The Annals of Statistics
2021, Vol. 49, No. 3, 1378–1406
https://doi.org/10.1214/20-AOS2004

LEARNING MODELS WITH UNIFORM PERFORMANCE VIA DISTRIBUTIONALLY ROBUST OPTIMIZATION

BY JOHN C. DUCHI[1] AND HONGSEOK NAMKOONG[2]

[1]*Departments of Statistics and Electrical Engineering, Stanford University, jduchi@stanford.edu*
[2]*Decision, Risk, and Operations Division, Columbia Business School, namkoong@gsb.columbia.edu*

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 [67], 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 [23, 67, 79]. 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 [5]. 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 [20, 45, 49, 81, 92].

In addition to latent heterogeneity in the population, distributional shifts in covariates [12, 87] or unobserved confounding variables (e.g., unmodeled temporal effects [46]) 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 [19, 29, 46, 80, 93] and even when new test data are constructed following identical data construction procedures [74]. Domain adaptation [11, 12, 87] and multitask learning methods [26] can be effective in situations where (potentially unlabeled) data points from the target

Received October 2018; revised July 2020.
MSC2020 subject classifications. 62F12, 68Q32, 62C20.
Key words and phrases. Robust optimization, minimax optimality, risk-averse learning.

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

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

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 [4, 28] 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 $X_i \overset{\text{i.i.d.}}{\sim} P_0$, our approach to minimizing objective (1) is via the plug-in estimator

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

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 nonworst-case distribution shifts, the worst-case formulation (1) *prima-facie* does not guarantee better performance 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 [15] 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 nonexchangeably 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

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

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 $\{Q : D_f(Q \| P_0) \le \rho\}$ 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} [16, 50, 87, 89, 90, 94]. 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 [46], and time-varying-coefficient models are effective when time indices are available [24, 38]. When one believes there may be latent subpopulations, mixture model approaches can model latent membership directly [3, 25, 39, 66]. 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 [67], which connects to minimax regret in linear models [14, 23, 37, 78, 79]. 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, James and Dupuis [72] build worst-case optimal controllers for systems whose uncertain dynamics are described by Kullback–Leibler (KL) divergence balls. In econometrics, Hansen and Sargent [47] 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 [9, 36, 42, 43, 59, 60]. 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 [15, 31, 54]. Other work [13, 15, 34, 59, 61, 70] 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, Glynn and Namkoong [34] and Lam and Zhou [61] show how such approximations correspond to generalized empirical likelihood [71] 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 $\{Q : D_f(Q\|P_0) \le \rho\}$ are Wasserstein balls [17, 18, 40, 65, 68, 73, 82, 88, 100, 101]. 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 [68, 82, 88], and they remain computationally challenging. Furthermore, most guarantees [17, 68, 82] for these problems also consider approximation only of the canonical (population) loss $\mathbb{E}_{P_0}[\ell(\theta; X)]$ using shrinking radius $\rho_n \to 0$. In comparison, our f-divergence formulation is computationally efficient to solve, even in large-scale learning scenarios [69, 70].

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 $Z_n \xrightarrow{p} Z$ 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*

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

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

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

where the supremum is over measurable functions. We now recall Ben-Tal et al. [13] and Shapiro's [85] dual reformulation of the quantity (3), where $f^*(s) := \sup_t\{st - f(t)\}$ is the usual Fenchel conjugate.

PROPOSITION 1 (Shapiro [85], Section 3.2). *Let P be a probability measure on $(\mathcal{X}, \mathcal{A})$ and $\rho > 0$. Then*

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

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

For convex losses $\theta \mapsto \ell(\theta; X)$, the dual form (4) is jointly convex in (θ, η, λ). While interior point methods [22] 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 [69, 70].

Divergence families. Much of our development centers on two families of divergences. The *Rényi α-divergence* [97] 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 $D_1(P\|Q) = D_{\rm kl}(P\|Q)$. For analytical reasons, we use the equivalent Cressie–Read family of f-divergences [27]. These are parameterized by $k \in (-\infty, \infty) \setminus \{0, 1\}$, $k_* = \frac{k}{k-1}$, with

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

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

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

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 $\lambda \ge 0$ 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 $\theta \in \Theta$,*

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

See Section 8.1 of the Supplementary Material [35] 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 $\mathcal{X}$. Equation (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 $\{Q : D_{f_k}(Q \| P) \le \rho\}$ 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 [2].[1]

An extensive literature on coherent risk measures defines utility functions that exhibit "sensible" tail risk preference [7, 57, 76, 86]; there is a duality between distributionally robust optimization and coherent risk measures (e.g., [86], Theorem 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 [57] proposes the dual form (8) as a higher-order generalization of the classical conditional value-at-risk [76], which corresponds to $\mathcal{R}_k(\theta; P)$ defined with $k = \infty$ (or $k_* = 1$) in our notation.

2.1. *Examples.* While—as we note in the Introduction—we do not provide precise recommendations for the choice of f-divergence, it is instructive to consider a few examples for motivation 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)

[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 8.2 [35].

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.

For well-specified linear models, it is frequently the case that the robust parameter $\theta_{\mathrm{dro}} \in \operatorname{argmin}_\theta \mathcal{R}_f(\theta; P)$ 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 [83]: 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 $\theta_\star \in \operatorname{argmin}_\theta \mathcal{R}_f(\theta; P)$ for all f-divergences, as $\mathcal{R}_f$ is a coherent risk measure (cf. [86], Chapter 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 [6, 41], Theorem 11.1, implies that

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

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

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.

EXAMPLE 3 (Worst-case minority performance and CVaR). For $0 < \alpha \le 1$, the conditional value-at-risk [76] (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 $D_\infty(P \| Q) := \lim_{\alpha \to \infty} D_\alpha(P \| Q) = \operatorname{ess\,sup} \log \frac{dP}{dQ}$, and if we define $f_{\infty, c}(t) = 0$ for $0 \le t \le c$ and $+\infty$ otherwise, then the uncertainty region

$$\mathcal{U}_{P_0} := \left\{P \mid D_\infty(P \| P_0) \le \log \frac{1}{\alpha}\right\} = \left\{P \mid D_{f_{\infty, \alpha^{-1}}}(P \| P_0) \le 1\right\}$$

$$= \left\{P \mid \text{there exists } Q, \beta \in [\alpha, 1] \text{ s.t. } P_0 = \beta P + (1 - \beta) Q\right\}$$

by a calculation [86], Example 6.19. The uncertainty set corresponds to distributions with minority subpopulations 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 [58, 84] of risk measures shows that the robust formulations (1) are worst-case CVaR mixtures, $\mathcal{R}_f(\theta; P_0) = \sup_{\mu \in \mathcal{M}_f} \int_0^1 \text{CVaR}_\alpha(\theta; P_0)\, d\mu(\alpha)$ for a set $\mathcal{M}_f$ of probability measures on $[0, 1]$. They thus correspond to drawing a random subpopulation 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 $\{P : D_f(P \| P_0) \le \rho\}$ 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 subpopulations. This recalls Meinshausen and Bühlmann [67], 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 subpopulations 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

$$Y = \theta_v + \varepsilon, \varepsilon \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, I_d).$$

