on U. We leave characterizing the precise form of such interventions as an open question. $\diamond$

For well-specified linear models, it is frequently the case that the robust parameter $\theta_{\mathrm{dro}} \in \arg\min_\theta \mathcal{R}_f(\theta; P)$ minimizing the objective (1) coincides with the true parameter, though its plug-in estimator may be less efficient than standard ordinary least-squares estimators (we do not discuss this efficiency here).

Example 2 (Regression and stochastic domination): To make things precise, recall stochastic orders [85]: for two $\mathbb{R}$-valued random variables U and V, we say that V stochastically dominates U if $\mathbb{P}(U \ge t) \le \mathbb{P}(V \ge t)$ for all $t \in \mathbb{R}$, written $U \preceq V$; this is equivalent to the condition that $\mathbb{E}[g(U)] \le \mathbb{E}[g(V)]$ for all nondecreasing g. For any problem with data in pairs (X, Y) and a loss $\ell(\theta; X, Y)$, if there exists a parameter $\theta_\star$ such that $\ell(\theta_\star; X, Y) \preceq \ell(\theta; X, Y)$ for all θ, we then have $\theta_\star \in \arg\min_\theta \mathcal{R}_f(\theta; P)$ for all f-divergences, as $\mathcal{R}_f$ is a coherent risk measure [cf. 88, Ch. 6.3]. Existence of such $\theta_\star$ is a strong condition, but holds in a few important cases.

For concreteness consider linear regression, where $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ and $\ell(\theta; x, y) = \frac{1}{2}(\theta^T x - y)^2$. First, we consider the case that the model is well-specified, so that $Y = X^T\theta_\star + \varepsilon$, where $\mathbb{E}[\varepsilon \mid X] = 0$. If the distribution of ε given $X = x$ is symmetric and log quasiconcave (unimodal), then Anderson's theorem [7, 43, Thm. 11.1] implies that

$$\mathbb{P}(|x^T\theta - Y| \ge t \mid X = x) = \mathbb{P}(|x^T(\theta - \theta_\star) - \varepsilon| \ge t \mid X = x) \ge \mathbb{P}(|\varepsilon| \ge t \mid X = x),$$

for all $t \in \mathbb{R}$, and so $\ell(\theta_\star; X, Y) \preceq \ell(\theta; X, Y)$ for all θ, and $\theta_\star \in \arg\min_\theta \mathcal{R}_f(\theta; P)$.

In a different vein, we can consider the case that X, Y are jointly Gaussian and mean zero,

$$(X, Y) \sim \mathsf{N}\left(\mathbf{0}, \begin{bmatrix} \Sigma & \gamma \\ \gamma^T & \sigma^2 \end{bmatrix}\right).$$

Then for any θ we have $(X^T\theta - Y) \sim \mathsf{N}(0, \theta^T\Sigma\theta - 2\theta^T\gamma + \sigma^2)$, and the ordinary least-squares solution $\theta_{\mathrm{ols}} = \Sigma^{-1}\gamma = \mathbb{E}[XX^T]^{-1}\mathbb{E}[XY]$ evidently uniformly minimizes the variance of $(X^T\theta - Y)$. Once again, we thus have the stochastic dominance $\ell(\theta_{\mathrm{ols}}; X, Y) \preceq \ell(\theta; X, Y)$ for all θ, and so the robust solutions coincide with standard estimators. $\diamond$

Example 3 (Worst-case minority performance and CVaR): For $0 < \alpha \leq 1$, the conditional value-at-risk [78] (CVaR) is

$$\mathrm{CVaR}_\alpha(\theta; P_0) := \inf_{\eta \in \mathbb{R}} \left\{ \alpha^{-1} \mathbb{E}_{P_0} \left[(\ell(\theta; X) - \eta)_+ \right] + \eta \right\}.$$

This corresponds to an uncertainty set arising out of limiting f- or Rényi divergences. Recalling the Rényi divergence (5), we have $D_\infty(P\|Q) := \lim_{\alpha \to \infty} D_\alpha(P\|Q) = \mathrm{ess\,sup} \log \frac{dP}{dQ}$, and if we define $f_{\infty,c}(t) = 0$ for $0 \leq t \leq c$ and $+\infty$ otherwise, then the uncertainty region

$$\mathcal{U}_{P_0} := \left\{ P \mid D_\infty(P\|P_0) \leq \log \frac{1}{\alpha} \right\} = \left\{ P \mid D_{f_{\infty,\alpha^{-1}}}(P\|P_0) \leq 1 \right\}$$
$$= \{ P \mid \text{ there exists } Q, \ \beta \in [\alpha, 1] \text{ s.t. } P_0 = \beta P + (1 - \beta)Q \}$$

by a calculation [88, Example 6.19]. The uncertainty set corresponds to distributions with minority sub-populations of size at least α, and $\mathrm{CVaR}_\alpha(\theta; P_0) = \sup_{P \in \mathcal{U}_{P_0}} \mathbb{E}_P[\ell(\theta; X)]$ is the expected loss of the worst α-sized subpopulation. ◇

The Kusuoka representation [86, 61] of risk measures shows that the robust formulations (1) are worst-case CVaR mixtures, $\mathcal{R}_f(\theta; P_0) = \sup_{\mu \in \mathcal{M}_f} \int_0^1 \mathrm{CVaR}_\alpha(\theta; P_0) d\mu(\alpha)$ for a set $\mathcal{M}_f$ of probability measures on $[0, 1]$. They thus correspond to drawing a random sub-population size α and measuring the loss of the worst subpopulation of P_0 mass at least α. Precisely connecting the subpopulation size and robustness set $\{P : D_f(P\|P_0) \leq \rho\}$ is challenging.

We now consider two examples in which data comes from latent *mixtures* of populations, where within each subpopulation a model is well-specified, though it is not globally. In both of these cases—mean estimation and a linear regression problem—we see that as the robustness parameter $\rho \uparrow \infty$ in the DRO formulation (1), the robust estimator converges to the minimax estimator minimizing the worst-case loss across all sub-populations. This recalls Meinshausen and Bühlmann [70], who consider min/max effects in heterogeneous regression problems with known group identities, but here the DRO estimator recovers a minimax estimator *without* such knowledge. The examples are stylized to give explicit limits, though they convey the intuition that the robust estimators seek to do well on unknown sub-populations in a reasonably precise way. In each example, we consider the conditional value at risk (Ex. 3) for simplicity; the results for higher-order robustness measures are similar but tedious.

Example 4 (Mixtures in mean estimation): Consider a finite number of distinct populations on $\mathbb{R}^d$ indexed by $v \in V$, each appearing with probability $p_v > 0$, where under population v, we observe

$$Y = \theta_v + \varepsilon, \quad \varepsilon \overset{\mathrm{iid}}{\sim} \mathsf{N}(0, I_d).$$

Letting the loss $\ell(\theta; y) = \frac{1}{2} \|\theta - y\|_2^2$, we define the minimax estimator

$$\theta_{\mathrm{minimax}} := \underset{\theta}{\mathrm{argmin}} \max_{v \in V} \|\theta - \theta_v\|_2^2 = \underset{\theta}{\mathrm{argmin}} \max_{v \in V} \mathbb{E}_v[\|\theta - Y\|_2^2].$$

The unique vector $\theta_{\mathrm{minimax}}$ coincides with the Chebyshev center of the vectors $\{\theta_v\}$ [24, Ch. 8.5]; it also requires knowledge of the groups $v \in V$. In Appendix B.1, we show that if $\theta_\alpha = \mathrm{argmin}_\theta \, \mathrm{CVaR}_\alpha(\ell(\theta; Y))$, then

$$\theta_1 = \sum_v p_v \theta_v \quad \text{and} \quad \lim_{\alpha \downarrow 0} \theta_\alpha = \theta_{\mathrm{minimax}}.$$

8

Recalling from Ex. 3 that the parameter α is inversely proportional to the robustness in the DRO formulation, we see the expected behavior: as robustness increases, the DRO estimator converges to an estimator minimizing the worst sub-population expected loss. $\diamond$

Example 5 (Mixtures in linear regression)**:** We expand the previous example to allow covariates and potentially infinite subgroups. For groups indexed by $v \in V$, we draw $v \in V$ according to a probability measure μ on V, and then conditional on v draw

$$X \sim \mathsf{N}(0, \Sigma_v), \;\; \varepsilon_v \sim \mathsf{N}(0, \sigma_v^2), \;\; Y = X^T \theta_v + \varepsilon_v, \tag{9}$$

assuming implicitly that all parameters are v-measurable. (To show the result in the most straightforward way, we make the simplifying assumptions that $0 < \inf_v \sigma_v^2 \leq \sup_v \sigma_v^2 < \infty$, that the eigenvalues of Σ_v are finite and bounded away from 0 uniformly in v, that $\sup_v \|\theta_v\| < \infty$, and we also assume that for each $\theta \in \mathbb{R}^d$, we have ess $\sup_v (\theta - \theta_v)^T \Sigma_v (\theta - \theta_v) + \sigma_v^2 = \sup_v (\theta - \theta_v)^T \Sigma_v (\theta - \theta_v) + \sigma_v^2$. Each of these assumptions is trivial when there are a finite number of groups.)

Letting $\mathbb{E}_v$ denote expectation according to the model (9), let $\ell(\theta; x, y) = \frac{1}{2}(x^T \theta - y)^2$ be the standard squared error and consider the conditional value at risk

$$\mathrm{CVaR}_\alpha(\ell(\theta; X, Y)) = \inf_\eta \left\{ \frac{1}{\alpha} \int \mathbb{E}_v \left[(\ell(\theta; X, Y) - \eta)_+ \right] d\mu(v) + \eta \right\}.$$

We define the minimax estimator to minimize the worst sub-population risk

$$\theta_{\mathrm{minimax}} = \operatorname*{argmin}_\theta \sup_{v \in V} \left\{ \mathbb{E}_v[(\theta^T X - Y)^2] = (\theta - \theta_v)^T \Sigma_v (\theta - \theta_v) + \sigma_v^2 \right\}.$$

In this case, for the distributionally robust parameter $\theta_\alpha := \operatorname{argmin}_\theta \mathrm{CVaR}_\alpha(\ell(\theta; X, Y))$ and ordinary least squares solution $\theta_{\mathrm{ols}} = \operatorname{argmin}_\theta \mathbb{E}[\ell(\theta; X, Y)]$, we show in Appendix B.2 that

$$\theta_{\mathrm{ols}} = \theta_1 = \int \theta_v d\mu(v) \quad \text{and} \quad \lim_{\alpha \downarrow 0} \theta_\alpha = \theta_{\mathrm{minimax}}.$$

We again see the interpolation from an average parameter to one that minimizes the worst-case subpopulation risk as the robustness increases (i.e. $\alpha \downarrow 0$). $\diamond$

3 Empirical analysis, validation, and choice of uncertainty set

As this paper proposes and argues for alternatives to empirical risk minimization and standard M-estimation—workhorses of much of machine learning and statistics [102, 103, 55]—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

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

For concreteness, we focus on the Cressie-Read (equivalently Rényi) divergence family (6) with $k \in (1, \infty)$, experimenting on three related challenges:

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

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 subpopulations).

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

Since the choice of f and ρ governs the trade-off between average and tail performance, we propose two heuristics for choosing ρ and k, evaluating their performance on simulated examples. Our heuristics aim to provide uniform performance over difficult inputs by considering proxy sub-populations constructed from the training data, though to be clear, the only formal guarantees on robustness they provide is robustness to shifts contained in specified by f_k for the chosen and k (the duality relationships (4) and (8) makes the robustness less sensitive to ρ). Our first heuristic splits the training dataset into s equi-sized groups based on the values of the response variable Y, where Y has highest values in the first group, and the lowest values in the last sth group. We split each of the s groups into 80%/20% training/validation splits, and re-unify all of the 80% splits to give a new training dataset with 80% of the original data. We train our robust models (2) (varying ρ and k) on the new training dataset, evaluating these models on the unused data from each group (20%), giving s different empirical losses for a given model. A model's score is then its empirical loss on the *worst* of the s held-out sets. We use $s = 5$ groups since this consistently gives a good selection procedure across different settings. As our second heuristic, we consider scenarios where more is known about the problem. If a small auxiliary dataset collected from a worse-than-average subpopulation is available, we tune ρ and k on this auxiliary dataset so that heuristically, the re-sulting model performs uniformly well against all subpopulations of a *similar size* (the worst-case formulation (2) optimizes performance only over large enough subpopulations e.g. Example 3). Empirically, we observe that the second heuristic performs well even on rare subgroups that are far from the subpopulation generating the auxiliary dataset. On simulation examples, we observe good worst-case subpopulation performance for both procedures, with moderate degradation in the average-case performance.

We begin with simulation experiments that touch on all three of above challenges in Section 3.1. To investigate these challenges on different real-world datasets, in Section 3.2 we study domain adaptation in the context of predictors trained to recognize handwritten digits, then test them to recognize typewritten digits. 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 subpopulation performance.

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
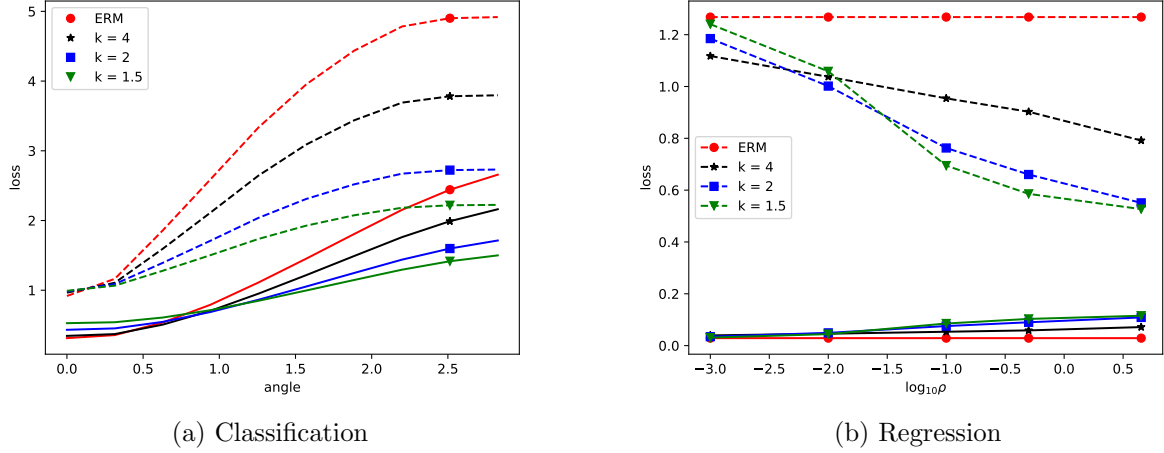
10

(a) Classification   (b) Regression

Figure 1. (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 (solid-line) and majority group (dotted-line) under the distribution (11). We define the minority group as those with $X^1 \leq z_{.95}$.

using a conic interior point solver; we extended the open-source Julia package `convex.jl` to implement power cone solvers [100] (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 [24]. 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 [51, Theorem VI.4.4.2] with

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

We use a fast bisection method [72] to compute Q^* at every iteration of our first-order method; see https://github.com/hsnamkoong/robustopt for the implementation.

