Evaluating model performance under worst-case subpopulations

Anonymous Author(s)
Affiliation
Address
email

Abstract

Typical ML models suffer a substantial performance degradation under subpopulation shifts—when the training population is different to that seen under operation. We develop principled methods for evaluating the worst-case performance of a model over *all* subpopulations of a given size, defined with respect to core attributes Z. Our conservative notion of performance can consider arbitrary Z (in particular, continuous), and automatically accounts for intersectionality in disadvantaged groups by virtue of being agnostic to demographic groupings. As performance of a model conditional on Z is never observed, we propose a scalable two-stage estimation procedure that allows measuring performance against complex subpopulations. We prove a number of finite-sample convergence guarantees for our procedure. In particular, our estimator enjoys *dimension-free* convergence that depend on the dimension of Z only through the empirical out-of-sample error for estimating the performance conditional on Z, instead of overly conservative notions based on Rademacher complexities. On real datasets, we demonstrate that our method certifies the robustness of a model, and allows only deploying models that perform uniformly well across subpopulations.

1 Introduction

In most applications of machine learning, the training population does not accurately represent what the model will encounter under operation. Model performance has been observed to substantially degrade in the presence of such distribution shift—which we term *subpopulation shifts* [16, 28, 72, 83, 53]. Examples are innumerable, ranging from speech recognition [52] and automated essay scoring [4] to wildlife conservation [10]. Similar trends persist for state-of-the-art NLP and computer vision models [81, 77], and even on new data constructed under a near-identical process [60, 69]. Heavily engineered commercial models are no exception [19], performing poorly on rare entities in named entity linking and examples that require abstraction and distillation in summarization tasks [38].

Our data collection systems form the infrastructure on which we build prediction models [30]. They inherit socioeconomic and political inequities of society, often against already marginalized communities. As an example, out of 10,000+ cancer clinical trials the National Cancer Institute funds, less than 5% of participants were non-white [22]. Typical models replicate and perpetuate such biases, and their performance drops significantly on underrepresented groups. Speech recognition systems work poorly for Blacks [52], or those with minority accents [3]. Across demographic attributes such as race, gender, or age, similar performance degradation has been observed in facial recognition, video captioning, language identification, and academic recommender systems [41, 47, 17, 75, 82, 19].

As ML models increasingly affect consequential systems, it is crucial to certify model performance across different subpopulations prior to deployment. Ensuring models achieve uniformly good perfor-

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

mance over all users is simultaneously critical for reliability, fairness, satisfactory user experience, and long term business goals. However, disadvantaged subpopulations are hard to define a priori because of *intersectionality*: the most adversely affected are often determined by a complex combination of variables such as race, income, and gender [19]. For example, performance on summarization tasks vary across demographic characteristics, as well as document specific traits such as abstractiveness, distillation, and location and dispersion of information [38].

Motivated by these challenges, we study the worst-case subpopulation performance across *all* subpopulations of a given size. Let Z be a core set of variables that we wish to guarantee uniform performance over. It may include protected demographic variables such as race, gender, income, age, or task-specific information such as length of the prompt, or metadata on the input; notably, it can contain any continuous or discrete variables. We let $X \in \mathcal{X}$ be the input / covariate, and $Y \in \mathcal{Y}$ be the label. In NLP and vision applications, X is very high-dimensional, and typically $\dim(Z) \ll \dim(X)$. We use $\theta(X)$ to denote a fixed prediction model, and consider flexible and abstract losses $\ell(\theta(x); y)$.

Our goal is to ensure that the model θ performs well over all subpopulations defined over Z. We evaluate model losses on a mixture component, which we call a subpopulation. Postulating a lower bound $\alpha \in (0, 1]$ on the demographic proportion (mixture weight), we consider the set of subpopulations of the data-generating distribution P_Z

$$\mathcal{Q}_\alpha := \{Q_Z \mid P_Z = aQ_Z + (1-a)Q'_Z \text{ for some } a \geq \alpha, \text{ and subpopulation } Q'_Z\}\,. \tag{1}$$

The demographic proportion (mixture weight) a represents how underrepresented the subpopulation is under the data-generating distribution P_Z.

Prior to deploying the model θ, we wish to ensure that its worst-case subpopulation performance

$$\mathsf{W}_\alpha(\theta) := \sup_{Q_Z \in \mathcal{Q}_\alpha} \mathbb{E}_{Z \sim Q_Z}\left[\mathbb{E}[\ell(\theta(X), Y) \mid Z]\right] \tag{2}$$

is below an acceptable level, which guarantees uniform performance over subpopulations (1). The conservative notion of performance (2) evaluates robustness to unanticipated distribution shifts in Z, and it is agnostic to demographic groupings. The worst-case subpopulation performance (2) has a clear interpretation, and its implications can be easily communicated with different stakeholders.