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

$$\theta_{\text{minimax}} := \operatorname*{argmin}_\theta \max_{v \in V} \|\theta - \theta_v\|_2^2 = \operatorname*{argmin}_\theta \max_{v \in V} \mathbb{E}_v\big[\|\theta - Y\|_2^2\big].$$

The unique vector θ_{minimax} coincides with the Chebyshev center of the vectors $\{\theta_v\}$ [22], Chapter 8.5; it also requires knowledge of the groups $v \in V$. In Supplementary Appendix 9.1 [35], we show that if $\theta_\alpha = \operatorname{argmin}_\theta \text{CVaR}_\alpha(\ell(\theta; Y))$, then

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

Recalling from Example 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 subpopulation expected loss.

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

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

assuming implicitly that all parameters are v-measurable. (To show the result in the most straightforward way, we make the simplifying assumptions that $0 < \inf_v \sigma_v^2 \le \sup_v \sigma_v^2 < \infty$, 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 $\operatorname{ess\,sup}_v (\theta - \theta_v)^T \Sigma_v (\theta - \theta_v) + \sigma_v^2 = \sup_v (\theta - \theta_v)^T \Sigma_v (\theta - \theta_v) + \sigma_v^2$. 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

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

We define the minimax estimator to minimize the worst subpopulation risk

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

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 Supplementary Appendix 9.2 [35] 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., $\alpha \downarrow 0$).

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 [52, 98, 99]—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

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

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 $\{Q : D_f(Q\|P_0) \le \rho\}$ 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 subpopulations 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 reunify 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 resulting 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 using a conic interior point solver; we extended the open-source Julia package `convex.jl` to implement power cone solvers [95] (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 [22]. 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 nonconstant, which it is in all of our applications, so $\mathcal{R}_k$ is differentiable [48], Theorem VI.4.4.2, with

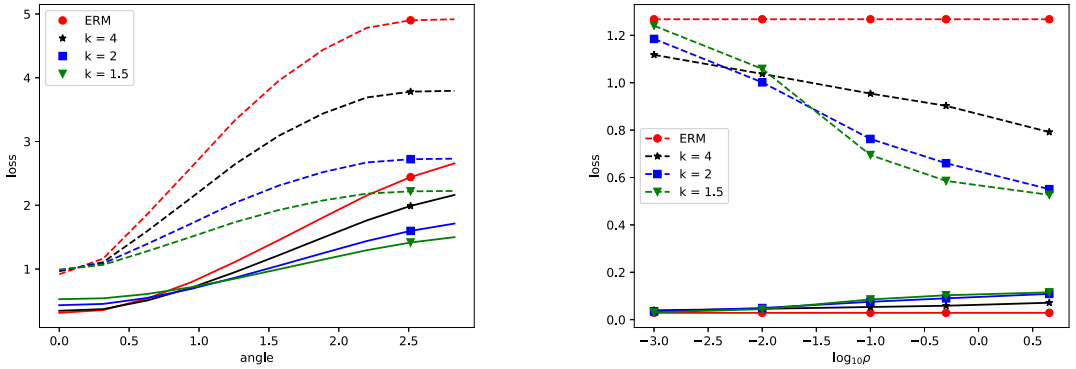
$$\nabla \mathcal{R}_k(\theta, \widehat{P}_n) = \sum_{i=1}^n q_i^* \nabla \ell(\theta; X_i) \quad \text{and} \quad Q^* = \operatorname*{argmax}_{Q: D_{f_k}(Q \| \widehat{P}_n) \le \rho} \left\{ \sum_{i=1}^n q_i \ell(\theta; X_i) \right\}.$$

We use a fast bisection method [70] 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 $\ell(\theta; (x, y)) = (1 - yx^\top\theta)_+$, 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

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



(a) Classification (b) Regression

FIG. 1. (a) *Hinge losses* (*average and* 90*th percentile in solid and dashed lines*, *resp.*) *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* $X^1 \le z_{0.95}$.

(Our below observations still hold when varying these probabilities.) We train our models on $n_{\text{train}} = 100$ training data points, where we use $\rho = 0.5$ 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 $X \overset{\text{i.i.d.}}{\sim} N(0, I_d)$, $\varepsilon \sim N(0, 0.01)$,

$$Y = \begin{cases} X^\top\theta^\star + \varepsilon & \text{if} X^1 \le z_{0.95} = 1.645, \\ X^\top\theta^\star + X^1 + \varepsilon & \text{otherwise}, \end{cases} \tag{11}$$

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 $z_{0.95}$), we expect ERM to have poor performance on rare events when $X^1 \ge 1.645$, or in the tails generally. In addition, a fully robust solution 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 $\rho \in \{0.001, 0.01, 0.1, 0.5, 4.5\}$, 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) improves 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 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.

3.1.3.1. *Two groups*. In our first scenario, for $\theta_0^\star = (1, 0.1), \theta_1^\star = (1, 1)$, we generate

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

where $X \sim N(0, I_2)$, $\varepsilon \sim \mathsf{N}(0, 0.01)$, 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

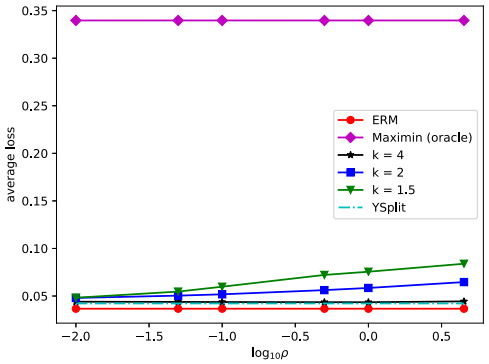
$$G = \begin{cases} 0 & \text{with probability } 0.1 \text{ (minority)}, \\ 1 & \text{with probability } 0.9 \text{ (majority)}. \end{cases} \tag{13}$$

In this two-group setting, we also consider the maximin effects estimator [67]

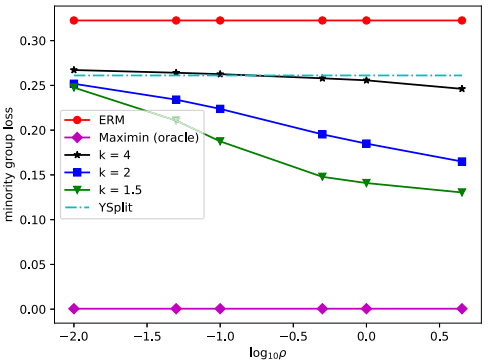
$$\widehat{\theta}_n^{\text{maximin}} = \operatorname*{argmax}_\theta \min_{g=0,1} \big\{2\theta^\top \widehat{\Sigma}_{n,g}\theta_g^\star - \theta^\top \widehat{\Sigma}_{n,g}\theta\big\}$$

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 [67]. 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 (ERM) solution—which has the best average loss and worst minority group loss—and the maximin estimator $\widehat{\theta}_n^{\text{maximin}}$, which sacrifices performance on the average loss for strong minority group performance. The distributionally robust estimators $\widehat{\theta}_n$ exhibit tradeoffs between the two regimes, improving performance on the minority population at smaller degradation in the average loss. The parameters ρ and k allow flexibility in achieving these tradeoffs, though they of course must be set appropriately in applications. 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.