3.1 Simulation

Our first experiments use simulated data, where we fit linear models 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 $d = 5$. We choose a vector $\theta_0^\star \in \mathbb{R}^5$ uniformly on the unit sphere and generate data

$$X \overset{\text{iid}}{\sim} \mathsf{N}(0, I_d) \quad \text{and} \quad Y \mid X = \begin{cases} \operatorname{sign}(X^\top \theta_0^\star) & \text{w.p. } 0.9 \\ -\operatorname{sign}(X^\top \theta_0^\star) & \text{w.p. } 0.1. \end{cases} \tag{10}$$

11

(Our below observations still hold when varying these probabilities.) 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 $s \in [0, \pi]$ define $\theta_s^\star = \theta_0^\star \cdot \cos s + v \cdot \sin s$, so that $\theta_\pi^\star = -\theta_0^\star$. For each perturbation, we generate $n_{\text{test}} = 100,000$ test examples using the same scheme (10) 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 $f_k(t) \propto t^k$), the better the fitted model should perform on high quantiles of the loss, with potentially worse average performance. Moreover, for $s = 0$, we should see that ERM and large k solutions exhibit the best average performance, with growing s reversing this behavior. In Figure 1(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 $s \in [0, \pi]$. The plot bears out our intuition: the distributionally robust solution $\widehat{\theta}_n$ has worse *mean* loss on the original distribution than empirical risk minimization (ERM) while achieving significantly smaller loss on the distributional shifts. The ordering of the mean performance of the different solutions inverts as the perturbation grows: under no perturbation ($s = 0$), the least robust method (ERM) has the best performance, while the most robust method (corresponding to $k = \frac{3}{2}$) performs the best under large distributional perturbations (s large).

3.1.2 Tail performance

We transition now to regression, investigating performance on rare examples, 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 take $d = 5$ and generate data $X \overset{\text{iid}}{\sim} N(0, I_d)$, $\varepsilon \sim N(0, .01)$,

$$Y = \begin{cases} X^\top \theta^\star + \varepsilon & \text{if } X^1 \leq z_{.95} = 1.645 \\ X^\top \theta^\star + X^1 + \varepsilon & \text{otherwise,} \end{cases} \tag{11}$$

where 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 $X^1 \geq 1.645$, or in the tails generally. In addition, a fully robust solution is $\theta^{\text{rob}} = \theta^\star + \frac{1}{2}e_1$, as this minimizes worst-case expected loss across the two cases (11); 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 1(b), we plot the mean loss under the data generation scheme (11) as solid lines and the 90%-quantile as a dotted line. We see once again that the robust solutions trade tail performance for average-case performance. The tail performance (90%-quantile loss) improve with increasing robustness level ρ, with slight degradation in average case performance.

3.1.3 Performance on different subpopulations

For our final small-scale simulation, we study item 3 (subpopulation performance) by considering a two-dimensional regression problem with heterogeneous subpopulations. We consider two scenarios: a two-group setting and an infinite number of groups. In each scenario, we demonstrate
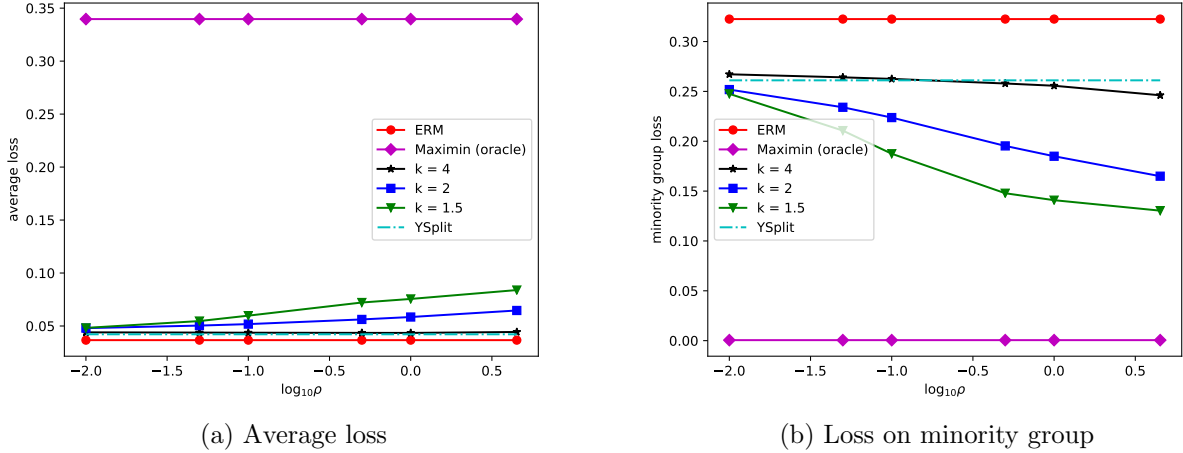
(a) Average loss   (b) Loss on minority group

Figure 2. Two groups: Figures (a) and (b) plots average and minority group losses under the distribution (13). "YSplit" is the performance of the model whose ρ and k was chosen based on groups formed by sorted values of Y.

the performance of our heuristic procedure for choosing ρ and k; these subpopulation scenarios are appropriate for succinctly characterizing the trade-off between average and tail subpopulations. Our tuning procedure provides good performance on the (latent) worst-case subpopulation even when the proxy subpopulation for tuning ρ and k is far from the rare subpopulation. In what follows, we denote by "YSplit" our first proposal that chooses ρ and k based on sorted values of Y.

Two groups In our first scenario, for $\theta_0^\star = (1, .1)$, $\theta_1^\star = (1, 1)$, we generate

$$Y = X^\top((1 - G)\theta_0^\star + G\theta_1^\star) + \varepsilon \tag{12}$$

where $X \sim N(0, I_2)$, $\varepsilon \sim \mathsf{N}(0, .01)$, and $G \in [0, 1]$ indicates a random *latent group*. We assume that X, G and ε are mutually independent. Both the distributionally robust procedure (2) and ERM are oblivious to the label G, where we think of $G = 1$ as the *majority* group, and $G = 0$ as the *minority* group. We simulate $n_{\text{train}} = 1000$ training data points, and train ERM and robust models (2) on varying values of k and ρ. We let

$$G = \begin{cases} 0 & \text{with probability } .1 \quad (\text{minority}) \\ 1 & \text{with probability } .9 \quad (\text{majority}) \end{cases} \tag{13}$$

In this two-group setting, we also consider the maximin effects estimator [70]

$$\widehat{\theta}_n^{\text{maximin}} = \operatorname*{argmax}_\theta \min_{g=0,1} \left\{ 2\theta^\top \widehat{\Sigma}_{n,g} \theta_g^\star - \theta^\top \widehat{\Sigma}_{n,g} \theta \right\}$$

as a benchmark, where $\widehat{\Sigma}_{n,g}$ is the empirical covariance matrix of the X_i with $G_i = g$, which maximizes the explained variance for each group [70]. The oracle estimator $\widehat{\theta}_n^{\text{maximin}}$ requires knowledge of the labels G_i and the group-specific regressors $\theta_g^\star$ for $g = 0, 1$.

In Figure 2(a) and (b), we plot the average and minority group losses for the different methods, respectively. Here the robust methods interpolate between the empirical risk minimizing
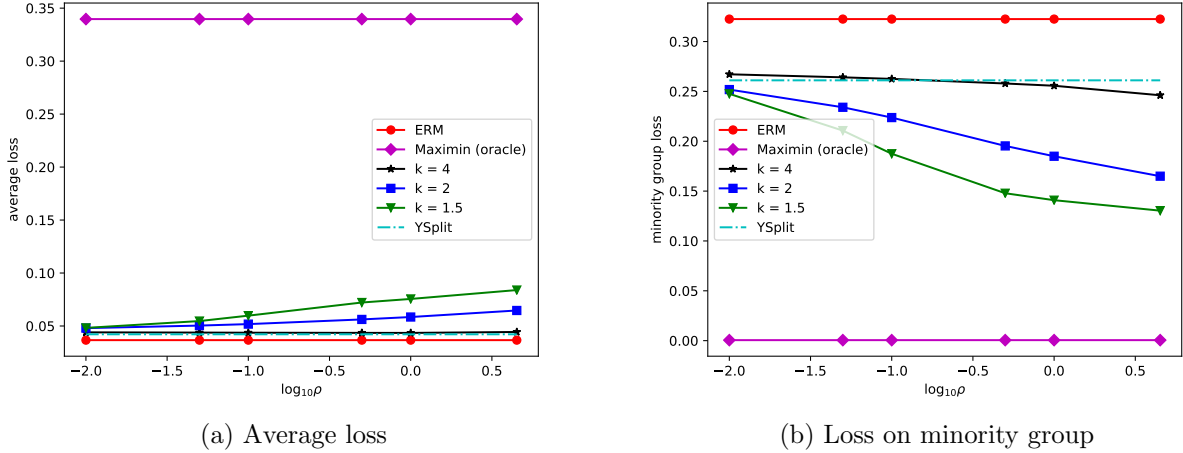
13

(a) Average loss

(b) Loss on minority group

Figure 3. Infinite groups: Figures (a) and (b) plot average and minority group losses under the distribution (14). "YSplit" is the performance of the model whose ρ and k was chosen based on groups formed by sorted values of Y, and "$G = .5$" chose k and ρ based on auxiliary data with intervention $G = 0.5$.

(ERM) solution—which has the best average loss and worst minority group loss—and the maximin estimator $\widehat{\theta}_n^{\mathrm{maximin}}$, 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. Our first heuristic ("YSplit") chooses ρ and k based on groups formed by sorted values of Y, and improves minority performance while sacrificing very little average-case performance.

Infinite groups For our last scenario, we again generate X and Y following the equation (12), but with

$$G \sim P_G \text{ with density } p_G(g) \propto (1 - g)^{-\frac{1}{3}}, \tag{14}$$

so small values of G again correspond to rare minority subpopulations. To study how k and ρ can be tuned if a small auxiliary dataset is available, we generate a small auxiliary dataset from the distribution (12) with group $G = .5$, which we interpret as a particular group intervention; we simulate $n_{\mathrm{auxiliary}} = 100$ observations from $G = .5$, which is small compared to $n_{\mathrm{train}} = 1000$ training examples. We refer to choosing k and ρ with the least prediction error on this auxiliary validation data as the "$G = 0.5$" method.

As earlier, we plot in Figure 3(a) and (b) the average and minority group ($G = 0$) losses for the different methods. The minority group $G = 0$ now never appears in the training set, and small values of G are rare under the distribution (14). Our first heuristic "YSplit" chooses a model that balance average and minority performance, although it is somewhat conservative. Our second proposal, the $G = 0.5$ method, achieves good performance on the rare minority group while sacrificing little average performance, despite the fact that the auxiliary data was collected from the group $G = 0.5$ that is far from the minority group $G = 0$.

14

3.2 Domain generalization for classification and digit recognition

(a) MNIST hand-written Digits

(b) All type-written Digits

(c) Type-written Digit 3 (easy class)
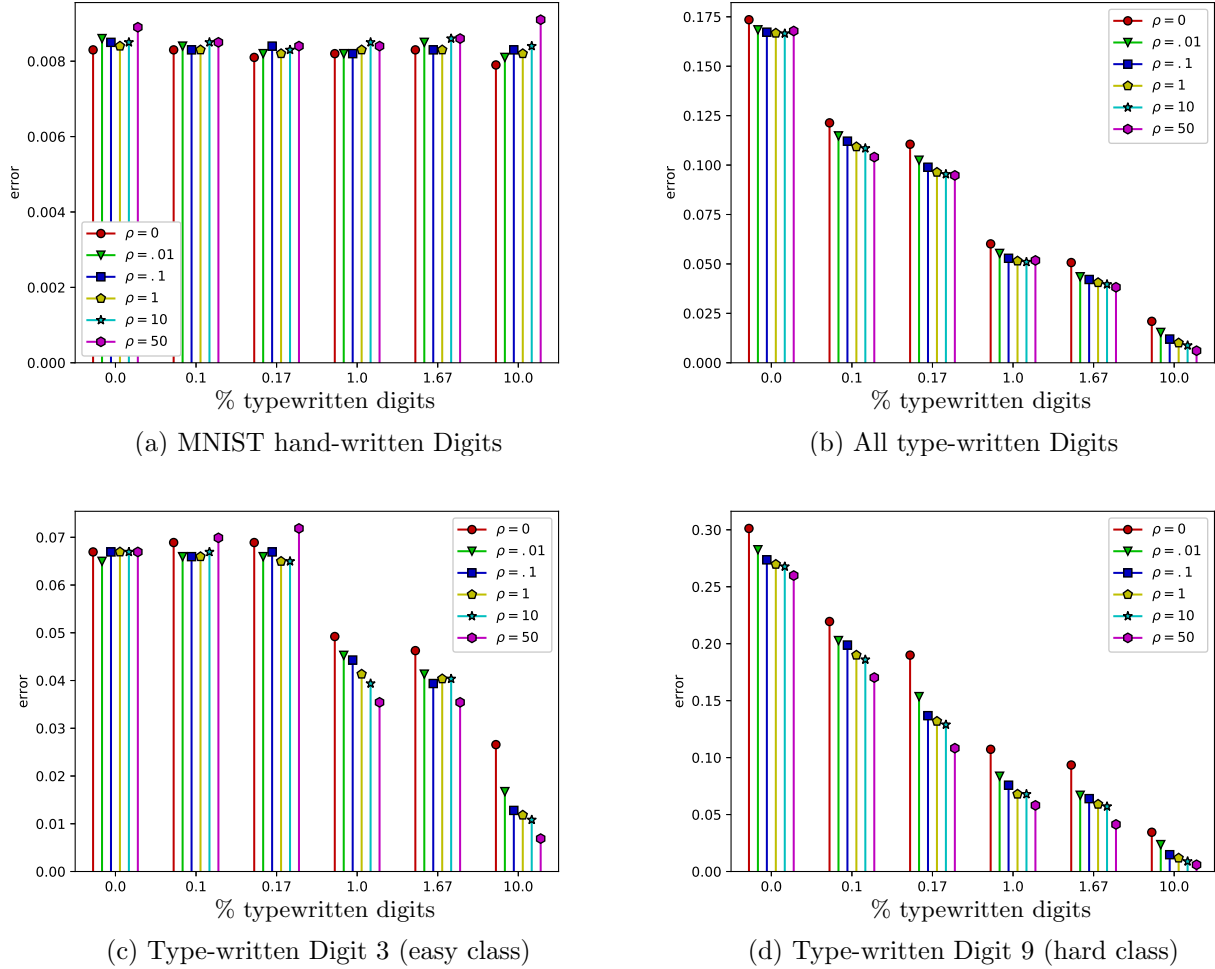
(d) Type-written Digit 9 (hard class)

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.

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 [34] (majority population) and type-written digits consisting of different fonts [32] (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).

15

| 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 (%)

Our classifiers have no knowledge of whether an 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 $\widehat{\theta}_n^{\mathrm{erm}}$, where we vary ρ and the latent minority proportion. We evaluate our classifiers on both hand- and type-written digits on held-out test sets.

For $y \in \{0, \ldots, 9\}$ and $x \in \mathbb{R}^d$, we use the multi-class logistic loss $\ell(\theta; (x, y)) = \log(\sum_{i=0}^{k} \exp((\theta_i - \theta_y)^\top x))$, where $\theta_i \in \mathbb{R}^d$. For our feature vector X, we use the $d = 4509$-dimensional output of the final fully connected layer of LeNet [66] after 10^4 stochastic gradient steps on the training dataset (see [56] for detailed hyper-parameter settings). We constrain our parameter matrix $[\theta_0, \ldots, \theta_9]$ to lie in the Frobenius norm ball of radius $r = 5$, chosen by cross validation on ERM ($\rho = 0$).

Returning to the justification for our development, we expect our robust models to exhibit better performance on rare and difficult test data when compared against ERM models. This prediction is mostly consistent with our observations, though the effects are not always strong. We suspect this is because the test data we construct is different from the worst-case scenario; the procedure (2) can be conservative as it guarantees uniform performance by optimizing the worst-case performance. In Figure 4, we plot the classification errors over the minority proportion as we vary ρ (so that $\rho = 0$ corresponds to ERM), 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 $\widehat{\theta}_n^{\mathrm{erm}}$ 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 [77, 9], studying the performance 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 [77]). We use the absolute loss $\ell(\theta; (x, y)) = |\theta^\top x - y|$ and compare method (2) with constrained forms of
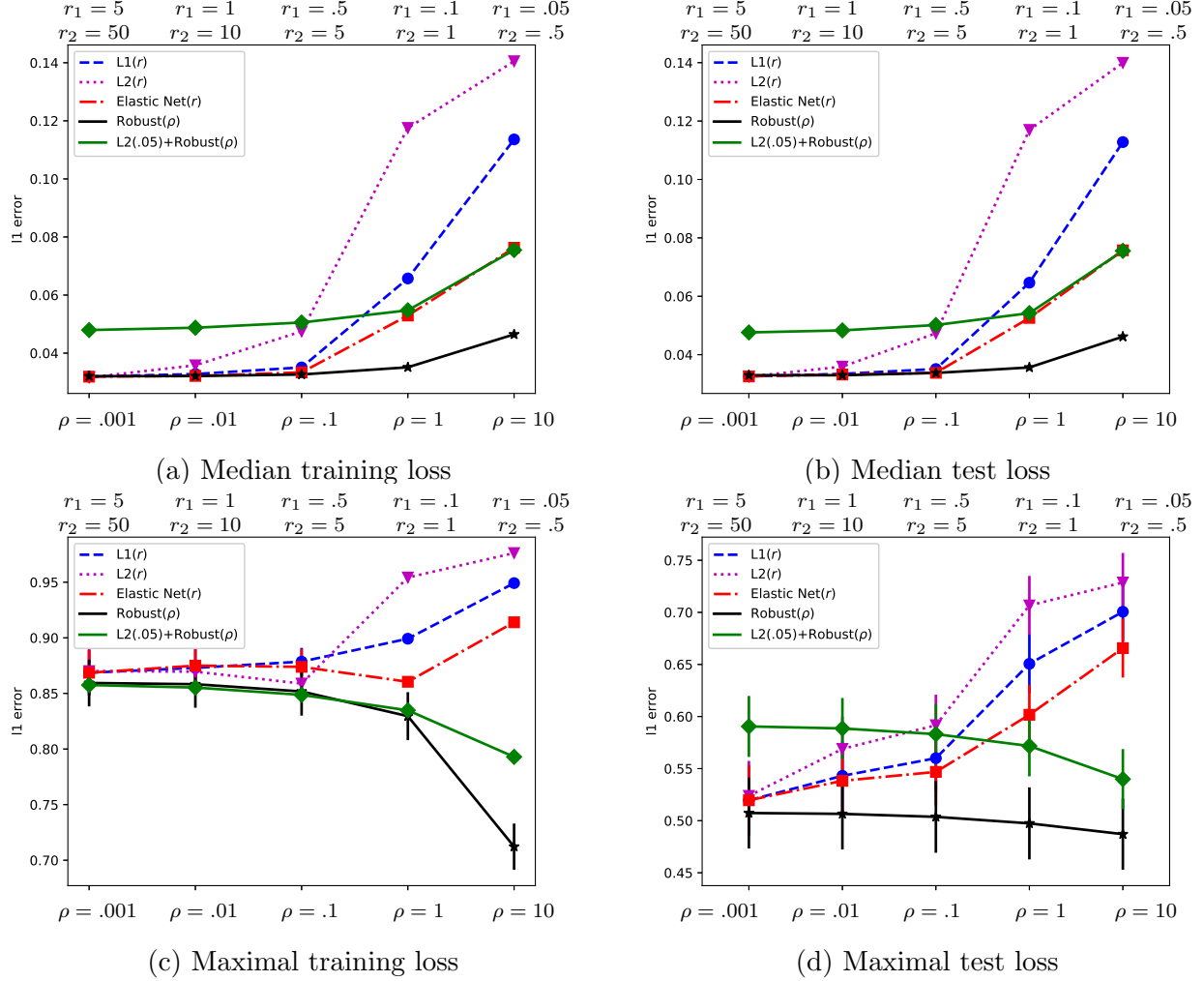
(a) Median training loss

(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.

lasso, ridge, and elastic net regularization [108], taking constraints of the form

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

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 $\rho \in \{.001, .01, .1, 1, 10\}$.
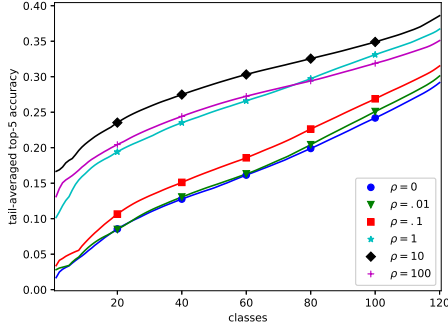
In Figure 5, we plot the quantiles of the training and test losses with respect to different values of regularization or ρ. The horizontal axis in each figure indexes our choice of regularization value. We observe that $\widehat{\theta}_n$ shows very different behavior than other regularizers; $\widehat{\theta}_n$ attains median
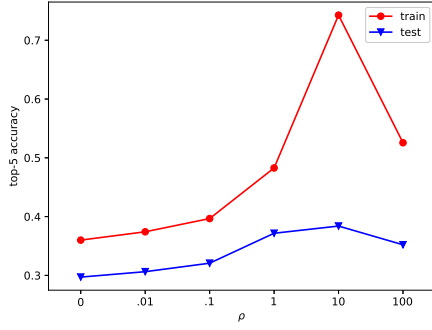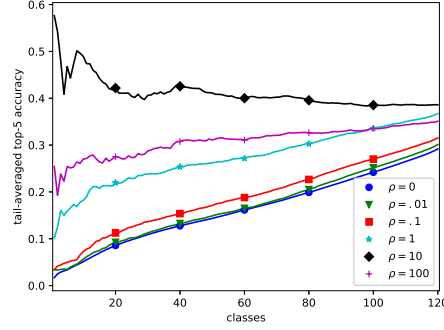
(a) Overall top-5 accuracy
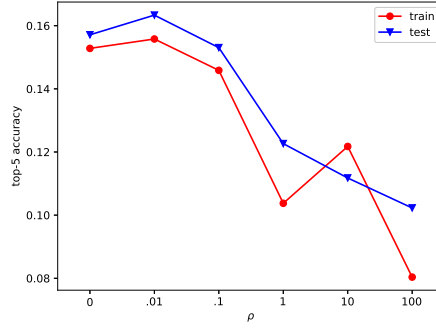
(b) Standard deviation of top-5 accuracy

(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 ($\rho = 0$).

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 $n_{\text{train}} = 1800$ and a test set with $n_{\text{test}} = 318$.

3.4 Fine-grained recognition and challenging sub-groups

Finally, we consider the fine-grained recognition task of the `Stanford Dogs` dataset [58], 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 [101], resulting in vectors $x \in \mathbb{R}^d$ with $d =12,000$.

We train 120 one-versus-rest classifiers, one each class, 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 $\Theta_{\text{one−vs−rest}} = \left\{ \theta \in \mathbb{R}^d : \|\theta\|_1 \leq r \right\}$. Thus, for each class i, we represent a pair (x, y) by $y = 1$ if x 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 m highest scores, we measure

18

performance with respect to top-m accuracy, which counts the number of test examples in which the true label was among these m predictions. As ρ grows larger, we expect better performance on challenging classes, sacrificing performance on easier classes, and due to uniform performance, for the variance in the class-wise accuracies to be smaller, though we do not necessarily expect that average accuracies should improve as ρ increases.

In Figure 6, we present top-5 accuracies; top-1 and top-3 accuracies are similar. Overall accuracy *improves* moderately as ρ grows (Figure 6(a)), and the *standard deviation* of the top-5 accuracy across the classes decreases as ρ increases (Figure 6(b)), consistent with our hypothesis that the robust formulations should yield more uniform performance across different subpopulations. In Figure 6(c), we plot the accuracy averaged over 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 $\widehat{\theta}_n^{\mathrm{erm}}$'s. In Figure 6(d), we plot the accuracy averaged over the first c-classes that have the lowest accuracy under the ERM model. We see that robust solutions $\widehat{\theta}_n$ improve 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. 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. Although it is hard to draw conclusions from this experiment due to improved overall performance when increasing ρ, we conjecture that is due to the regularization effect for relatively small values of ρ described by many previous authors [47, 62, 36, 64, 72].

4 Convergence Guarantees

Our empirical experiments in the previous section evidence the potential statistical benefits of the distributionally robust estimator (2). As a consequence, we view it as important to develop some of its theoretical properties, so we investigate its performance under a variety of conditions on the f-divergence, providing finite sample convergence guarantees for f-divergences with $f(t) \asymp t^k$ with $k \in (1, \infty)$. Recalling the definition (7) of worst-case risk $\mathcal{R}_k(\theta; P_0)$ for the Cressie-Read divergences (6), we show that the empirical minimizer $\widehat{\theta}_n$ for the plug-in (2) satisfies $\mathcal{R}_f(\widehat{\theta}_n; P_0) - \inf_{\theta \in \Theta} \mathcal{R}_f(\theta; P_0) \le Cn^{-\frac{1}{k_* \vee 2}}$ 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^{k_*}-norms of random variables X for $k_* > 2$; that is, the minimax rate for such estimation is n^{-1/k_*} for $k_* > 2$.

Throughout this section, we assume that for any $\theta \in \Theta$ and $x \in \mathcal{X}$, we have $\ell(\theta; x) \in [0, M]$ for some $M \ge 1$, 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 $\mathcal{R}_k(\theta; P_0)$; we use convex concentration inequalities [23, 95] to show concentration of $\mathcal{R}_k(\theta; \widehat{P}_n)$ to $\mathbb{E}[\mathcal{R}_k(\theta; \widehat{P}_n)]$, and then carefully bound the bias of $\mathbb{E}[\mathcal{R}_k(\theta; \widehat{P}_n)]$ in estimating 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 $c_k(\rho) := (k(k - 1)\rho + 1)^{1/k}$. For a fixed $\theta \in \Theta$ and $t > 0$, whenever $n \geq k \vee 3$, with probability at least $1 - 2e^{-t}$*

$$\left| \mathcal{R}_k(\theta; \widehat{P}_n) - \mathcal{R}_k(\theta; P_0) \right| \leq 10 n^{-\frac{1}{k_* \vee 2}} c_k(\rho)^2 M \left(\frac{c_k(\rho)}{c_k(\rho) - 1} \vee 2 \right) \left(\frac{1}{k} + \sqrt{t + 2 \log n} \right).$$

See Section C.1 for the proof. Relaxing the boundedness assumption $\ell(\theta; x) \in [0, M]$ to sub-Gaussian or sub-exponential tails, or providing similar finite-sample guarantees for general f-divergences are topics of future research.

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 $\{\ell(\theta; \cdot) : \theta \in \Theta\}$ (e.g. [103]). A collection $v_1, \ldots, v_N$ is an ϵ-*cover* of a set 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* 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 $\|h\|_{L^\infty(\mathcal{X})} := \sup_{x \in \mathcal{X}} |h(x)|$, a covering argument gives a uniform concentration result, where we use

$$\epsilon_{t,n,k}(\rho) := n^{-\frac{1}{k_* \vee 2}} c_k(\rho)^2 \left(\frac{c_k(\rho)}{c_k - 1} \vee 2 \right) \left(\frac{1}{k} + \sqrt{t + 2 \log n} \right).$$

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

$$\sup_{\theta \in \Theta} \left| \mathcal{R}_k(\theta; \widehat{P}_n) - \mathcal{R}_k(\theta; P_0) \right| \leq 30 M \epsilon_{t,n,k}(\rho).$$

See Section C.2 for the proof. From Corollary 1, we immediately get below.

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

$$\mathcal{R}_k(\widehat{\theta}_n; P_0) \leq \inf_{\theta \in \Theta} \mathcal{R}_k(\theta; P_0) + 60 n^{-\frac{1}{k_* \vee 2}} c_k^2 M \left(\frac{c_k}{c_k - 1} \vee 2 \right) \left(\frac{1}{k} + \sqrt{t + 2 \log n} \right).$$

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 [103, 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

$$\mathcal{R}_k(\widehat{\theta}_n; P_0) \leq \inf_{\theta \in \Theta} \mathcal{R}_k(\theta; P_0) + 60 n^{-\frac{1}{k_* \vee 2}} c_k^2 (LD + M_0) \left(\frac{c_k}{c_k - 1} \vee 2 \right) \left(\frac{1}{k} + \sqrt{t + 2d \log(2n)} \right)$$

with probability at least $1 - 2 \exp(-t)$. Replacing covering numbers in the above guarantees with Rademacher averages or their localized variants [11] and leveraging Rademacher contraction inequalities [67] remain open.

5 Lower Bounds

To complement our uniform upper bounds, 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 risk

$$\mathfrak{M}_n(\mathcal{P}, f, \ell) := \inf_{\widehat{\theta}_n} \sup_{P_0 \in \mathcal{P}} \; \mathbb{E}_{P_0^n} \left[\mathcal{R}_f \left(\widehat{\theta}_n(X_1^n); P_0 \right) - \inf_{\theta \in \Theta} \mathcal{R}_f \left(\theta; P_0 \right) \right] \tag{15}$$

where the outer infimum is over all $(X_1, \dots, X_n)$-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 (15) where $k_* = k/(k-1)$. Thus there is a necessary transition from parametric $\sqrt{n}$-type rates to n^{1/k_*} when k is small—that is, when we seek protection against large distributional shifts.

It is of interest both to *estimate* the value of the risk $\mathcal{R}_f$—see the literature on risk measures we reference in the introduction—and to minimize it. Consequently, we divide our lower bounds into estimation rates on the value $\mathcal{R}_f(\theta; P_0)$ and on the actual minimax risk (15) for the optimization problem (1), 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 provide lower bounds for general f-divergences using the same techniques. The rough intuition for our approach is as follows: we consider Bernoulli variables $Z \in \{0, M\}$, 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 of the robust risk value

For the rest of this subsection, we fix any $\theta \in \Theta$, and consider $Z(x) := \ell(\theta; x)$, abusing notation by writing $\mathcal{R}_f(Z) := \sup_{D_f(Q\|P_0) \le \rho} \mathbb{E}_Q[Z]$ and $\mathcal{R}_k(Z) := \mathcal{R}_f(Z)$ if $f = f_k$ is a Cressie-Read divergence (6). We are interested here in the minimax error for estimating the robust risk $\mathcal{R}_f(Z)$ itself, rather than any optimization over θ (justifying our abuse $Z(x) = \ell(\theta; x)$), studying

$$\mathfrak{M}_n(\mathcal{P}, f) := \inf_{\widehat{R}} \sup_{P_0 \in \mathcal{P}} \; \mathbb{E}_{P_0^n} \left| \widehat{R}(Z_1^n) - \mathcal{R}_f(Z) \right|, \tag{16}$$

where $Z \sim P_0$ and $Z_1^n \overset{\text{iid}}{\sim} P_0$, and the outer infimum is over $\widehat{R} : \{0, M\}^n \to \mathbb{R}$. Throughout this section, we let $\mathcal{P}$ be the collection of distributions on $Z \in \{0, M\}$ for a fixed $M > 0$.

We first establish a lower bound for estimating $\mathcal{R}_k(Z) = \mathcal{R}_k(\theta; P_0)$ under the Cressie-Read family f_k (6); see Section D.1 for the proof. Our proof uses Le Cam's method [107, 65], by noting 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. Define $c_k(\rho) := (1 + k(k-1)\rho)^{1/k}$, $p_k := (1 + k(k-$*

$1)\rho)^{-1/(k-1)}$, and $\beta_k = \frac{k(k-1)\rho}{2(1+k(k-1)\rho)}$. Then

$$\mathfrak{M}_n(\mathcal{P}, f_k) \geq M \max \left\{ \frac{1}{8k_* p_k} \left(\sqrt{\frac{p_k(1-p_k)}{8n}} \wedge \frac{1}{2}(1-p_k) \wedge p_k \right), \right.$$

$$\left. \frac{1}{8} \beta_k^{\frac{1}{k}} c_k(\rho) \left(\frac{1}{4n} \wedge p_k \wedge (1-(1-\beta_k)^{1-k_*} p_k) \right)^{\frac{1}{k_*}} \right\}.$$

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. First, we give canonical $\Omega(n^{-1/2})$ lower bounds. We assume 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 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

$$q(p) := \sup_{q \geq p}\{q : h_f(q; p) \leq \rho\}. \tag{17}$$

(Moreover, q is nondecreasing and concave in p, so it is a.e. differentiable.) We then have the following $\Omega(n^{-1/2})$ lower bound.