(a) Average loss



(b) Loss on minority group

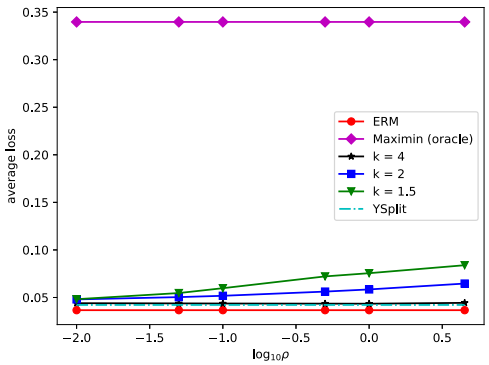
FIG. 2. *Two groups*: *Figures* (a) *and* (b) *plots average and minority group losses under the distribution* (13). *"YSplit" is the performance of the model whose* ρ *and* k *was chosen based on groups formed by sorted values of* Y.

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

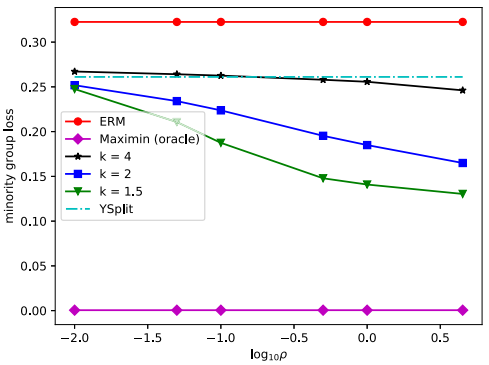
$$G \sim P_G \quad \text{with density } p_G(g) \propto (1-g)^{-\frac{1}{3}}, \tag{14}$$

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 $G = 0.5$, which we interpret as a particular group intervention; we simulate $n_{\text{auxiliary}} = 100$ observations from $G = 0.5$, which is small compared to $n_{\text{train}} = 1000$ 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 $G = 0$.



(a) Average loss



(b) Loss on minority group

FIG. 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 "*$G = 0.5$*" chose* k *and* ρ *based on auxiliary data with intervention* $G = 0.5$.

3.2. *Domain generalization for classification and digit recognition.* In this first of our real experiments, we consider a multiclass 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 handwritten digits [32] (majority population) and typewritten digits consisting of different fonts [30] (minority population). We fix the number of training examples, and vary the minority proportions of typewritten digits from 0–10% of the training data. In the MNIST handwritten 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 typewritten dataset (with the same label).

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

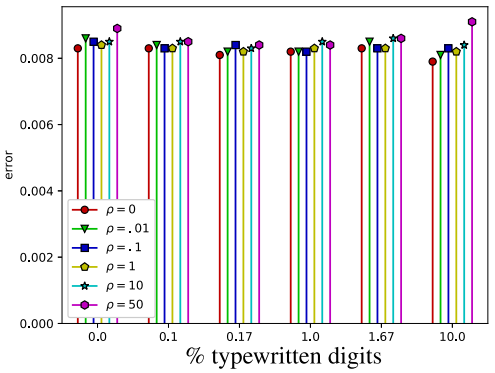
For $y \in \{0, \ldots, 9\}$ and $x \in \mathbb{R}^d$, we use the multiclass logistic loss $\ell(\theta; (x, y)) = \log(\sum_{i=0}^{k} \exp((\theta_i - \theta_y)^\top x))$, where $\theta_i \in \mathbb{R}^d$. For our feature vector X, we use the $d = 4509$-dimensional output of the final fully connected layer of LeNet [63] after 10^4 stochastic gradient steps on the training dataset (see [53] for detailed hyperparameter 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 handwritten test set (majority) across different radii ρ (error below 1%, with a decrease in accuracy of at most 0.1–0.2%). On a test set of all typed digits (Figure 4(b)), the robust solutions exhibit a 1–2% improvement over the nonrobust (ERM) solution in each mixture of typewritten digits (minority proportions) into the training data, which is larger than the persistent 0.1–0.2% degradation on handwritten recognition. The trend of robust improvements on typewritten digits is more pronounced on the harder classes: the gap between $\widehat{\theta}_n^{\text{erm}}$ and $\widehat{\theta}_n$ widens up to 9% on the digit 9 (see Table 1 and Figure 4(d)). We observe that $\widehat{\theta}_n$ consistently performs well on the latent minority (typewritten) subpopulation by virtue of upweighting the hard instances in the training set.

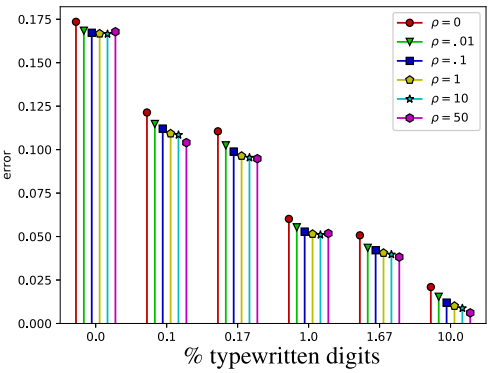
3.3. *Tail performance in a regression problem.* We consider a linear regression problem using the `communities and crime` dataset [8, 75], studying the performance of distributionally robust methods on tail losses. Given a 122-dimensional attribute vector X describing a community, the goal is to predict per capita violent crimes Y (see [75]). We use the absolute loss $\ell(\theta; (x, y)) = |\theta^\top x - y|$ and compare method (2) with constrained forms of lasso, ridge and elastic net regularization [103], taking constraints of the form

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

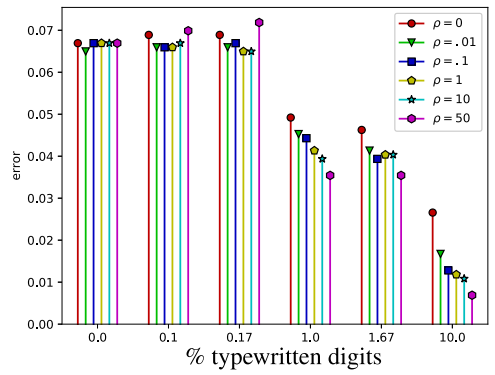
We vary a_1, a_2, and r: for ℓ_1-constraints we take $a_1 = 1, a_2 = 0$ and vary $r_1 \in \{0.05, 0.1, 0.5, 1, 5\}$; for ℓ_2-constraints we take $a_1 = 0, a_2 = 1$ and vary $r_2 \in \{0.5, 1, 5, 10, 50\}$; for elastic net we take $a_1 = 1, a_2 = 10$ and set $r = r_1 + r_2$. We compare these regularizers 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 $r = 0.05$, where we vary $\rho \in \{0.001, 0.01, 0.1, 1, 10\}$.



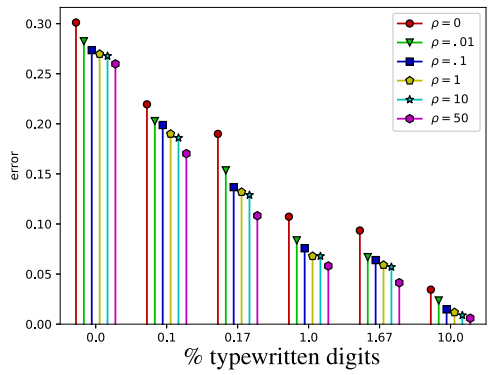
(a) MNIST hand-written Digits



(b) All type-written Digits



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



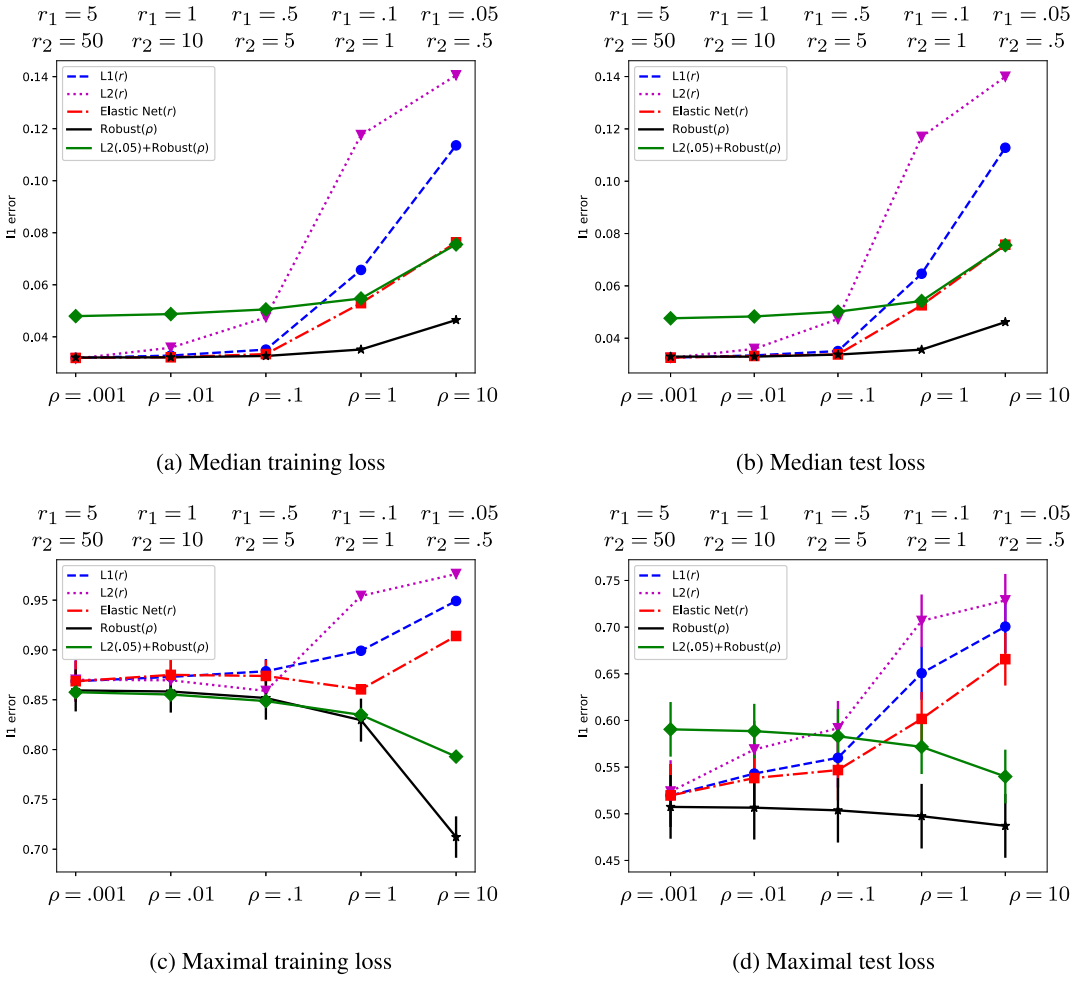
(d) Type-written Digit 9 (hard class)

FIG. 4. (a) *Test error on the* handwritten digits (*MNIST test dataset*). (b)–(d) *Test errors on* typewritten digits. *Models were trained on data consisting of MNIST handwritten digits with* 0–10% *replaced by typewritten digits. The horizontal axis of each plot denotes percentage of typewritten 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 typewritten digits*. (c) *Classification error on digit* 3 *of the typewritten digits*. (d) *Classification errors for digit* 9 *of the typewritten digits*.

In Figure 5, we plot the quantiles of the training and test losses with respect to different values of regularization or ρ. The horizontal axis in each figure indexes our choice of regularization value. We observe that $\widehat{\theta}_n$ shows very different behavior than other regularizers; $\widehat{\theta}_n$ attains median losses similar or slightly higher than the regularized ERM solutions, and achieves much smaller loss on the tails of the inputs. As ρ grows, the robust solution exhibits

TABLE 1
Test error on typewritten digits (%)

Minority proportion	All digits		Digit 9 (hard)		Digit 6 (hard)		Digit 3 (easy)	
	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



(a) Median training loss

(b) Median test loss

(c) Maximal training loss

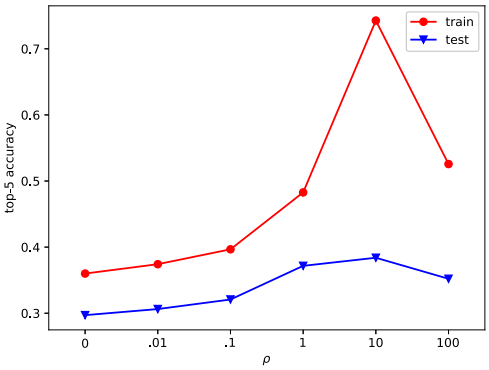
(d) Maximal test loss

FIG. 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* $\rho = 0.001$*. Error bars correspond to standard error.*

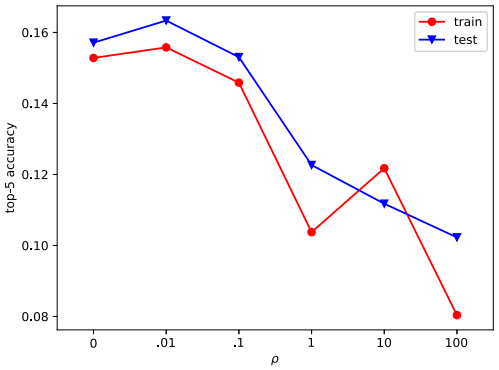
increasing median loss—though slowly—and decreasing maximal loss. To validate our experiments, we made 50 independent random partitions of our dataset with $n = 2118$ samples. For each random partition, we divide the dataset into training set with $n_{\text{train}} = 1800$ and a test set with $n_{\text{test}} = 318$.