Proposition 4. *Let $f : (0, \infty) \to \mathbb{R} \cup \{+\infty\}$ be strictly convex at $t = 1$. 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,*

$$\liminf_{n \to \infty} \sqrt{n}\mathfrak{M}_n(\mathcal{P}, f) \geq M\frac{\sqrt{p(1-p)}}{8}\frac{-\partial_p h_f(q(p); p)}{\partial_q h_f(q(p); p)} > 0.$$

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

If the asymptotic growth of f is at most t^k, we can give an $\Omega(n^{-1/k_*})$ lower bound, which we prove in Section D.3. Letting $f^{-1}(s) := \inf\{t \in [0, 1] : f(t) \leq s\}$ and $m > 0$, define

$$C_{f,\rho,m} := \frac{m}{\rho}\left(1 \wedge \left(\frac{\rho}{2m}\right)^{-k_*}\left(1 - f^{-1}\left(\frac{\rho}{2}\right)\right)^{k_*}\right)^{-1}. \tag{18}$$

Proposition 5. *Let $m > 0$ and $k \in (1, \infty)$. If $f(t) \leq mt^k$ for $t \geq \{(n \vee C_{f,\rho,m})\rho m^{-1}\}^{\frac{1}{k}}$, then*

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

5.2 Lower bounds on optimization

Our lower bounds on optimization build on those for estimating $\mathcal{R}_f$. We consider linear losses, which makes the situation closest to the estimation of the risk results in the previous section (as we must still estimate kth norms of random variables), providing analogous lower bounds for optimizing the worst-case objective $\mathcal{R}_f(\cdot; P_0)$. Using a standard notion of distance for proving lower bounds in stochastic optimization [2, 35], we construct a reduction from distributionally robust optimization to hypothesis testing. Throughout, we let $\mathcal{P}$ be the set of distributions with $x \in [-1, 1]$ almost surely. We begin by considering the lower bound for the Cressie-Read family (6) f_k, whose proof we give in Section D.4.

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

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

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 D.5.

Proposition 7. *Let $\ell(\theta; x) = \theta x$ where $\theta \in \Theta = [-M, M]$ and $X \in [-1, 1]$. If the conditions on f of Proposition 4 hold,*

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

For f-divergences with $f(t) = O(t^k)$ as $t \to \infty$, we can again prove a $\Omega(n^{-1/k_*})$ lower bound on optimizing $\mathcal{R}_f(\cdot; P_0)$. Recalling the definition (18) of $C_{f,\rho,m}$, we obtain the following result, whose proof we give in Section D.6.

Proposition 8. *Let $\ell(\theta; x) = \theta x$ where $\theta \in \Theta = [-M, M]$ and $X \in [-1, 1]$. If the conditions on f of Proposition 5 hold,*

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

In terms of rates in n, there is a tradeoff between convergence rates and robustness, as measured by the asymptotic growth of the function f defining the robustness set $\{P : D_f(P\|P_0) \leq \rho\}$. In this sense, our finite sample convergence guarantees of Section 4 are sharp. All results in this section can be stated in a probabilistic form that matches our high probability guarantees in the previous section; see the remark in the beginning of Section D.

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 [79, 59]. Our first results shows that $\mathcal{R}_f(\theta; \widehat{P}_n)$ is pointwise consistent for its population counterpart $\mathcal{R}_f(\theta; P_0)$. See Section E.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 $\mathcal{R}_f(\theta; \widehat{P}_n) \overset{a.s.}{\to} \mathcal{R}_f(\theta; P_0) < \infty$.*

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 and 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, though computation of the plug-in (2) may be difficult. Coercivity guarantees the existence and compactness of the set of optima for $\mathcal{R}_f(\theta; P_0)$.

Define the *inclusion distance*, or the *deviation*, from a set A to B as

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

This is an one-sided notion of the Hausdorff distance $d_H(A, B) = \max\{d_{\subset}(A, B), d_{\subset}(B, A)\}$. For any $\varepsilon \geq 0$ and distribution P, define the set 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. The following consistency result shows that approximate empirical optimizers are eventually nearly in the population optima $S_{P_0}(\Theta)$; we provide its proof in Section E.2.

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

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

24

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

The worst-case minimax results are sometimes pessimistic, so we provide a central limit result for the empirical optimizer $\widehat{\theta}_n \in \operatorname{argmin}_{\theta \in \mathbb{R}^d} \mathcal{R}(\theta; \widehat{P}_n)$ to the population optimizer $\theta^\star = \operatorname{argmin}_{\theta \in \mathbb{R}^d} \mathcal{R}(\theta; \widehat{P}_n)$ under appropriate smoothness conditions on the risk. Given that in the general formulation of our problem, the supremum over distributions 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 simpler conditions that allow an essentially classical treatment with a brief proof, based on the dual formulation (4).

Throughout this section, we assume that the population optimizer $\theta^\star = \operatorname{argmin}_{\theta \in \mathbb{R}^d} \mathcal{R}(\theta; \widehat{P}_n)$ is unique. We begin with a smoothness assumption.

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

1. *There exists $L : \mathcal{X} \to \mathbb{R}_+$ such that $|\ell(\theta_0; x) - \ell(\theta_1; x)| \leq L(x) \|\theta_0 - \theta_1\|_2$ for all $\theta_i \in U$, where $\mathbb{E}[L(X)^{2k_*}] < \infty$ (we again use $k_* = \frac{k}{k-1}$).*

2. *$\mathbb{E}[|\ell(\theta^\star; X)|^{2k_*}] < \infty$, and the function $\theta \mapsto \ell(\theta; x)$ is differentiable on U.*

Recalling the dual (4), for shorthand define

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

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

The second condition of Assumption C guarantees $\lambda^\star > 0$. For Cressie-Read divergences (6), a sufficient condition for uniqueness of $(\eta^\star, \lambda^\star)$ follows.

Lemma 2. *Let f be the Cressie-Read divergence (6) with parameter $k \in (1, \infty)$, and $\theta_0 \in \Theta$. If $\ell(\theta_0; X)$ is non-constant under P and $\mathbb{E}_P[|\ell(\theta; X)|^{k_*}] < \infty$ near θ_0, then $(\lambda_0, \eta_0) = \operatorname{argmin}_{\lambda \geq 0, \eta} g_{P_0}(\theta_0, \lambda, \eta)$ is unique.*

See Appendix F.1 for a proof. Sufficient conditions for differentiability are similar to the classical conditions for asymptotic normality of quantile estimators [102]; 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_0} is $\mathcal{C}^2$ in a neighborhood of $(\theta_0, \eta_0, \lambda_0)$ whenever $\lambda_0 > 0$. With this brief discussion, we now provide an asymptotic normality result.

Theorem 11. *Let Assumptions B and C hold. Let $\widehat{\theta}_n$ be any sequence of approximate optimizers to the empirical plug-in satisfying $\mathcal{R}_f(\widehat{\theta}_n; \widehat{P}_n) \leq \inf_\theta \mathcal{R}_f(\theta; \widehat{P}_n) + o_P(1/n)$. Then*

$$\sqrt{n} \left(\widehat{\theta}_n - \theta^\star \right) \overset{d}{\rightsquigarrow} \mathsf{N} \left(0, V \operatorname{Cov} \left(f^{*\prime} \left(\frac{\ell(\theta^\star; X) - \eta^\star}{\lambda^\star} \right) \nabla \ell(\theta^\star; X) \right) V \right) \tag{19}$$

where V is the first d-by-d block of $\left(\nabla^2 g_{P_0}(\theta^\star, \lambda^\star, \eta^\star) \right)^{-1} \in \mathbb{R}^{(d+2) \times (d+2)}$.

See Section F.2 for the proof. Under the same assumptions, it is straightforward to see that plug-in estimators for V and $\mathrm{Cov}(f^{*\prime}(\frac{\ell(\theta^\star;X)-\eta^\star}{\lambda^\star})\nabla\ell(\theta^\star;X))$ are consistent. Combining these estimators with Theorem 11 gives an asymptotically pivotal confidence region for $\theta^\star$ by Slutsky's lemmas.

We can relax the assumption that $\nabla^2 g_{P_0}(\theta^\star,\lambda^\star,\eta^\star)\succ 0$ in Assumption C 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

$$
\begin{aligned}
(k-1)\mathbb{E}B^{k_*-2}&\left(\mathbb{E}B^{k_*}\mathbb{E}B^{k_*-2}-(\mathbb{E}B^{k_*-1})^2\right)\mathbb{E}[B^{k_*-1}\nabla^2\ell(\theta^\star;X)]\\
&-\left(\mathbb{E}B^{k_*-1}\right)^2\mathbb{E}[B^{k_*-2}\nabla\ell(\theta^\star;X)]\mathbb{E}[B^{k_*-2}\nabla\ell(\theta^\star;X)]^\top\succ 0,
\end{aligned}
\tag{20}
$$

and $k\in(1,2)$. For $k=2$, the relaxed condition holds if in addition to the bound (20), there is a neighborhood of $(\theta^\star,\eta^\star)$ such that $\mathbb{P}(\ell(\theta;X)=\eta)=0$. Assumption C also requires identifiability of nuisance variables $\lambda^\star,\eta^\star$. Whether directly analyzing the primal formulation (1)—rather than our proof via the dual (4)—can relax this assumption remains open.

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 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, $\{Q:D_f(Q\|P_0)\le\rho\}$. 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 [72, 36, 15, 62, 64] show that the robustness provides a type of regularization by variance of the loss when f is smooth, no matter what 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. Although we provide heuristics for choosing f and ρ in Section 3, a principled understanding of these adaptive procedures is an important future direction of research.

The minimax guarantees demonstrate 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 match up to rates in n of n^{-1/k_*} (up to logarithmic factors), though not in dimension dependence, so 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 [11] is also a topic of future research. In our asymptotic results (Section 6), we require an identifiability assumption on the dual formulation, and it is open whether this assumption can be relaxed by analyzing the primal problem directly.

The robust formulation (1) and its empirical formulation (2) are complementary to traditional robustness approaches in statistics arising out of Huber's work [54, 55]. In the classical notions of Huber robustness, one wishes to obtain an estimate of a parameter θ of a distribution P_0 contaminated by some 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 outcome Y. The distributionally robust formulation (1) considers shifts in the joint distribution $(X, Y) \sim P_0$. Traditional domain adaptation 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). In causal data analyses, one wishes to perturb only the distribution of the covariates X, observing the effect of such interventions on Y. Connecting these ideas and developing variants of the formulation (1) that only protect against covariate shift or structural shifts on X may be useful in many scenarios.

Acknowledgments

References

[1] M. Abramowitz and I. Stegun, editors. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables.* Dover, 1965.

[2] 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.

[3] A. Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012.

[4] 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.

[5] 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.

[6] 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.

[7] T. W. Anderson. The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society*, 6 (2):170–176, 1955.

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

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

[10] R. Atar, K. Chowdhary, and P. Dupuis. Robust bounds on risk-sensitive functionals via Rényi divergence. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):18–33, 2015.

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

[12] 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.

[13] 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.

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

[15] 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.

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

[17] 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.

[18] 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.

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

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

[21] 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.

[22] 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.

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

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

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

[26] 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.

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

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

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

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

[31] H. Daume III and D. Marcu. Domain adaptation for statistical classifiers. *Journal of artificial*

Intelligence research, 26:101–126, 2006.

[32] 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.

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

[34] 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.

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

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

[37] P. Dupuis, M. A. Katsoulakis, Y. Pantazis, and P. Plechác. Path-space information bounds for uncertainty quantification and sensitivity analysis of stochastic dynamics. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):80–111, 2016.

[38] 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.

[39] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming, Series A*, 171(1–2):115–166, 2018.

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

[41] 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.

[42] R. Gao and A. J. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *arXiv:1604.02199 [math.OC]*, 2016.

[43] R. J. Gardner. The Brunn-Minkowski inequality. *Bulletin of the American Mathematical Society*, 39(3):355–405, 2002.

[44] W. Gautschi. *The Incomplete Gamma Functions since Tricomi*, volume 147 of *Atti dei Convegni Lincei*. Accademia Nazionale dei Lincei, 1997.

[45] S. Ghosh and H. Lam. Robust analysis in stochastic simulation: Computation and performance guarantees. *Operations Research*, 2019.

[46] P. Glasserman and X. Xu. Robust risk measurement and model risk. *Quantitative Finance*, 14(1):29–58, 2013.

[47] J.-y. Gotoh, M. J. Kim, and A. Lim. Robust empirical optimization is almost the same as mean-variance optimization. *Available at SSRN 2827400*, 2015.

[48] P. J. Grother, G. W. Quinn, and P. J. Phillips. Report on the evaluation of 2d still-image face recognition algorithms. *NIST Interagency/Internal Reports (NISTIR)*, 7709, 2010.

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

[50] L. P. Hansen and T. J. Sargent. *Robustness*. Princeton University Press, 2008.

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

[52] 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.

[53] 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.

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

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

[56] 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.

[57] R. Jiang and Y. Guan. Data-driven chance constrained stochastic program. *Mathematical Programming*, 158(1-2):291–327, 2016.

[58] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, page 1, 2011.

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

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

[61] S. Kusuoka. On law invariant coherent risk measures. In *Advances in Mathematical Economics*, pages 83–95. Springer, 2001.

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

[63] H. Lam. Sensitivity to serial dependency of input processes: A robust approach. *Management Science*, 64(3):1311–1327, 2017.

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

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

[66] 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.

[67] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

[68] J. Lee and M. Raginsky. Minimax statistical learning and domain adaptation with Wasserstein distances. *arXiv:1705.07815 [cs.LG]*, 2017.

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

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

[71] 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.

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

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

[74] I. R. Petersen, M. R. James, and P. Dupuis. Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Transactions on Automatic Control*, 45(3): 398–412, 2000.

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

[76] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[77] 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.

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

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

[80] 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.

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

[82] 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.

[83] 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.

[84] 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.

[85] M. Shaked and J. G. Shanthikumar. *Stochastic Orders*. Springer Series in Statistics. Springer, 2007.

[86] A. Shapiro. On Kusuoka representation of law invariant risk measures. *Mathematics of Operations Research*, 38(1):142–152, 2013.

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

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

[89] 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.

[90] M. Simon and M.-S. Alouini. *Digital Communication Over Fading Channels: A Unified Approach to Performance Analysis*. John Wiley & Sons, New York, 2000.

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

[92] 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.

[93] 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.

[94] Y. Sun, Árpád Baricz, and S. Zhou. On the monotonicity, log-concavity and tight bounds of the generalized Marcum and Nuttall Q-functions. *IEEE Transactions on Information Theory*, 56(3):1166–1186, 2010.

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

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

[97] 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.

[98] 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.

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

[100] 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.

[101] 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.