3.4. *Fine-grained recognition and challenging subgroups.* Finally, we consider the fine-grained recognition task of the `Stanford Dogs` dataset [55], 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 [96], resulting in vectors $x \in \mathbb{R}^d$ with $d = 12{,}000$.

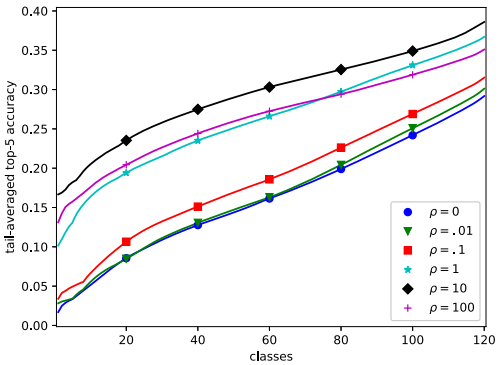
We train 120 one-versus-rest classifiers, one each class, and combine their predictions by taking the k predictions with largest scores for a given example x. For each binary classification problem, we use the binary logistic loss, regularized with lasso (in constrained form) so that $\Theta_{\text{one-vs-rest}} = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \le r\}$. 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



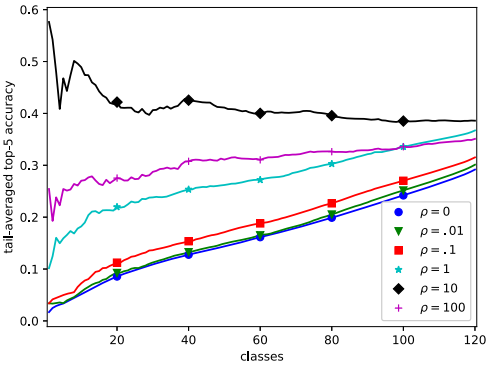
(a) Overall top-5 accuracy



(b) Standard deviation of top-5 accuracy



(c) Test top-5 on worst c classes



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

FIG. 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, that is, c classes with lowest accuracy under each model.* (d) *Test top-5 accuracy on the worst-c classes ordered by accuracy of* ERM *model* ($\rho = 0$).

using the m highest scores, we measure 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 classwise 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 $\widehat{\theta}_n^{\text{erm}}$'s. In Figure 6(d), we plot the accuracy averaged over the first c-classes that have the lowest accuracy under the ERM model. 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 nonrobust classifier goes

from 0.17 vs. 0.03 (hardest class accuracy) to 0.38 vs. 0.28 (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 [34, 44, 59, 61, 70].

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 [21, 91] 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 $n \ge k \vee 3$, with probability at least $1 - 2e^{-t}$*

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

See Section 10.1 of the Supplementary Material [35] for the proof. Relaxing the boundedness assumption $\ell(\theta; x) \in [0, M]$ to sub-Gaussian or subexponential 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., [99]). 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 $\|v - v_i\| \le \epsilon$. 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 $n \ge k \vee 3$, with probability at least $1 - 2N(\mathcal{F}, \frac{\epsilon_{t,n,k}(\rho)}{3}, \|\cdot\|_{L^\infty(\mathcal{X})})e^{-t}$*

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

See Section 10.2 of the Supplementary Material [35] 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 [99], Chapter 2.7.4, is

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

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 [10] and leveraging Rademacher contraction inequalities [64] 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

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

where the outer infimum is over all $(X_1, \ldots, X_n)$-measurable functions and the inner supremum is over probability measures in $\mathcal{P}$, where the loss is implicit in the risk $\mathcal{R}_f$. Whenever $f(t) \lesssim t^k$ as $t \uparrow \infty$, we show there exist losses for which $n^{-1/(k_* \vee 2)}$ is a lower bound on the minimax distributionally robust risk (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, resp.). 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

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

where $Z \sim P_0$ and $Z_1^n \overset{\text{i.i.d.}}{\sim} P_0$, 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 11.1 of the Supplementary Material [35] for the proof. Our proof uses Le Cam's method [62, 102], 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*

$$\mathfrak{M}_n(\mathcal{P}, f_k) \ge M \max\Bigg\{ \frac{1}{8k_* p_k}\bigg(\sqrt{\frac{p_k(1-p_k)}{8n}} \wedge \frac{1}{2}(1-p_k) \wedge p_k\bigg), \\ \frac{1}{8}\beta_k^{\frac{1}{k}} c_k(\rho)\bigg(\frac{1}{4n} \wedge p_k \wedge \big(1 - (1-\beta_k)^{1-k_*} p_k\big)\bigg)^{\frac{1}{k_*}} \Bigg\}.$$

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

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

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

$$q(p) := \sup_{q\ge p}\big\{q : h_f(q; p) \le \rho\big\}. \tag{17}$$

(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,*

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

See Section 11.2 of the Supplementary Material [35] 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 11.3 of the Supplementary Material [35]. Letting $f^{-1}(s) := \inf\{t \in [0, 1] : f(t) \le s\}$ 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 $f(t) \le mt^k$ for $t \ge \{(n \vee C_{f,\rho,m})\rho m^{-1}\}^{\frac{1}{k}}$, then*

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

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 [1, 33], 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 11.4 of the Supplementary Material [35].

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*

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

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 11.5 of the Supplementary Material [35].

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,*

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

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 11.6 of the Supplementary Material [35].

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,*

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

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 $\{P : D_f(P\|P_0) \le \rho\}$. 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 11 of the Supplementary Material [35].

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 [56, 77]. 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 12.1 of the Supplementary Material [35] 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 $\mathcal{R}_f(\theta; \widehat{P}_n) \xrightarrow{\text{a.s.}} \mathcal{R}_f(\theta; P_0) < \infty$.*

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 nonrobust 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

$$d_\subset(A, B) := \sup_{y\in A} \operatorname{dist}(y, B) = \inf_\epsilon\{\epsilon \ge 0 : A \subset \{y : \operatorname{dist}(y, B) \le \epsilon\}\}.$$

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 $\varepsilon \ge 0$ and distribution P, define the set of ε-approximate minimizers

$$S_P(\Theta, \varepsilon) := \Big\{\theta \in \Theta \,\Big|\, \mathcal{R}_f(\theta; P) \le \inf_{\theta\in\Theta} \mathcal{R}_f(\theta; P) + \varepsilon\Big\},$$

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 12.2 of the Supplementary Material [35].

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,*

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

and for any sequence $\varepsilon_n \downarrow 0$, with probability 1 we have $S_{\widehat{P}_n}(\Theta, \varepsilon_n) \neq \varnothing$ eventually and $d_\subset(S_{\widehat{P}_n}(\Theta, \varepsilon_n), S_{P_0}(\Theta)) \to 0$.

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 $|\ell(\theta_0; x) - \ell(\theta_1; x)| \le L(x)\|\theta_0 - \theta_1\|_2$ 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 nonconstant under P and $\mathbb{E}_P[|\ell(\theta; X)|^{k_*}] < \infty$ near θ_0, then $(\lambda_0, \eta_0) = \operatorname{argmin}_{\lambda\ge 0,\eta} g_{P_0}(\theta_0, \lambda, \eta)$ is unique.*

See Supplementary Appendix 13.1 [35] for a proof. Sufficient conditions for differentiability are similar to the classical conditions for asymptotic normality of quantile estimators [98]; 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 $\mathcal{R}_f(\widehat{\theta}_n; \widehat{P}_n) \le \inf_\theta \mathcal{R}_f(\theta; \widehat{P}_n) + o_P(1/n)$. Then*

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

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

See Section 13.2 of the Supplementary Material [35] for the proof. Under the same assumptions, it is straightforward to see that plug-in estimators for V and $\operatorname{Cov}(f^{*\prime}(\frac{\ell(\theta^\star;X)-\eta^\star}{\lambda^\star}) \times \nabla\ell(\theta^\star; X))$ 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

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

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 nonasymptotic considerations [13, 34, 59, 61, 70] 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 [10] 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 [51, 52]. 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.

Funding. Both authors were supported by the SAIL-Toyota Center for AI Research. J. C. Duchi was supported by National Science Foundation Award NSF-CAREER-1553086 and Office of Naval Research YIP Award N00014-19-2288. H. Namkoong was supported by Samsung Fellowship.

SUPPLEMENTARY MATERIAL

Proofs of results (DOI: 10.1214/20-AOS2004SUPP; .pdf). The supplementary material contains proofs of our results.

REFERENCES

[1] AGARWAL, A., BARTLETT, P. L., RAVIKUMAR, P. and WAINWRIGHT, M. J. (2012). Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Trans. Inf. Theory* **58** 3235–3249. MR2952543 https://doi.org/10.1109/TIT.2011.2182178

[2] AHMADI-JAVID, A. (2012). Entropic value-at-risk: A new coherent risk measure. *J. Optim. Theory Appl.* **155** 1105–1123. MR3000633 https://doi.org/10.1007/s10957-011-9968-2

[3] AITKIN, M. and RUBIN, D. B. (1985). Estimation and hypothesis testing in finite mixture models. *J. Roy. Statist. Soc. Ser. B* **47** 67–75.

[4] ALI, S. M. and SILVEY, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B* **28** 131–142. MR0196777

[5] AMODEI, D., ANANTHANARAYANAN, S., ANUBHAI, R., BAI, J., BATTENBERG, E., CASE, C., CASPER, J., CATANZARO, B., CHENG, Q. et al. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. In *Proceedings of the 33rd International Conference on Machine Learning* 173–182.

[6] ANDERSON, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proc. Amer. Math. Soc.* **6** 170–176. MR0069229 https://doi.org/10.2307/2032333

[7] ARTZNER, P., DELBAEN, F., EBER, J.-M. and HEATH, D. (1999). Coherent measures of risk. *Math. Finance* **9** 203–228. MR1850791 https://doi.org/10.1111/1467-9965.00068

[8] ASUNCION, A. and NEWMAN, D. J. (2007). UCI Machine Learning Repository.

[9] ATAR, R., CHOWDHARY, K. and DUPUIS, P. (2015). Robust bounds on risk-sensitive functionals via Rényi divergence. *SIAM/ASA J. Uncertain. Quantificat.* **3** 18–33. MR3299141 https://doi.org/10.1137/130939730

[10] BARTLETT, P. L., BOUSQUET, O. and MENDELSON, S. (2005). Local Rademacher complexities. *Ann. Statist.* **33** 1497–1537. MR2166554 https://doi.org/10.1214/009053605000000282

[11] BEN-DAVID, S., BLITZER, J., CRAMMER, K., KULESZA, A., PEREIRA, F. and VAUGHAN, J. W. (2010). A theory of learning from different domains. *Mach. Learn.* **79** 151–175. MR3108150 https://doi.org/10.1007/s10994-009-5152-4

[12] BEN-DAVID, S., BLITZER, J., CRAMMER, K. and PEREIRA, F. (2007). Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems* 20 137–144.

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

[14] Ben-Tal, A., El Ghaoui, L. and Nemirovski, A. (2009). *Robust Optimization*. *Princeton Series in Applied Mathematics*. Princeton Univ. Press, Princeton, NJ. MR2546839 https://doi.org/10.1515/9781400831050

[15] Bertsimas, D., Gupta, V. and Kallus, N. (2018). Data-driven robust optimization. *Math. Program.* **167** 235–292. MR3755733 https://doi.org/10.1007/s10107-017-1125-8

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

[17] Blanchet, J., Kang, Y. and Murthy, K. (2019). Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Probab.* **56** 830–857. MR4015639 https://doi.org/10.1017/jpr.2019.49

[18] Blanchet, J. and Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Math. Oper. Res.* **44** 565–600. MR3959085 https://doi.org/10.1287/moor.2018.0936

[19] Blitzer, J., McDonald, R. and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the* 2006 *Conference on Empirical Methods in Natural Language Processing* 120–128. Association for Computational Linguistics, Stroudsburg, PA.

[20] Blodgett, S. L., Green, L. and O'Connor, B. (2016). Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of Empirical Methods for Natural Language Processing* 1119–1130.

[21] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[22] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575 https://doi.org/10.1017/CBO9780511804441

[23] Bühlmann, P. and Meinshausen, N. (2016). Magging: Maximin aggregation for inhomogeneous large-scale data. *Proc. IEEE* **104** 126–135.

[24] Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.* **95** 888–902. MR1804446 https://doi.org/10.2307/2669472

[25] Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in Hidden Markov Models*. *Springer Series in Statistics*. Springer, New York. MR2159833

[26] Caruana, R. (1998). Multitask learning. In *Learning to Learn* 95–133. Springer, Berlin.