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

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

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

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

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

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

[108] 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 Lemma 1

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

Lemma 3.

$$f_k^*(s) = \frac{1}{k} \left((k-1)s + 1 \right)_+^{k_*} - \frac{1}{k} \tag{21}$$

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 $t = ((k-1)s + 1)^{1/(k-1)}$, 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 $1 - 1/k = 1/k_*$. $\qquad\square$

From the dual formulation (4), we have

$$\sup_{P \ll P_0} \left\{ \mathbb{E}_P[Z] \text{ s.t. } D_f(P\|P_0) \leq \rho \right\} = \inf_{\lambda \geq 0, \eta} \left\{ \lambda \mathbb{E}_{P_0} f^* \left(\frac{Z - \eta}{\lambda} \right) + \lambda \rho + \eta \right\}$$

$$= \inf_{\lambda \geq 0, \eta} \left\{ \frac{(k-1)^{k_*}}{k} \lambda^{1-k_*} \mathbb{E}_{P_0} \left(Z - \eta + \frac{\lambda}{k-1} \right)_+^{k_*} + \lambda(\rho - \frac{1}{k}) + \eta \right\}$$

$$= \inf_{\lambda \geq 0, \tilde{\eta}} \left\{ (k-1)^{k_*} k^{-1} \mathbb{E}_{P_0} (Z - \tilde{\eta})_+^{k_*} \lambda^{1-k_*} + \left(\rho + \frac{1}{k(k-1)} \right) \lambda + \tilde{\eta} \right\}$$

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 $(k_* - 1)/k_* = 1/k$)

$$\lambda = (k-1)(k(k-1)\rho + 1)^{-\frac{1}{k_*}} \left(\mathbb{E}_{P_0} (Z - \tilde{\eta})_+^{k_*} \right)^{\frac{1}{k_*}} \tag{22}$$

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

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

A.2 Moments and duality

We discuss the norm-like behavior of the robust risk $\mathcal{R}_f$ when f behaves asymptotically as t^k for some $k \in (1, \infty)$; we treat c_j and C_j as constants whose values may change from line to line. For simplicity we assume that f is differentiable, though subdifferential calculus [51] allows immediate extension to the non-differentiable case. Assume that $0 < \liminf_{t \to \infty} f(t)/t^k \leq \limsup_{t \to \infty} f(t)/t^k < \infty$. Then as $t \mapsto f'(t)$ is non-decreasing, there exist $0 < c_0 \leq c_1 < \infty$ such

that $c_0 t^{k-1} \leq f'(t) \leq c_1 t^{k-1}$ for all large enough t. Then for all large s the t solving $f'(t) = s$ satisfies $c_0 t^{k-1} \leq s \leq c_1 t^{k-1}$, that is, $(s/c_1)^{\frac{1}{k-1}} \leq (f')^{-1}(s) \leq (s/c_0)^{\frac{1}{k-1}}$. Recall that the conjugate $f^*(s) := \sup_t \{st - f(t)\}$ satisfies the duality $(f')^{-1}(s) = (f^*)'(s)$, and as $\operatorname{dom} f \subset \mathbb{R}_+$, f^* is non-decreasing [51]. Then there are constants C_0, C_1 such that for all large s, we evidently have

$$C_0 s^{\frac{1}{k-1}} \leq (f^*)'(s) \leq C_1 s^{\frac{1}{k-1}},$$

and so by an integration argument

$$C_0 s^{k_*} \leq f^*(s) \leq C_1 s^{k_*} \quad \text{for large } s.$$

In particular, for some threshold τ_f depending on f, if we define the shorthand $Z = \ell(\theta; X)$ then the dual (4) satisfies

$$
\begin{aligned}
&\inf_{\lambda \geq 0, \eta \in \mathbb{R}} \left\{ C_0 \lambda^{1-k_*} \mathbb{E}\left[(Z - \eta)_*^k \mathbf{1}\left\{ Z \geq \lambda \tau_f \right\} \right] + \lambda \mathbb{E}\left[f^*\left(\frac{Z - \eta}{\lambda} \right) \mathbf{1}\left\{ Z < \lambda \tau_f \right\} \right] + \eta + \lambda \rho \right\} \\
&\leq \mathcal{R}_f(\theta; P_0) \\
&\leq \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \left\{ C_1 \lambda^{1-k_*} \mathbb{E}\left[(Z - \eta)_*^k \mathbf{1}\left\{ Z \geq \lambda \tau_f \right\} \right] + \lambda \mathbb{E}\left[f^*\left(\frac{Z - \eta}{\lambda} \right) \mathbf{1}\left\{ Z < \lambda \tau_f \right\} \right] + \eta + \lambda \rho \right\}.
\end{aligned}
$$

That $\inf_{\lambda \geq 0}\{\lambda \rho + C\lambda^{1-k_*}\} = (k_* - 1)^{1/k_*}(1 + \frac{1}{k_*})\rho^{1/k} C^{1/k_*}$ shows that once again, we have the dependence of $\mathcal{R}_f$ on k_*th moments of the loss.

B Proofs of Examples

B.1 Proof of Example 4

That $\theta_1 = \sum_v p_v \theta_v$ is immediate. For the second claim, we begin with a characterization of the Chebyshev center and a few of its properties. We have

$$\theta_{\text{minimax}} = \sum_v s_v \theta_v \quad \text{where} \quad \mathbf{1}^T s = 1, s \succeq 0,$$

and by the KKT conditions for optimality, $s_v > 0$ only if $\max_{w \in V} \|\theta_{\text{minimax}} - \theta_w\|_2 = \|\theta_v - \theta_w\|_2$. We recall that a function h is c-strongly if $\langle \nabla h(\theta) - \nabla h(\tau), \theta - \tau \rangle \geq c \|\theta - \tau\|_2^2$, and θ^* minimizes h over Θ if and only if for some $g \in \partial h(\theta^*)$ we have $\langle g, \theta - \theta^* \rangle \geq 0$ for all $\theta \in \Theta$. Let $h(\theta) = \frac{1}{2} \max_{v \in V} \|\theta - \theta_v\|_2^2$. Then letting $\theta \neq \theta_{\text{minimax}}$, we immediately see that for any $v \in V$ for which $\|\theta - \theta_v\|_2 = \max_{w \in V} \|\theta - \theta_w\|_2$, we have for $g = \theta - \theta_v \in \partial h(\theta)$ and $g' \in \partial h(\theta_{\text{minimax}})$ that

$$
\begin{aligned}
\langle \theta - \theta_v, \theta - \theta_{\text{minimax}} \rangle &= \langle g, \theta - \theta_{\text{minimax}} \rangle \\
&\geq \langle g', \theta - \theta_{\text{minimax}} \rangle + \|\theta - \theta_{\text{minimax}}\|_2^2 \geq \|\theta - \theta_{\text{minimax}}\|_2^2.
\end{aligned}
\tag{23}
$$

Fix $\theta \in \mathbb{R}^d$ with $\theta \neq \theta_{\text{minimax}}$. For fixed $\eta > 0$, consider the objectives $h_v(\cdot; \eta) : \mathbb{R} \to \mathbb{R}$,

$$h_v(t; \eta) := \mathbb{E}\left[\left((1/2)\|(1-t)\theta + t\theta_{\text{minimax}} - \theta_v - \varepsilon\|_2^2 - \eta \right)_+ \right].$$

By the continuity of the density of ε, h_v is differentiable in t, and we have

$$h_v'(t; \eta) = \mathbb{E}\left[\mathbf{1}\left\{ \|(1-t)\theta + t\theta_{\text{minimax}} - \theta_v - \varepsilon\|_2^2 \geq 2\eta \right\} \langle \theta_{\text{minimax}} - \theta, (1-t)\theta + t\theta_{\text{minimax}} - \theta_v - \varepsilon \rangle \right]$$

and
$$h'_v(0; \eta) = \mathbb{E}\left[\mathbf{1}\left\{\|\theta - \theta_v - \varepsilon\|_2^2 \geq 2\eta\right\}\langle\theta_{\text{minimax}} - \theta, \theta - \theta_v - \varepsilon\rangle\right]. \tag{24}$$

For any vector $\mu \in \mathbb{R}^d$, define the constant $c(\eta; \mu) > 0$ such that

$$\mathbb{E}[\mathbf{1}\left\{\|\mu + \varepsilon\|_2^2 \geq 2\eta\right\}(\mu + \varepsilon)] = c(\eta; \mu)\mu,$$

which must exist by the rotational symmetry of the Gaussian. Now we claim that for any μ_1, μ_2 with $\|\mu_1\|_2 > \|\mu_2\|_2$,

$$\lim_{\eta \to \infty} \frac{c(\eta; \mu_1)}{c(\eta; \mu_2)} = \infty. \tag{25}$$

Deferring the proof of the claim (25), let us see how it yields the theorem.

We shall show that for $\theta \neq \theta_{\text{minimax}}$, if we define

$$R(t; \eta) := \sum_v p_v \mathbb{E}\left[\left((1/2)\|(1-t)\theta + t\theta_{\text{minimax}} - \theta_v - \varepsilon\|_2^2 - \eta\right)_+\right],$$

then for all large η, we have $R'(0; \eta) < 0$, so θ cannot minimize $\sum_v p_v\mathbb{E}[((1/2)\|\theta - \theta_v - \varepsilon\|_2^2 - \eta)_+]$. That this gives the theorem is nearly immediate, because for all $0 < \alpha \leq 1$, the η minimizing the CVaR risk $\alpha^{-1}\mathbb{E}[(\ell(\theta; Y) - \eta)_+] + \eta$ is the $1 - \alpha$ quantile of $\ell(\theta; Y)$, which for our setting evidently tends to ∞ as $\alpha \downarrow 0$. Thus, if $\eta(\theta, \alpha) = \operatorname{argmin}_\eta\{\alpha^{-1}\mathbb{E}[(\ell(\theta; Y) - \eta)_+] + \eta\}$, we have $\eta(\theta, \alpha) \to \infty$ uniformly in θ as $\alpha \downarrow 0$, and so $R'(0; \eta) < 0$ implies that θ cannot minimize $\text{CVaR}_\alpha(\ell(\theta; Y))$. To see that $R'(0; \eta) < 0$, simply note that

$$R'(0; \eta) = \sum_v p_v c(\eta; \|\theta - \theta_v\|_2)\langle\theta_{\text{minimax}} - \theta, \theta - \theta_v\rangle$$

by Eq. (24). Let $V^\star = \{v : \|\theta - \theta_v\|_2 = \max_w \|\theta - \theta_w\|_2\}$. Then inequality (23) implies that

$$R'(0; \eta) \leq -\left(\sum_{v \in V^\star} p_v c(\eta; \|\theta - \theta_v\|_2)\right)\|\theta - \theta_{\text{minimax}}\|_2^2$$
$$+ \sum_{v \notin V^\star} p_v c(\eta; \|\theta - \theta_v\|_2)c(\eta; \|\theta - \theta_v\|_2)\langle\theta_{\text{minimax}} - \theta, \theta - \theta_v\rangle.$$

In particular, for any $v \in V^\star$ we have

$$\frac{1}{c(\eta; \|\theta - \theta_v\|_2)}R'(0; \eta) \leq -\left(\sum_{v \in V^\star} p_v\right)\|\theta - \theta_{\text{minimax}}\|_2^2 + o(1)$$

as $\eta \to \infty$, giving the theorem.

Finally, we return to prove the claim (25), note that $\|\mu + \varepsilon\|_2^2$ follows a non-central χ^2 distribution with $\mathbb{P}(\|\mu + \varepsilon\|_2^2 \geq t) = Q_{d/2}(\|\mu\|_2, \sqrt{t})$ for Q the Marcum Q-function. Letting Φ be the standard Gaussian CDF, the Marcum Q-function satisfies the asymptotics (e.g. [90, p. 81] or [94, Eq. (4)]) that as $t \to \infty$,

$$Q_k(a, t) = (1 + o(1))\left(\frac{t}{a}\right)^{k-1/2}(1 - \Phi(t - a)) = (1 + o(1))\frac{1}{t-a}\exp\left(-\frac{(t-a)^2}{2}\right)\left(\frac{t}{a}\right)^{k-1/2}.$$

Thus we obtain

$$\frac{\mathbb{P}(\|\mu_1 + \varepsilon\|_2^2 \geq t^2)}{\mathbb{P}(\|\mu_2 + \varepsilon\|_2^2 \geq t^2)} = (1 + o(1)) \left(\frac{\|\mu_2\|_2}{\|\mu_1\|_2} \right)^{\frac{d-1}{2}} \exp\left(-\frac{(t - \|\mu_1\|_2)^2}{2} + \frac{(t - \|\mu_2\|_2)^2}{2} \right)$$

$$= (1 + o(1)) \left(\frac{\|\mu_2\|_2}{\|\mu_1\|_2} \right)^{\frac{d-1}{2}} \exp\left(t(\|\mu_1\|_2 - \|\mu_2\|_2) + \frac{1}{2}(\|\mu_2\|_2^2 - \|\mu_1\|_2^2) \right) \to \infty$$

as $t \to \infty$, giving the claim (25).

B.2 Proof of Example 5

That $\theta_{\mathrm{ols}} = \theta_1 = \int \theta_v d\mu(v)$ is immediate.

We begin with a technical lemma on the expectations of Gaussian variables whose proof we defer to Sec. B.2.1.

Lemma 4. *Let $Z \sim \mathsf{N}(0,1)$. Then*

$$\mathbb{E}\left[(Z^2 - t)_+ \right] = \sqrt{\frac{2}{\pi}} \left(t - 1 + O(t^{-1}) \right) e^{-\frac{1}{2}t^2}.$$

Defining the shorthand $\tau_v^2(\theta) := (\theta - \theta_v)^T \Sigma_v (\theta - \theta_v) + \sigma_v^2$, for $Z \sim \mathsf{N}(0,1)$ we evidently have

$$\mathrm{CVaR}_\alpha(\ell(\theta; X, Y)) = \inf_\eta \left\{ \frac{1}{2\alpha} \int \mathbb{E}\left[(\tau_v^2(\theta) Z^2 - 2\eta)_+ \right] d\mu(v) + \eta \right\}. \tag{26}$$

For large η, Lemma 4 gives

$$\mathbb{E}\left[(\tau_v^2(\theta) Z^2 - 2\eta)_+ \right] = \tau_v^2(\theta) \mathbb{E}\left[(Z^2 - 2\eta/\tau_v^2(\theta))_+ \right]$$

$$= C(1 + O(\tau_v^2(\theta)/\eta))\eta \exp\left(-\frac{2\eta^2}{\tau_v^4(\theta)} \right) \tag{27}$$

uniformly in $\tau_v^2(\theta)$, where $C = 2\sqrt{2/\pi}$ is a fixed constant.

We now compute normalized asymptotics of the mixture CVaR (26). For a measure μ on V and measurable $g : V \to \mathbb{R}$, we define the quantile $\mathsf{quant}_p(g, \mu) := \inf\{t \in \mathbb{R} : p \leq \mu(g^{-1}((-\infty, t]))\}$, which gives the following.

Lemma 5. *Let $\epsilon > 0$ and let $\mathrm{ess\,sup}_v\, g(v) = \inf\{t \mid \mu(\{v : g(v) > t\}) = 0\}$ be the essential supremum of g. Then for all $t > 0$,*

$$t\,\mathsf{quant}_{1-\epsilon}(g, \mu) - \log\frac{1}{\epsilon} \leq \log\left(\int e^{tg(v)} d\mu(v) \right) \leq t \,\mathrm{ess\,sup}_{v \in V}\, g(v).$$

Proof Clearly $\log \int e^{tg(v)} d\mu(v) \leq t \,\mathrm{ess\,sup}_{v \in V}\, g(v)$. For the lower bound, letting $q = \mathsf{quant}_{1-\epsilon}(g, \mu)$, we have $\log \int e^{tg(v)} d\mu(v) \geq \log \int e^{tq} \mathbf{1}\{g(v) \geq q\} d\mu(v) \geq \log \epsilon + tq$. $\qquad \square$

For $\theta \in \mathbb{R}^d$ and $\eta > 0$ define the normalized logarithmic risk

$$R(\theta; \eta) := \frac{1}{2\eta^2} \log \int \mathbb{E}_v\left[(\tau_v^2(\theta) Z^2 - 2\eta)_+ \right] d\mu(v),$$

36

which by Lemma 5 satisfies

$$\mathsf{quant}_{1-\epsilon}\left(-\frac{1}{\tau_v^4(\theta)},\mu\right) - \frac{\log\frac{1}{\epsilon}}{2\eta^2} + \frac{\log\left(C(1+O(1/\eta))\eta\right)}{2\eta^2}$$

$$\leq R(\theta;\eta) \leq \sup_v\left\{-\frac{1}{\tau_v^4(\theta)}\right\} + \frac{\log\left(C(1+O(1/\eta))\eta\right)}{2\eta^2},$$

where we have used the boundedness assumptions on Σ_v, σ_v, and θ_v.

Assume now that $\theta \neq \theta_{\mathrm{minimax}}$. Then $\sup_v \tau_v(\theta) > \sup_v \tau_v(\theta_{\mathrm{minimax}})$ because $\Sigma_v \succ 0$ and so $\theta_{\mathrm{minimax}}$ must be unique. By the assumption that the essential suprema and suprema over v are equal, $\mathsf{quant}_{1-\epsilon}(-1/\tau_v^4(\theta)) \to \sup_v\{-1/\tau_v^4(\theta)\}$ as $\epsilon \to 0$, for all large enough η we have

$$R(\theta_{\mathrm{minimax}};\eta) < R(\theta;\eta).$$

Notably, if $\theta_\eta \in \operatorname{argmin}_\theta R(\theta;\eta)$, then evidently $\theta_\eta \to \theta_{\mathrm{minimax}}$.

Finally, we consider the quantity

$$\eta_\alpha(\theta) := \operatorname*{argmin}_\eta\left\{\frac{1}{\alpha}\int \mathbb{E}_v\left[(\ell(\theta;X,Y)-\eta)_+\right]d\mu(v) + \eta\right\}.$$

By definition, we have

$$\operatorname*{argmin}_\theta \mathrm{CVaR}_\alpha(\ell(\theta;X,Y)) = \operatorname*{argmin}_\theta \int \mathbb{E}_v[(\ell(\theta;X,Y)-\eta_\alpha(\theta))_+]d\mu(v),$$

and moreover, $\eta_\alpha(\theta) = \mathsf{quant}_{1-\alpha}(\ell(\theta;X,Y))$, where the quantile is computed jointly over $v \sim \mu$ and (X,Y). As by assumption $\inf_v \sigma_v^2 > 0$, it is evident that $\liminf_{\alpha\downarrow 0}\inf_\theta \eta_\alpha(\theta) = \infty$. In particular, for all small enough $\alpha > 0$, we obtain $R(\theta_{\mathrm{minimax}}, \eta_\alpha(\theta)) < R(\theta, \eta_\alpha(\theta))$, or

$$\int \mathbb{E}_v\left[(\ell(\theta_{\mathrm{minimax}};X,Y)-\eta_\alpha(\theta))_+\right] < \int \mathbb{E}_v\left[(\ell(\theta;X,Y)-\eta_\alpha(\theta))_+\right],$$

giving the result we claim in Example 5.

B.2.1 Proof of Lemma 4

Let $\Phi(t) = \mathbb{P}(Z \leq t)$ be the standard normal CDF and $\Gamma(a,x) = \int_x^\infty z^{a-1}e^{-z}dz$ be the incomplete Gamma function. For $t \geq 0$ we have

$$\mathbb{E}[(Z^2-t)_+] = \sqrt{\frac{2}{\pi}}\int_t^\infty (z^2-t)e^{-\frac{1}{2}z^2}dz$$

$$\overset{(i)}{=} \frac{2}{\sqrt{\pi}}\int_{t^2/2}^\infty u^{1/2}e^{-u}du - 2t(1-\Phi(t))$$

$$= \frac{2}{\sqrt{\pi}}\Gamma\left(\frac{3}{2},\frac{t^2}{2}\right) - 2t(1-\Phi(t))$$

where inequality (i) is via the substitution $u = z^2/2$. Now we use asymptotics of the normal CDF and incomplete Gamma function to approximate the preceding display for large t. By standard normal approximations [1, Eq. (7.1.13)], we have for $t \geq 0$ that

$$\frac{t}{t^2+1}e^{-\frac{1}{2}t^2} \leq \frac{2}{t+\sqrt{t^2+4}}e^{-\frac{1}{2}t^2} \leq \sqrt{2\pi}(1-\Phi(t)) \leq \frac{2}{t+\sqrt{t^2+8/\pi}}e^{-\frac{1}{2}t^2} \leq \frac{1}{t}e^{-\frac{1}{2}t^2}.$$

37

As $\sqrt{t^2 + c} = t + c/2t + O(t^{-2})$ for any constant c, we have

$$(1 - \Phi(t)) = \frac{1}{\sqrt{2\pi}} \left(t^{-1} - O(1)t^{-3} \right) e^{-\frac{1}{2}t^2},$$

while we also have the well-known asymptotic expansion

$$\Gamma(a + 1, x) = \frac{e^{-x} x^{a+1}}{x - a} \left[1 - \frac{a}{(x - a)^2} + \frac{2a}{(x - a)^3} + O\left(\frac{a^2}{(x - a)^4} \right) \right]$$

as $\sqrt{a}/(x - a) \to 0$ [44, Eq. (2.12)].

Substituting these above (with $x = t^2/2$ and $a = 1/2$) yields for large t that

$$\begin{aligned}
\sqrt{\pi/2} e^{\frac{1}{2}t^2} \mathbb{E}\left[(Z^2 - t)_+ \right] &= \frac{t^3}{t^2 - 1} \left[1 - \frac{2}{(t^2 - 1)^2} + O(t^{-6}) \right] - t\left(\frac{1}{t} - \frac{O(1)}{t^3} \right) \\
&= t - 1 + O(t^{-1}),
\end{aligned}$$

giving the lemma.

C   Proof of Upper Bounds

C.1   Proof of Theorem 2

To ease notation, for any fixed $\theta \in \Theta$, let $Z(x) = \ell(\theta; x)$ and

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

so that $\mathcal{R}_k(Z; P) = \inf_\eta g_k(\eta; P)$ from Proposition 1. We begin by showing pointwise concentration of $g_k(\eta; \widehat{P}_n)$ to $g_k(\eta; P_0)$ for each bounded η. First, we begin by recalling a standard convex Lipschitz concentration inequality for bounded random variables.

Lemma 6 (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 $t \geq 0$,*

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

To apply Lemma 6, we verify that $g_k(\eta; \widehat{P}_n)$ is Lipschitz in the data vector Z_1^n by using the following elementary result.

Lemma 7. *The map $\mathbb{R}^n \ni y \mapsto \left(\frac{1}{n} \sum_{i=1}^n |y_i|^{k_*} \right)^{\frac{1}{k_*}}$ is $n^{-\frac{1}{2 \vee k_*}}$-Lipschitz with respect to the $\|\cdot\|_2$-norm.*

Proof of Lemma   We denote $\|Y\|_{L^p(\widehat{P}_n)} = \left(\frac{1}{n} \sum_{i=1}^n |y_i|^p \right)^{\frac{1}{p}}$ to ease notation. Noting that

$$\|\psi_n(y)\|_2 = n^{-\frac{1}{2}} \left(\frac{\|Y\|_{L^{2(k_*-1)}(\widehat{P}_n)}}{\|Y\|_{L^{k_*}(\widehat{P}_n)}} \right)^{k_*-1},$$

38

we proceed in two cases. If $k_* \leq 2$, the result follows from $\|Y\|_{L^{2(k_*-1)}(\widehat{P}_n)} \leq \|Y\|_{L^{k_*}(\widehat{P}_n)}$. If $k_* \geq 2$, $\left(\sum_{i=1}^n |y_i|^{2(k_*-1)}\right)^{\frac{1}{2(k_*-1)}} \leq \left(\sum_{i=1}^n |y_i|^{k_*}\right)^{\frac{1}{k_*}}$ implies

$$\|Y\|_{L^{2(k_*-1)}(\widehat{P}_n)} \leq n^{-\frac{1}{k_*}+\frac{1}{2(k_*-1)}} \|Y\|_{L^{k_*}(\widehat{P}_n)},$$

which gives the result. $\qquad\square$

Lemma 7 implies $g_k(\eta; \widehat{P}_n)$ is a $c_k n^{-\frac{1}{2\vee k_*}}$-Lipschitz function of the data vector Z_1^n with respect to the $\|\cdot\|_2$-norm. Applying Lemma 6, for any fixed $\eta \in [-\frac{1}{c_k-1}M, M]$

$$|g_k(\eta; \widehat{P}_n) - \mathbb{E}_{P_0}[g_k(\eta; \widehat{P}_n)]| \leq \sqrt{2t}c_k \left(\frac{c_k}{c_k-1} \vee 2\right) Mn^{-\frac{1}{k_*\vee 2}} \tag{28}$$

with probability at least $1 - 2e^{-t}$.

To establish pointwise concentration of $g_k(\eta; \widehat{P}_n)$ to $g_k(\eta; P_0)$, it remains to see that $\mathbb{E}_{P_0}[g_k(\eta; \widehat{P}_n)]$ and $g_k(\eta; P_0)$ are close. We use the following lemma, whose proof we defer to Section C.1.1.

Lemma 8. *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*

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n |Y_i|^{k_*}\right)^{\frac{1}{k_*}}\right] \geq \mathbb{E}[|Y|^{k_*}]^{\frac{1}{k_*}} - \frac{2}{k}\sqrt{C}n^{-\frac{1}{k_*\vee 2}} \tag{29}$$

Since $\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n |Y_i|^{k_*}\right)^{\frac{1}{k_*}}\right] \leq \mathbb{E}[|Y|^{k_*}]^{\frac{1}{k_*}}$ by Jensen's inequality, Lemma 8 implies

$$\left|\mathbb{E}_{P_0}[g_k(\eta; \widehat{P}_n)] - g_k(\eta; P_0)\right| \leq \frac{2c_k}{k}\sqrt{\left(\frac{c_k}{c_k-1} \vee 2\right) M}n^{-\frac{1}{k_*\vee 2}}$$

for any fixed $\eta \in [-\frac{1}{c_k-1}M, M]$. Combining the bound with the concentration result (28), we conclude that with probability at least $1 - 2e^{-2t}$,

$$|g_k(\eta; \widehat{P}_n) - g_k(\eta; P_0)| \leq n^{-\frac{1}{k_*\vee 2}}Mc_k\left(\frac{c_k}{c_k-1} \vee 2\right)\left(\sqrt{2t} + \frac{2}{k}\right) =: \epsilon_t. \tag{30}$$

We now show uniform concentration by using a simple covering argument. The following lemma restricts the domain of η to a compact set, which is essential to this argument.

Lemma 9. *If $Z \in [0, M]$, then for any distribution P*

$$\inf_{\eta \in \mathbb{R}} g(\eta; P) = \inf_{\eta} \left\{g(\eta; P) : \eta \in \left[-\frac{1}{c_k-1}M, M\right]\right\}.$$

Proof of Lemma By definition, $g(\eta; P) = \eta$ for $\eta \geq M$, and

$$g\left(-\frac{1}{c_k-1}M; P\right) \geq c_k\frac{M}{c_k-1} - \frac{M}{c_k-1} = M = g(M; P).$$

39

Since $\eta \mapsto g(\eta; P)$ is convex, this implies the result. $\qquad\square$

Recalling the shorthand $\epsilon_{t,n} := n^{-\frac{1}{k_* \vee 2}} M c_k \left(\frac{c_k}{c_k - 1} \vee 2 \right) \left(\sqrt{2t} + \frac{2}{k} \right)$, define the sequence

$$\eta_i := -(c_k - 1)^{-1} M + i\epsilon_{t,n}$$

for nonnegative integers $i \leq \frac{c_k}{c_k - 1} \frac{M}{\epsilon_{t,n}}$. Then, for any $\eta \in [-(c_k - 1)^{-1} M, M]$, there exists $1 \leq i(\eta) \leq \frac{c_k}{c_k - 1} \frac{M}{\epsilon_{t,n}}$ such that $|\eta - \eta_{i(\eta)}| \leq \epsilon_{t,n}$.

$$\sup_{\eta \in [-(c_k-1)^{-1}M, M]} |g(\eta; \widehat{P}_n) - g(\eta; P_0)|$$

$$\leq \sup_{\eta \in [-(c_k-1)^{-1}M, M]} \left\{ |g(\eta; \widehat{P}_n) - g(\eta_{i(\eta)}; \widehat{P}_n)| + |g(\eta_{i(\eta)}; \widehat{P}_n) - g(\eta_{i(\eta)}; P_0)| + |g(\eta_{i(\eta)}; P_0) - g(\eta; P_0)| \right\}$$

$$\leq \max_{1 \leq i \leq \frac{c_k}{c_k-1} \frac{M}{\epsilon_{t,n}}} |g(\eta_{i(\eta)}; \widehat{P}_n) - g(\eta_{i(\eta)}; P_0)| + 2(1 + c_k)\epsilon_{t,n}$$

where we used $(1+c_k)$-Lipschitzness of $\eta \mapsto g(\eta; P_0)$ and $\eta \mapsto g(\eta; \widehat{P}_n)$ in the last inequality. Taking the union bound over the pointwise concentration result (30) with $\eta = \eta_i$, conclude from Lemma 9

$$\left| \mathcal{R}_k(Z; \widehat{P}_n) - \mathcal{R}_k(Z; P_0) \right| = \left| \inf_\eta g_k(\eta; \widehat{P}_n) - \inf_\eta g_k(\eta; P_0) \right|$$

$$= \left| \inf_{\eta \in [-(c_k-1)^{-1}M, M]} g_k(\eta; \widehat{P}_n) - \inf_{\eta \in [-(c_k-1)^{-1}M, M]} g_k(\eta; P_0) \right|$$

$$\leq \sup_{\eta \in [-(c_k-1)^{-1}M, M]} |g(\eta; \widehat{P}_n) - g(\eta; P_0)|$$

$$\leq (2c_k + 3)\epsilon_{t,n}$$

with probability at least $1 - 2\exp\left(-t + \log \frac{c_k}{c_k - 1} \frac{M}{\epsilon_{t,n}} \right)$. Doing a change of variables $t(s) = s + \left(\frac{1}{k_* \vee 2} + 1 \right) \log n$, we obtain the final result.

C.1.1   Proof of Lemma 8

First, we claim that it suffices to show

$$\mathbb{E}[|Y|^q]^{1/q} \geq \mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^n |Y_i|^q \right)^{1/q} \right]$$

$$\geq \mathbb{E}[|Y|^q]^{1/q} - 2\frac{q-1}{q} \begin{cases} (C^{q/2} \vee 1) \cdot n^{-1/q} & \text{if } q \geq 2 \\ (C \vee C^{1-q/2}) \cdot n^{-1/2} & \text{if } q < 2. \end{cases} \tag{31}$$

where the last inequality holds for $n \geq C^q$ when $q \geq 2$. To see how our desired bound (29) follows from (31), we use a quick scaling argument. Let $\alpha > 0$, and note that $\mathbb{E}[|\alpha Y|^{2q}] \leq (C\alpha^2)^q \mathbb{E}[|Y|^q]$ 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 $q \geq 2$, we have $(\alpha^2 C)^{q/2} \geq 1$ if $\alpha \geq C^{-\frac{1}{2}}$, and we obtain

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

40

Choosing $\alpha = C^{-\frac{1}{2}}$ gives the result (29) when $q \geq 2$. For $q < 2$, we similarly obtain that $C\alpha^2 \geq (C\alpha^2)^{1-q/2}$ for $\alpha \geq C^{-\frac{1}{2}}$, whence we have the lower bound

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

for $\alpha \geq C^{-\frac{1}{2}}$. Choosing $\alpha = C^{-\frac{1}{2}}$ thus gives the desired result (29).

Now, we proceed to show the bound (31). Let

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

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

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

$$\geq \frac{\frac{1}{n}\sum_{i=1}^{n}|Y_i|^q}{q\gamma^{q-1}} + \frac{q-1}{q}\gamma + \left(\frac{q-1}{q}-\frac{(q-1)\frac{1}{n}\sum_{i=1}^{n}|Y_i|^q}{q\gamma^q}\right)(\gamma_n-\gamma). \qquad (32)$$

We now show how to provide a bound on magnitude of the final term in expression (32).