[27] Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* **46** 440–464. MR0790631

[28] Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2** 299–318. MR0219345

[29] Daumé, H. III and Marcu, D. (2006). Domain adaptation for statistical classifiers. *J. Artificial Intelligence Res.* **26** 101–126. MR2306416 https://doi.org/10.1613/jair.1872

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

[31] Delage, E. and Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* **58** 595–612. MR2680566 https://doi.org/10.1287/opre.1090.0741

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

[33] Duchi, J. C. (2018). Introductory lectures on stochastic optimization. In *The Mathematics of Data*. *IAS/Park City Math. Ser.* **25** 99–185. Amer. Math. Soc., Providence, RI. MR3839168

[34] Duchi, J. C., Glynn, P. W. and Namkoong, H. (2016). Statistics of robust optimization: A generalized empirical likelihood approach. Available at arXiv:1610.03425.

[35] Duchi, J. C. and Namkoong, H. (2021). Supplement to "Learning models with uniform performance via distributionally robust optimization." https://doi.org/10.1214/20-AOS2004SUPP

[36] Dupuis, P., Katsoulakis, M. A., Pantazis, Y. and Plecháč, P. (2016). Path-space information bounds for uncertainty quantification and sensitivity analysis of stochastic dynamics. *SIAM/ASA J. Uncertain. Quantificat.* **4** 80–111. MR3455143 https://doi.org/10.1137/15M1025645

[37] Eldar, Y. C., Ben-Tal, A. and Nemirovski, A. (2004). Linear minimax regret estimation of deterministic parameters with bounded data uncertainties. *IEEE Trans. Signal Process.* **52** 2177–2188. MR2085579 https://doi.org/10.1109/TSP.2004.831144

[38] Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27** 1491–1518. MR1742497 https://doi.org/10.1214/aos/1017939139

[39] FIGUEIREDO, M. A. T. and JAIN, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 381–396.
[40] GAO, R. and KLEYWEGT, A. J. (2016). Distributionally robust stochastic optimization with Wasserstein distance. Available at arXiv:1604.02199.
[41] GARDNER, R. J. (2002). The Brunn–Minkowski inequality. *Bull. Amer. Math. Soc.* (*N.S.*) **39** 355–405. MR1898210 https://doi.org/10.1090/S0273-0979-02-00941-2
[42] GHOSH, S. and LAM, H. (2019). Robust analysis in stochastic simulation: Computation and performance guarantees. *Oper. Res.* **67** 232–249. MR3919867 https://doi.org/10.1287/opre.2018.1765
[43] GLASSERMAN, P. and XU, X. (2014). Robust risk measurement and model risk. *Quant. Finance* **14** 29–58. MR3175968 https://doi.org/10.1080/14697688.2013.822989
[44] GOTOH, J.-Y., KIM, M. J. and LIM, A. (2015). Robust empirical optimization is almost the same as mean-variance optimization. Available at https://ssrn.com/abstract=2827400.
[45] GROTHER, P. J., QUINN, G. W. and PHILLIPS, P. J. (2010). Report on the evaluation of 2D still-image face recognition algorithms. NIST Interagency/Internal Reports (NISTIR) 7709.
[46] HAND, D. J. (2006). Classifier technology and the illusion of progress. *Statist. Sci.* **21** 1–34. MR2275965 https://doi.org/10.1214/088342306000000060
[47] HANSEN, L. P. and SARGENT, T. J. (2008). *Robustness*. Princeton Univ. Press, Princeton, NJ. MR3617628 https://doi.org/10.1515/9781400829385
[48] HIRIART-URRUTY, J.-B. and LEMARÉCHAL, C. (1993). *Convex Analysis and Minimization Algorithms. I*: *Fundamentals. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **305**. Springer, Berlin. MR1261420
[49] HOVY, D. and SØGAARD, A. (2015). Tagging performance correlates with author age. In *Proceedings of the* 53*rd Annual Meeting of the Association for Computational Linguistics* (*Short Papers*) **2** 483–488.
[50] HUANG, J., GRETTON, A., BORGWARDT, K. M., SCHÖLKOPF, B. and SMOLA, A. J. (2007). Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems* 20 601–608.
[51] HUBER, P. J. (1981). *Robust Statistics. Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0606374
[52] HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust Statistics*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2488795 https://doi.org/10.1002/9780470434697
[53] JIA, Y., SHELHAMER, E., DONAHUE, J., KARAYEV, S., LONG, J., GIRSHICK, R., GUADARRAMA, S. and DARRELL, T. (2014). Caffe: Convolutional architecture for fast feature embedding. Available at arXiv:1408.5093.
[54] JIANG, R. and GUAN, Y. (2016). Data-driven chance constrained stochastic program. *Math. Program.* **158** 291–327. MR3511385 https://doi.org/10.1007/s10107-015-0929-7
[55] KHOSLA, A., JAYADEVAPRAKASH, N., YAO, B. and LI, F.-F. (2011). Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization*, *IEEE Conference on Computer Vision and Pattern Recognition* **2** 1–2.
[56] KING, A. J. and WETS, R. J.-B. (1991). Epi-consistency of convex stochastic programs. *Stoch. Stoch. Rep.* **34** 83–92. MR1104423 https://doi.org/10.1080/17442509108833676
[57] KROKHMAL, P. A. (2007). Higher moment coherent risk measures. *Quant. Finance* **7** 373–387. MR2354775 https://doi.org/10.1080/14697680701458307
[58] KUSUOKA, S. (2001). On law invariant coherent risk measures. In *Advances in Mathematical Economics*, *Vol.* 3. *Adv. Math. Econ.* 83–95. Springer, Tokyo. MR1886557 https://doi.org/10.1007/978-4-431-67891-5_4
[59] LAM, H. (2016). Robust sensitivity analysis for stochastic systems. *Math. Oper. Res.* **41** 1248–1275. MR3544795 https://doi.org/10.1287/moor.2015.0776
[60] LAM, H. (2017). Sensitivity to serial dependency of input processes: A robust approach. *Manage. Sci.* **64** 1311–1327.
[61] LAM, H. and ZHOU, E. (2017). The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Oper. Res. Lett.* **45** 301–307. MR3671280 https://doi.org/10.1016/j.orl.2017.04.003
[62] LE CAM, L. and YANG, G. L. (2000). *Asymptotics in Statistics*: *Some Basic Concepts*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1784901 https://doi.org/10.1007/978-1-4612-1166-2
[63] LECUN, Y., BOSER, B., DENKER, J. S., HENDERSON, D., HOWARD, R. E., HUBBARD, W. and JACKEL, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1** 541–551.
[64] LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces*: *Isoperimetry and Processes. Ergebnisse der Mathematik und Ihrer Grenzgebiete* (3) [*Results in Mathematics and Related Areas* (3)] **23**. Springer, Berlin. MR1102015 https://doi.org/10.1007/978-3-642-20212-4

[65] LEE, J. and RAGINSKY, M. (2017). Minimax statistical learning and domain adaptation with Wasserstein distances. Available at arXiv:1705.07815.