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

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

$$= \left(\frac{q-1}{q}\right)^2\left[(1-\sigma^q/\gamma^q)^2 + \frac{1}{\gamma^{2q}n}\operatorname{Var}(|Y|^q)\right],$$

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

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

and

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

Now we provide an upper bound on the $(\gamma_n - \gamma)$ term in the product in inequality (32). By inspection, we have

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

$$\leq \left(\frac{1}{n}\sum_{i=1}^{n}|Y_i|^q\right)^{\frac{2}{q}} + \max\left\{n^{-\alpha},\sigma^q\right\}^{\frac{2}{q}}. \qquad (33)$$

41

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

Lemma 10. *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 10, 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

$$\mathbb{E}[X^2] \leq C^{2-q}\sigma^q. \tag{34}$$

To see this, substitute $a = 2/q \in [1, 2]$ in Lemma 10, 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 (34) that

$$\mathbb{E}[(\gamma_n - \gamma)^2] \leq \mathbb{E}[|Y|^2] + \max\{n^{-2\alpha/q}, \sigma^2\}$$

$$\leq \begin{cases} \sigma^2 + \max\{n^{-2\alpha/q}, \sigma^2\} & \text{if } q \geq 2 \\ C^{2-q}\sigma^q + \max\{n^{-2\alpha/q}, \sigma^2\} & \text{if } q < 2 \end{cases}$$

$$\leq 2 \begin{cases} \max\{n^{-2\alpha/q}, \sigma^2\} & \text{if } q \geq 2 \\ \max\{C^{2-q}\sigma^q, n^{-2\alpha/q}\} & \text{if } q < 2, \end{cases}$$

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

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

In particular, we have by Hölder's inequality that

$$\mathbb{E}\left[\left(1 - \frac{\frac{1}{n}\sum_{i=1}^n |Y_i|^q}{\gamma^q}\right)(\gamma_n - \gamma)\right]^2 \leq \mathbb{E}\left[\left(1 - \frac{\frac{1}{n}\sum_{i=1}^n |Y_i|^q}{\gamma^q}\right)^2\right]\mathbb{E}[(\gamma_n - \gamma)^2] \tag{35}$$

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

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

Lemma 11. *For any $\sigma \geq 0$, we have*

$$(1 - n^\alpha\sigma^q)_+^2 \cdot \begin{cases} \max\{n^{-2\alpha/q}, \sigma^2\} & \text{if } q \geq 2 \\ \max\{n^{-2\alpha/q}, C^{2-q}\sigma^q\} & \text{if } q < 2. \end{cases} \leq \begin{cases} n^{-2\alpha/q} & \text{if } q \geq 2 \\ C^{2-q}\min\{\sigma^q, n^{-\alpha}\} & \text{if } q < 2. \end{cases} \tag{36a}$$

and

$$C^q \min\left\{\frac{\sigma^q}{n^{1-2\alpha}}, \frac{1}{n\sigma^q}\right\} \cdot \begin{cases} \max\{n^{-2\alpha/q}, \sigma^2\} & \text{if } q \geq 2 \\ \max\{n^{-2\alpha/q}, C^{2-q}\sigma^q\} & \text{if } q < 2. \end{cases} \leq \begin{cases} C^q \frac{1}{n^{1-\alpha+2\alpha/q}} & \text{if } q \geq 2 \\ \max\left\{\frac{C^2}{n}, \frac{C^q}{n^{1-\alpha+2\alpha/q}}\right\} & \text{if } q < 2. \end{cases} \tag{36b}$$

We now use Lemma 11 to give the remainder of the proof. First, consider the case that $q \geq 2$. Then choosing $\alpha = 1$ we have $\gamma^q = \max\{n^{-1}, \sigma^q\}$, and

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

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

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

(Asymptotically, then, we obtain $4\max\{C^2, C^{2-q}\}/n$.) Referring to inequality (32), we thus have

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

which was the desired result.

C.1.2   Proof of Lemma 10

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

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

This gives the result of the lemma.

C.1.3   Proof of Lemma 11

We begin with inequality (36a). 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 (36b). 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 (36b) 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 (36b), 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.

C.2 Proof of Corollary 1

Let $\mathcal{F} := \{\ell(\theta; \cdot) : \mathcal{X} \to \mathbb{R} \text{ for } \theta \in \Theta\}$ be our function class. Fix $t > 0$ and let $N = N(\frac{\epsilon_{t,n}}{3}, \mathcal{F}, \|\cdot\|_{L^\infty(\mathcal{X})})$ 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 $\frac{\epsilon_{t,n}}{3}$-cover of $\mathcal{F}$. For any $\theta \in \Theta$, let $i(\theta)$ be such that $\|\ell(\theta; \cdot) - \ell(\theta_{i(\theta)}; \cdot)\|_{L^\infty(\mathcal{X})} \leq \frac{\epsilon_{t,n}}{3}$. We have

$$\sup_{\theta \in \Theta}\left|\mathcal{R}_k(\theta; \widehat{P}_n) - \mathcal{R}_k(\theta; P_0)\right|$$

$$\leq \sup_{\theta \in \Theta}\left\{\left|\mathcal{R}_k(\theta; \widehat{P}_n) - \mathcal{R}_k(\theta_{i(\theta)}; \widehat{P}_n)\right| + \left|\mathcal{R}_k(\theta_{i(\theta)}; \widehat{P}_n) - \mathcal{R}_k(\theta_{i(\theta)}; P_0)\right| + \left|\mathcal{R}_k(\theta_{i(\theta)}; P_0) - \mathcal{R}_k(\theta; P_0)\right|\right\}$$

$$\leq \max_{i=1,\ldots,N}\left|\mathcal{R}_k(\theta_i; \widehat{P}_n) - \mathcal{R}_k(\theta_i; P_0)\right| + \frac{2\epsilon_{t,n}}{3},$$

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

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

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

D Proof of Lower Bounds

All results in Section 5 can be alternatively stated as a probabilistic lower bound on the estimation or optimization error

$$P_0\left(|\widehat{R}(Z_1^n) - \mathcal{R}_f(Z)| \geq n^{-\frac{1}{2\vee k_*}}\right), \text{ or } P_0\left(\mathcal{R}_f\left(\widehat{\theta}(X_1^n); P_0\right) - \inf_{\theta \in \Theta}\mathcal{R}_f\left(\theta; P_0\right) \geq n^{-\frac{1}{2\vee k_*}}\right).$$

These results follow by using the below identical proofs by noting that

$$\inf_{\widehat{\theta}(X_1^n)} \sup_{P_0 \in \mathcal{P}} P_0 \left(|\widehat{R}(Z_1^n) - \mathcal{R}_f(Z)| \geq \delta \right) \geq \frac{1}{2} \left(1 - \|P_1^n - P_2^n\|_{\mathrm{TV}} \right)$$

whenever $|\mathcal{R}_f(Z_1) - \mathcal{R}_f(Z_2)| \geq 2\delta$ for $Z_1 \sim P_1$ and $Z_2 \sim P_2$ (and similarly for the optimization error).

In the coming proofs related to Section 5.1, we define

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

for shorthand, and use it without comment.

D.1 Proof of Theorem 3

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 $Z_1, \ldots, Z_n \overset{\mathrm{iid}}{\sim} P_v$. Assuming that $|\mathcal{R}_k(P_0) - \mathcal{R}_k(P_1)| \geq 2\delta > 0$ for some δ, Le Cam's classical reduction from estimation to testing [65, 107] yields that

$$\mathfrak{M}_n(\mathcal{P}, f) \geq \frac{\delta}{2} \left(1 - \|P_0^n - P_1^n\|_{\mathrm{TV}} \right). \tag{37}$$

We use the bound (37) to give the lower bound by choosing P_0 and P_1 so that $\|P_0^n - P_1^n\|_{\mathrm{TV}} \leq \frac{1}{2}$ and δ is as large as possible.

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

Lemma 12. *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 $z_0 \leq z_1$, let Z be such that*

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

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

See Section D.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 $0 < \delta \leq p_k \wedge (1 - p_k)$ 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 12, we have that $\mathcal{R}_k(Z_1) = M$ and $\mathcal{R}_k(Z_2) \leq M c_k(p_k - \delta)^{\frac{1}{k_*}}$. Consequently, P_1 and P_2 are separated in the robust objective

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

where we used Taylor's theorem

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

It suffices to show that $\|P_1^n - P_2^n\|_{\mathrm{TV}} \leq \frac{1}{2}$ 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 $\|P_1^n - P_2^n\|_{\mathrm{TV}}^2 \leq \frac{n}{2} D_{\mathrm{kl}}(P_2\|P_1)$ so it is enough to show $D_{\mathrm{kl}}(P_2\|P_1) \leq \frac{1}{n}$ for the given value of δ. To this end, we note that for $\delta \leq \frac{1}{2}(1 - p_k)$,

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

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

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

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

for $0 < \delta \leq p_k \wedge (1 - (1 - \beta_k)^{1-k_*} p_k)$ by Lemma 12. This gives the the separation $|\mathcal{R}_k(P_1) - \mathcal{R}_k(P_2)| \geq M \beta_k^{\frac{1}{k}} c_k \delta^{\frac{1}{k_*}}$.

Noting that

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

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

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

where in the first inequality we used the reduction (37) and Pinsker's inequality as before. The desired result follows by setting $\delta = \frac{1}{4n} \wedge p_k \wedge (1 - (1 - \beta_k)^{1-k_*} p_k)$.

D.1.1 Proof of Lemma 12

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

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

46

If $c_k p^{\frac{1}{k*}} \geq 1$ then $\eta^* = \arg\min_\eta g(\eta)$ is attained at z_1 by convexity, and $R(P) = g(\eta^*) = z_1$. If $c_k p^{\frac{1}{k*}} < 1$, we have $\eta^* \leq z_0$ so that

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

which gives the second claim.

For the second inequality, noting that

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

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

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

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

D.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

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

and similarly, $\mathcal{R}_f(Z_2) = Mq(p + \delta)$. 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 $\|P_1^n - P_2^n\|_{\mathrm{TV}} \leq \frac{1}{2}$. By Pinsker's inequality and standard KL vs. χ^2-divergence inequalities [99, Lemmas 2.5–2.7], we have $\|P_1^n - P_2^n\|_{\mathrm{TV}}^2 \leq \frac{n}{2} D_{\mathrm{kl}}(P_1\|P_2)$; we will choose δ such that $D_{\mathrm{kl}}(P_1\|P_2) \leq \frac{1}{n}$. For $\delta \in [0, p]$, Lemma 2.7 of [99] yields

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

Setting $\delta_n = \sqrt{\frac{p(1-p)}{n}}$, we obtain from the reduction from estimation to hypothesis testing (37) that

$$\mathfrak{M}_n(\mathcal{P}, f) \geq \frac{M}{8} q'(p) \sqrt{\frac{p(1 - p)}{n}} + o\left(\frac{1}{\sqrt{n}}\right),$$

which gives the result.

D.3 Proof of Proposition 5

We use LeCam's method and proceed similarly as in the second part of Section D.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 13. *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* $D_f(Q\|P_2) \leq \rho$.

Proof   We have

$$\delta f\left(\frac{\left(\frac{\rho}{2m}\right)^{\frac{1}{k}} \delta^{\frac{1}{k*}}}{\delta}\right) + (1 - \delta)f\left(\frac{1 - \left(\frac{\rho}{2m}\right)^{\frac{1}{k}} \delta^{\frac{1}{k*}}}{1 - \delta}\right)$$

$$\overset{(a)}{\leq} \delta f\left(\left(\frac{\rho}{2m}\right)^{\frac{1}{k}} \delta^{-\frac{1}{k}}\right) + (1 - \delta)f\left(1 - \left(\frac{\rho}{2m}\right)^{\frac{1}{k}} \delta^{\frac{1}{k*}}\right) \overset{(b)}{\leq} \delta f\left(\left(\frac{\rho}{2m}\right)^{\frac{1}{k}} \delta^{-\frac{1}{k}}\right) + \frac{\rho}{2}$$

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 $f^{-1}(s) = \inf\{t \in [0, 1] : f(t) \leq s\}$.

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

As a consequence of Lemma 13, we have $\mathcal{R}_f(Z_2) \geq M\left(\frac{\rho}{2m}\right)^{1/k}\delta^{1/k*}$. As $\mathcal{R}_f(Z_1) = 0$, we have $|\mathcal{R}_f(Z_1) - \mathcal{R}_f(Z_2)| \geq M\left(\frac{\rho}{2m}\right)^{1/k}\delta^{1/k*}$. Proceeding similarly as in the last paragraph of Section D.1 we obtain the result.

D.4 Proof of Theorem 6

Define the optimization distance between two distributions P_0 and P_1 (cf. [2, 35]) 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 14 (Duchi [35, Chs. 5.1–5.2])**.** *If* $P_1, P_2 \in \mathcal{P}$ *are such that* $\mathrm{d}_{\mathrm{opt}}(P_1, P_2; f) \geq \delta$, *then*

$$\mathfrak{M}_n(\mathcal{P}, f, \ell) \geq \delta \inf_{\widehat{\theta}_n} \sup_{P_0 \in \mathcal{P}} P_0\left(\mathcal{R}_f\left(\widehat{\theta}_n(X_1^n); P_0\right) - \inf_{\theta \in \Theta} \mathcal{R}_f(\theta; P_0) \geq \delta\right)$$

$$\geq \frac{\delta}{2}\left(1 - \|P_1^n - P_2^n\|_{\mathrm{TV}}\right).$$

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 $0 < \delta \le p_k \wedge (1 - p_k)$ to be choosen later. Note that

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

For $\delta \le 1 - 2p_k$, we from Lemma 12 that $\mathcal{R}_k(\theta; P_1) = -\theta \mathbf{1}\{\theta < 0\} + \epsilon\theta \mathbf{1}\{\theta \ge 0\}$ and $\mathcal{R}_k(\theta; P_2) = -\theta$ when $\theta < 0$. Now, we have $\mathcal{R}_k(\theta; P_2) \le -\epsilon\theta$ when $\theta \ge 0$ since

$$\sup_{Q \ll P_2} \{\mathbb{E}_Q[X_2] : D_f(Q\|P_2) \le \rho\} \le \epsilon c_k(p_k - \delta)^{\frac{1}{k_*}} + (c_k(p_k - \delta)^{\frac{1}{k_*}} - 1)$$

$$\le \epsilon c_k(p_k - \delta)^{\frac{1}{k_*}} - \frac{\delta}{k_* p_k} \le \epsilon - \frac{\delta}{k_* p_k} = -\epsilon. \tag{38}$$

Here, we used Taylor's theorem

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

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 $\mathcal{R}_k(\theta_1^\star; P_1) = 0$, $\mathcal{R}_k(\theta_2^\star; P_2) \le -M\epsilon$. We then have the following lemma.

Lemma 15. *Let the above conditions hold. Then* $d_{\mathrm{opt}}(P_1, P_2; f_k) \ge \frac{\epsilon}{2}M$.

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

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

$$\ge -\theta\mathbf{1}\{\theta < 0\} - \epsilon\theta\mathbf{1}\{\theta \ge 0\} + M\epsilon$$

$$\ge -\theta\mathbf{1}\{\theta < 0\} - M\kappa\mathbf{1}\{\theta \ge 0\} + M\epsilon \ge \frac{M\epsilon}{2} \ge M\kappa$$

where we used the bound (38) 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 (38),

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

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 (38)

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

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

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

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

Now, we argue as in the proof of Theorem 3. Noting that $D_{\mathrm{kl}}(P_1\|P_2) \leq \frac{\delta^2}{p_k(1-p_k)}$ (e.g. [99, Lemma 2.7]) for $0 \leq \delta \leq (1 - p_k)$, 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 14 yields

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

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 $0 < \delta \leq p_k \wedge (1-p_k) \wedge (1-(1-\beta_k)^{1-k_*}p_k)$ to be choosen later. Now, we again show that $d_{\mathrm{opt}}(P_1, P_2; f_k) \geq \frac{\epsilon}{2}$. To this end, first observe that $\mathcal{R}_k(\theta; P_1) = -\epsilon\theta$. From the first part of Lemma 12, we have $\mathcal{R}_k(\theta; P_2) = -\epsilon\theta \geq 0$ when $\theta < 0$. For $\theta \geq 0$, the last inequality in Lemma 12 gives

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

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 $d_{\mathrm{opt}}(P_1, P_2; f_k) \geq \frac{M\epsilon}{2}$. 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

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

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

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

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

D.5 Proof of Proposition 7

For p given by hypothesis, recall the definition (17) 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 14. 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

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

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

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

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 $\theta_i^\star := \mathrm{argmin}_{\theta \in [-M, M]} \mathcal{R}_f(\theta; P_i)$ satisfies $\theta_1^\star = M$ and $\theta_2^\star = 0$.

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

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

$$= \frac{M}{-\epsilon_\delta + (1 + \epsilon_\delta)q(1 - p)} (2\Delta_\delta - \kappa) \ge 0.$$

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

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

$$\ge M (2\Delta_\delta - \kappa) \ge M\kappa.$$

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 $\mathrm{d}_{\mathrm{opt}}(P_1, P_2) \ge M\Delta_\delta$ as claimed.

We now pick δ such that $\|P_1^n - P_2^n\|_{\mathrm{TV}} \le \frac{1}{2}$. By Pinsker's inequality, we have $\|P_1^n - P_2^n\|_{\mathrm{TV}}^2 \le \frac{n}{2} D_{\mathrm{kl}}(P_1 \| P_2)$, and letting $\delta_n = \sqrt{\frac{p(1-p)}{n}}$, we get $D_{\mathrm{kl}}(P_1 \| P_2) \le \frac{1}{n}$ as for $\delta \in [0, p]$, we have as usual that $D_{\mathrm{kl}}(P_1 \| P_2) \le \frac{\delta^2}{p(1-p)}$. From the reduction from distributionally robust optimization to hypothesis testing (Lemma 14), we conclude

$$\mathfrak{M}_n(\mathcal{P}, f, \ell) \ge \frac{M}{4} \Delta_{\delta_n}.$$

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

D.6 Proof of Proposition 8

We proceed as in the second part of Section D.4. We use Lemma 14 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

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

to be choosen later. Now, we again show that $\mathrm{d}_{\mathrm{opt}}(P_1, P_2; f) \ge \frac{\epsilon}{2}$. To this end, first observe that $\mathcal{R}_f(\theta; P_1) = -\epsilon\theta$. When $\theta < 0$, we have $\mathcal{R}_f(\theta; P_2) \ge \theta\mathbb{E}[X_2] \ge -\theta(\epsilon(1-\delta) - \delta) \ge 0$ as $\epsilon \le \frac{1}{2}$ in the given range of δ. When $\theta \ge 0$, 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 $D_f(Q\|P_2) \le \rho$ by Lemma 13. Hence, we have for $\theta \ge 0$

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

Denoting $\theta_i^\star := \mathrm{argmin}_{\theta \in \Theta} \mathcal{R}_f(\theta; P_i)$ 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 D.4, we can show $\mathrm{d}_{\mathrm{opt}}(P_1, P_2; f) \ge \frac{M\epsilon}{2}$. 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 D.4, we obtain the result.

E Proofs of Consistency

We begin this section with a brief review of the theory of epi-convergence [59, 79], 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*

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

$$\liminf_n A_n := \left\{ v \in \mathbb{R}^d \mid \limsup_{n \to \infty} \mathrm{dist}(v, A_n) = 0 \right\}.$$

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

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

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 $h_n \overset{\mathrm{epi}}{\to} h$, if*

$$\mathrm{epi}\, h = \liminf_{n\to\infty} \mathrm{epi}\, h_n = \limsup_{n\to\infty} \mathrm{epi}\, h_n. \tag{39}$$

If $\mathrm{dom}\, h \neq \varnothing$, meaning that h is proper, epigraphical convergence (39) for closed convex functions has the following equivalent characterizations.

Lemma 16 (Theorem 7.17, Rockafellar and Wets [79]). *Let $h_n : \mathbb{R}^d \to \overline{\mathbb{R}}, h : \mathbb{R}^d \to \overline{\mathbb{R}}$ be closed convex and proper. Then $h_n \overset{\mathrm{epi}}{\to} h$ 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 $C \subset \mathrm{dom}\, h$ not containing a boundary point of $\mathrm{dom}\, h$,

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

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

Lemma 17 (Theorem 7.31, Rockafellar and Wets [79]). *Let $h_n : \mathbb{R}^d \to \overline{\mathbb{R}}, h : \mathbb{R}^d \to \overline{\mathbb{R}}$ satisfy $h_n \overset{\mathrm{epi}}{\to} h$ and $-\infty < \inf h < \infty$. Let $S_n(\varepsilon) = \{\theta \mid h_n(\theta) \leq \inf h_n + \varepsilon\}$ and $S(\varepsilon) = \{\theta \mid h(\theta) \leq \inf h + \varepsilon\}$. Then $\limsup_n S_n(\varepsilon) \subset S(\varepsilon)$ for all $\varepsilon \geq 0$, and $\limsup_n S_n(\varepsilon_n) \subset S(0)$ whenever $\varepsilon_n \downarrow 0$.*

Lemma 18 (Proposition 7.33, Rockafellar and Wets [79]). *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 $h_n \overset{\mathrm{epi}}{\to} h$, then $\inf_v h_n(v) \to \inf_v h(v)$.*

E.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 E.

By Proposition 1, strong duality (4) holds for both $P = P_0$ and $P = \widehat{P}_n$. For a probability measure P, define the dual objective

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

where by convention we use the *closure* of the perspective $(\lambda, t) \mapsto \lambda f^*(t/\lambda)$ (cf. [87, Sec. 3.2] and [51, Prop. IV.2.2.2]). 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 [51, Prop. I.6.1.2] imply $1 \in \partial f^*(0) = \mathrm{argmax}_t\{-f(t)\}$, 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

$$\liminf_{v \to v_0} \left\{ \mathbb{E}_P \left[\lambda f^* \left(\frac{Z - \eta}{\lambda} \right) - (Z - \eta) \right] + \rho\lambda + \eta \right\}$$

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

$$\geq \mathbb{E}_P \left[\lambda_0 f^* \left(\frac{Z - \eta_0}{\lambda_0} \right) - (Z - \eta_0) \right] + \rho\lambda_0 + \eta_0,$$

where the last inequality follows by the lower semicontinuity of the closure of the perspective. 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 16 implies that $g_{f,\widehat{P}_n} \overset{\text{epi}}{\to} g_{f,P_0}$ with probability 1.

Finally, we would like to apply Lemma 18; 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,

$$\limsup_{s \to -\infty} \frac{f^*(s)}{s} = \epsilon < 1 \quad \text{and} \quad \liminf_{s \to \infty} \frac{f^*(s)}{s} = \infty. \tag{40}$$

Deferring the proof of the claims (40), 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

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

and for $\eta > c$ we similarly have

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

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 (40), coupled with our assumption that $\mathbb{E}_{P_0}[|Z|] < \infty$, implies that g_{f,P_0} is coercive. Because $g_{f,\widehat{P}_n} \overset{\text{epi}}{\to} g_{f,P_0}$, we have uniform convergence of $g_{f,\widehat{P}_n}$ to g_{f,P_0} on compacta (Lemma 16), and thus $g_{f,\widehat{P}_n}$ is eventually coercive. Lemma 18 thus implies the result.

Finally, we return to the claim (40). 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 (40), 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.

E.2 Proof of Proposition 10

The epi-convergence theory of the beginning of Section E, 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 $\mathcal{R}_f(\cdot; \widehat{P}_n) \overset{\text{epi}}{\to} \mathcal{R}_f(\cdot; P_0)$ 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 $\sup_{\theta \in C} |\mathcal{R}_f(\theta; P_0) - \mathcal{R}_f(\theta; \widehat{P}_n)| \overset{a.s.}{\to} 0$ by Lemma 16. 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 18 implies that $\inf_{\theta \in \Theta} \mathcal{R}_f(\theta; \widehat{P}_n) \overset{a.s.}{\to} \inf_{\theta \in \Theta} \mathcal{R}_f(\theta; P_0)$.

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 17 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.

F Proof of Limit Theorems

F.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.1) 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. (22), with

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

where $c_k = (k(k-1)\rho + 1)^{1/k} > 1$. It is evident that g_P is convex and coercive in η. Now, for all $\eta \geq \operatorname{ess\,sup} Z$ we have $g_P(\eta) = \eta$, so that g_P is strictly increasing in $\eta \geq \operatorname{ess\,sup} Z$. On the set $(-\infty, \operatorname{ess\,sup} Z)$, we claim that g_P is strictly convex in η. Indeed, for $\eta_1 \neq \eta_2 \in (-\infty, \operatorname{ess\,sup} Z)$ and $\alpha \in (0, 1)$, we have

$$g_P(\alpha \eta_1 + (1-\alpha)\eta_2)$$
$$\overset{(i)}{\leq} c_k \left\| \alpha (Z - \eta_1)_+ + (1-\alpha)(Z - \eta_2)_+ \right\|_{k_*, P} + \alpha \eta_1 + (1-\alpha)\eta_2$$
$$\overset{(ii)}{<} c_k \alpha \left\| (Z - \eta_1)_+ \right\|_{k_*, P} + c_k(1-\alpha) \left\| (Z - \eta_2)_+ \right\|_{k_*, P} + \alpha \eta_1 + (1-\alpha)\eta_2$$
$$= \alpha g_P(\eta_1) + (1-\alpha)g_P(\eta_2),$$

where step (i) follows by convexity and that the norm $\|\cdot\|$ is increasing in positive arguments, while inequality (ii) follows because equality in Minkowski's inequality $\|Y_1 + Y_2\|_{k_*} \leq \|Y_1\|_{k_*} + \|Y_2\|_{k_*}$ for $k_* \in (1, \infty)$ holds if and only if there exists $c \in \mathbb{R}_+$ such that $Y_1 = cY_2$ with probability one.

F.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 $l \leq u$, the bracket $[l, u]$ is the set of functions $h : \mathcal{X} \to \mathbb{R}$ such that $l \leq h \leq u$, and $[l, u]$ is an ϵ-bracket if $\|l - u\| \leq \epsilon$. 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

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

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

Lemma 19 ([103, 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 $\epsilon_0 > 0$,

$$\int_0^\infty \sup_{\epsilon < \epsilon_0} \sqrt{\log N_{[]}\left(\delta \|M_\epsilon\|_{P_0, 2}, \mathcal{H}_\epsilon, L_2(P_0)\right)} d\delta < \infty. \tag{41}$$

(iii) There exists $C < \infty$ such that $\mathbb{E}[M_\epsilon(X)^2] \leq C\epsilon^2$ 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 $c, K > 0$,

$$\lim_{\epsilon \to 0} \epsilon^{-2} \mathbb{E}[M_\epsilon(X)^2 \mathbf{1}\{M_\epsilon(X) > c\}] = 0, \tag{42a}$$

$$\lim_{\epsilon \to 0} \limsup_{\delta \to 0} \sup_{\|u_1 - u_2\| < \epsilon, \|u_1\| \vee \|u_2\| \leq K} \delta^{-2} \mathbb{E}[(H(v^\star + \delta u_1; X) - H(v^\star + \delta u_2; X))^2] = 0 \tag{42b}$$

$$\lim_{\delta \to 0} \delta^{-2} \mathbb{E}[(H(v^\star + \delta u_1; X) - H(v^\star + \delta u_2; X))^2] = \mathbb{E}[(G(u_1) - G(u_2))^2]. \tag{42c}$$

Then, there exists a version of G with bounded, uniformly continuous sample paths on compacta. Further, if $\widehat{v}_n \in \mathcal{V}$ satisfies $\mathbb{E}_{\widehat{P}_n}[H(\widehat{v}_n; X)] \leq \inf_{v \in \mathcal{V}} \mathbb{E}_{\widehat{P}_n}[H(v; X)] + O_P(1/n)$ and $\widehat{v}_n \overset{a.s.}{\to} v^\star$, 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 19. 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

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

For $\epsilon > 0$, define the collection

$$\mathcal{H}_\epsilon := \{ x \mapsto H(\theta, \lambda, \eta; x) - H(\theta^\star, \lambda^\star, \eta^\star; x) : \|\theta - \theta^\star\| + |\lambda - \lambda^\star| + |\eta - \eta^\star| \leq \epsilon \}. \tag{43}$$

We claim that the envelope M_ϵ exists for the set (43). 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 [16]. 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

$$\nabla H(\theta, \lambda, \eta; x) = \begin{bmatrix} f^{*\prime} \left(\frac{\ell(\theta;x) - \eta}{\lambda} \right) \nabla\ell(\theta; x) \\ -f^{*\prime} \left(\frac{\ell(\theta;x) - \eta}{\lambda} \right) + 1 \\ f^* \left(\frac{\ell(\theta;x) - \eta}{\lambda} \right) - \frac{1}{\lambda} f^{*\prime} \left(\frac{\ell(\theta;x) - \eta}{\lambda} \right) (\ell(\theta; x) - \eta) + \rho \end{bmatrix} \tag{44}$$

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

Lemma 20. *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 $t(s) \in \partial f^*(s)$, $t(s) \geq 0$ 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 $f(t) \geq ct^k$ for all $t \geq t_0$. Thus for $s \geq 0$, we have

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

Now we show the second claim. To see this, recall the standard conjugacy result [51] that $t(s) \in \operatorname{argmax}\{st - f(t)\}$, so that $t(s) \geq 0$ always, and let $\hat{t} = (s/kc)^{\frac{1}{k-1}}$. Assume that s is large enough that $f(t) \geq ct^k$ for $t > \hat{t}$. Then for $t > \hat{t}$, we have

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

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

With Lemma 20 in hand, the next lemma follows.

Lemma 21. *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*

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

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. [51]). Thus, we bound the components of ∇H from Eq. (44); 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

$$\|\nabla_\theta H(\theta, \lambda, \eta; x)\| = f^{*\prime} \left(\frac{\ell(\theta; x) - \eta}{\lambda} \right) \|\nabla \ell(\theta; x)\|$$

$$\overset{(i)}{\leq} C \left| \frac{\ell(\theta; x) - \eta}{\lambda} \right|^{\frac{1}{k-1}} \|\nabla \ell(\theta; x)\|$$

$$\overset{(ii)}{\leq} C \left| \frac{\ell(\theta; x) - \eta}{\lambda} \right|^{\frac{k}{k-1}} + C \|\nabla \ell(\theta; x)\|^{k_*}$$

$$\overset{(iii)}{\leq} C \frac{|\eta|^{k_*}}{\lambda^{k_*}} + C \frac{|\ell(\theta^\star; x)|^{k_*}}{\lambda^{k_*}} + C L(x)^{k_*},$$

where inequality (i) follows from Lemma 20, (ii) follows by the Fenchel-Young inequality that $ab \leq (1/k)|a|^k + (1/k_*)|b|^{k_*}$, 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 $\lambda^\star$. $\square$

Finally, we show that each of the conditions of Lemma 19 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 (41), From a standard bound on bracketing numbers for Lipschitz functions [103, Theorem 2.7.11], we have

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

for ϵ small enough, so that the bracketing integral is bounded. Each of the quantities (42) follows by Lebesgue's dominated convergence theorem. For condition (42a), 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 (42b), we have for $v^\star = (\theta^\star, \lambda^\star, \eta^\star)$ that

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

by Lemma 21. Thus the dominated convergence theorem again implies the convergence (42b). For the covariance condition (42c), we use the differentiability of H as in Eq. (44) 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

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

as our Gaussian process. The theorem is then an immediate consequence of Lemma 19.