[66] MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. *Wiley Series in Probability and Statistics*: *Applied Probability and Statistics*. Wiley Interscience, New York. MR1789474 https://doi.org/10.1002/0471721182

[67] MEINSHAUSEN, N. and BÜHLMANN, P. (2015). Maximin effects in inhomogeneous large-scale data. *Ann. Statist.* **43** 1801–1830. MR3357879 https://doi.org/10.1214/15-AOS1325

[68] MOHAJERIN ESFAHANI, P. and KUHN, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Program.* **171** 115–166. MR3844536 https://doi.org/10.1007/s10107-017-1172-1

[69] NAMKOONG, H. and DUCHI, J. C. (2016). Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems* 29.

[70] NAMKOONG, H. and DUCHI, J. C. (2017). Variance regularization with convex objectives. In *Advances in Neural Information Processing Systems* 30.

[71] OWEN, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18** 90–120. MR1041387 https://doi.org/10.1214/aos/1176347494

[72] PETERSEN, I. R., JAMES, M. R. and DUPUIS, P. (2000). Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Trans. Automat. Control* **45** 398–412. MR1762853 https://doi.org/10.1109/9.847720

[73] PFLUG, G. and WOZABAL, D. (2007). Ambiguity in portfolio selection. *Quant. Finance* **7** 435–442. MR2354780 https://doi.org/10.1080/14697680701455410

[74] RECHT, B., ROELOFS, R., SCHMIDT, L. and SHANKAR, V. (2019). Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the* 36*th International Conference on Machine Learning*.

[75] REDMOND, M. and BAVEJA, A. (2002). A data-driven software tool for enabling cooperative information sharing among police departments. *European J. Oper. Res.* **141** 660–678.

[76] ROCKAFELLAR, R. T. and URYASEV, S. (2000). Optimization of conditional value-at-risk. *J. Risk* **2** 21–42.

[77] ROCKAFELLAR, R. T. and WETS, R. J.-B. (1998). *Variational Analysis*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **317**. Springer, Berlin. MR1491362 https://doi.org/10.1007/978-3-642-02431-3

[78] ROTHENHÄUSLER, D., BÜHLMANN, P., MEINSHAUSEN, N. and PETERS, J. (2018). Anchor regression: Heterogeneous data meets causality. Available at arXiv:1801.06229.

[79] ROTHENHÄUSLER, D., MEINSHAUSEN, N. and BÜHLMANN, P. (2016). Confidence intervals for maximin effects in inhomogeneous large-scale data. In *Statistical Analysis for High-Dimensional Data*. *Abel Symp.* **11** 255–277. Springer, Cham. MR3616272

[80] SAENKO, K., KULIS, B., FRITZ, M. and DARRELL, T. (2010). Adapting visual category models to new domains. In *Proceedings of the European Conference on Computer Vision* 213–226. Springer, Berlin.

[81] SAPIEZYNSKI, P., KASSARNIG, V. and WILSON, C. (2017). Academic performance prediction in a gender-imbalanced environment. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* **1** 48–51.

[82] SHAFIEEZADEH-ABADEH, S., ESFAHANI, P. M. and KUHN, D. (2015). Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems* 28 1576–1584.

[83] SHAKED, M. and SHANTHIKUMAR, J. G. (2007). *Stochastic Orders*. *Springer Series in Statistics*. Springer, New York. MR2265633 https://doi.org/10.1007/978-0-387-34675-5

[84] SHAPIRO, A. (2013). On Kusuoka representation of law invariant risk measures. *Math. Oper. Res.* **38** 142–152. MR3029482 https://doi.org/10.1287/moor.1120.0563

[85] SHAPIRO, A. (2017). Distributionally robust stochastic programming. *SIAM J. Optim.* **27** 2258–2275. MR3715383 https://doi.org/10.1137/16M1058297

[86] SHAPIRO, A., DENTCHEVA, D. and RUSZCZYŃSKI, A. (2009). *Lectures on Stochastic Programming*: *Modeling and Theory*. *MPS/SIAM Series on Optimization* **9**. SIAM, Philadelphia, PA; Mathematical Programming Society (MPS), Philadelphia, PA. MR2562798 https://doi.org/10.1137/1.9780898718751

[87] SHIMODAIRA, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference* **90** 227–244. MR1795598 https://doi.org/10.1016/S0378-3758(00)00115-4

[88] SINHA, A., NAMKOONG, H. and DUCHI, J. C. (2017). Certifiable distributional robustness with principled adversarial training. Available at arXiv:1710.10571.

[89] SUGIYAMA, M., KRAULEDAT, M. and MÜLLER, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.* **8** 985–1005.

[90] SUGIYAMA, M., NAKAJIMA, S., KASHIMA, H., BUENAU, P. V. and KAWANABE, M. (2008). Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems* 21 1433–1440.

[91] TALAGRAND, M. (1996). A new look at independence. *Ann. Probab.* **24** 1–34. MR1387624 https://doi.org/10.1214/aop/1042644705

[92] TATMAN, R. (2017). Gender and dialect bias in YouTube's automatic captions. In *First Workshop on Ethics in Natural Langauge Processing* **1** 53–59.

[93] TORRALBA, A. and EFROS, A. A. (2011). Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1521–1528. IEEE, Piscataway, NJ.

[94] TSUBOI, Y., KASHIMA, H., HIDO, S., BICKEL, S. and SUGIYAMA, M. (2009). Direct density ratio estimation for large-scale covariate shift adaptation. *J. Inf. Process.* **17** 138–155.

[95] UDELL, M., MOHAN, K., ZENG, D., HONG, J., DIAMOND, S. and BOYD, S. (2014). Convex optimization in Julia. In *First Workshop on High Performance Technical Computing in Dynamic Languages* 18–28. IEEE, Piscataway, NJ.

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

[97] VAN ERVEN, T. and HARREMOËS, P. (2014). Rényi divergence and Kullback–Leibler divergence. *IEEE Trans. Inf. Theory* **60** 3797–3820. MR3225930 https://doi.org/10.1109/TIT.2014.2320500

[98] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[99] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

[100] WALD, A. (1945). Statistical decision functions which minimize the maximum risk. *Ann. of Math.* (2) **46** 265–280. MR0012402 https://doi.org/10.2307/1969022

[101] WOZABAL, D. (2012). A framework for optimization under ambiguity. *Ann. Oper. Res.* **193** 21–47. MR2874755 https://doi.org/10.1007/s10479-010-0812-0

[102] YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer, New York. MR1462963

[103] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67** 301–320. MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x