We develop a principled and scalable procedure for estimating the worst-case subpopulation performance (2) for an arbitrary model $\theta(\cdot)$. A key technical challenge is that for each data point, we only observe the loss $\ell(\theta(X); Y)$, but we never observe the conditional risk evaluated at the attribute Z

$$h^\star(Z) := \mathbb{E}[\ell(\theta(X); Y) \mid Z]. \tag{3}$$

In Section 2, we propose a two-stage estimation approach based on sample splitting, where we partition the sample into two parts, S_1, and S_2. On S_1, we compute an estimate $\widehat{h}(\cdot)$ of the conditional risk $h^\star(\cdot)$ using any model class $\mathcal{H}$ of choice. On S_2, we compute a plug-in estimate of the worst-case subpopulation performance under $\widehat{h}(\cdot)$, which we derive using a dual reformulation of the supremum (2). To utilize the entire dataset, we switch the roles of the two samples and average the resulting estimate as our final estimator.

Typical model training for $\theta(\cdot)$ relies heavily on varied engineering practices (e.g. large-scale NLP and vision applications) whose theoretical behavior is currently poorly understood. Nevertheless, it is often highly desirable to *rigorously verify* its performance under subpopulation shifts. Towards this goal, we provide a number of theoretical guarantees for our procedure, showing that it produces accurate estimates of the worst-case subpopulation performance (2). Our first finite-sample result (Section 3.1) shows convergence at the rate $o_p\left(\sqrt{\mathfrak{Comp}_n(\mathcal{H})/n}\right)$, where $\mathfrak{Comp}_n$ denotes a notion of complexity of a well-specified model class for the conditional risk (3). In some applications, it may be natural to define Z using images or natural languages describing the input. Our second result (Section 3.2) shows data-dependent *dimension-free* concentration of our two-stage estimator: our bound depends on the complexity of the model class $\mathcal{H}$ only through the out-of-sample error (evaluated on S_2) for estimating $h^\star(\cdot)$. In practice, this error can be made very small using expressive model classes, whereas the complexity term $\mathfrak{Comp}_n(\mathcal{H})$ in the first result is prohibitively large for deep networks [9, 88]. This bound allows us to estimate $h^\star(\cdot)$ using even the largest deep networks and still obtain a theoretically principled upper confidence bound on $\mathsf{W}_\alpha(\theta)$.

Since an informed choice of α is not always feasible, it is often informative to study the threshold level of $\alpha^\star$ when $\alpha \mapsto \mathsf{W}_\alpha(\theta)$ crosses the *maximum level of acceptable loss* $\bar{\ell}$. The threshold $\alpha^\star$

provides a *certificate of robustness* on the model $\theta(\cdot)$, guaranteeing that all subpopulations large than $\alpha^\star$ enjoy acceptable performance. We develop principled procedures for generating such certificates of robustness (Section 3.3). In Section 4, we demonstate the effectiveness of our procedure on a number of real datasets. By evaluating model robustness under subpopulation shifts, our methods allow only deploying models that can maintain good performance under distribution shift.

Related work The long line of works on distributionally robust optimization (DRO) aims to *train models* that perform well under worst-case distribution shift. Previous approaches considered finite-dimensional worst-case regions, such as constraint sets [29, 39], and those based on notions of distances for probability measures such as f-divergences [11, 12, 57, 56, 63, 33, 32], Levy-Prokhorov [34], Wasserstein distances [35, 76, 15, 37, 14, 85], and reproducing kernels [80, 89]. The notion of distribution shift considered in these approaches are somewhat contrived and difficult to interpret, and often result in overly conservative models. Furthermore, these approaches do not currently scale to modern large-scale NLP or vision applications.

Our work is most closely related to Duchi et al. [31], who propose algorithms for *training* models with respect to the worst-case subpopulation performance (2), which is a more ambitious goal than our narrower viewpoint of *evaluating* model performance prior to deployment. Their (full-batch) training procedure requires solving a convex program with n^2 variables per gradient step, which is often prohibitively expensive. Furthermore, training with respect to reweighted (smoothed) losses $\mathbb{E}[\ell(\theta(X); Y) \mid Z]$ do not scale to deep networks that can overfit to the training data [73]. By contrast, our evaluation perspective aims to take advantage of the rapid rate of progress in deep learning. We build scalable evaluation methods that apply to arbitrary models, which allows leveraging state-of-the-art methods for training $\theta(\cdot)$. By focusing on the question of evaluation, our procedure enjoys convergence guarantees (Section 3.2) that exhibit advantageous scaling with Z, compared to the nonparametric $O_p(n^{-1/d})$ rates for training [31] that scale exponentially with the dimension d of Z.

Our notion of worst-case subpopulation performance is also related to the by now vast literature on fairness and ethics in ML. We give a necessarily abridged discussion and refer readers to Barocas et al. [7], Corbett-Davies and Goel [26] for a more comprehensive treatment of the topic. A large body of work studies *equalizing* a notion of performance over fixed, pre-defined demographic groups for classification tasks [23, 36, 6, 44, 51, 87]. Kearns et al. [49, 50], Hébert-Johnson et al. [46] consider finite subgroups defined by a structured class of functions over Z, and study methods of equalizing performance across them. By contrast, our approach instantiates Rawls' theory of distributive justice [67, 68], where we consider allocation of the loss $\ell(\cdot; \cdot)$ as a resource. Rawls' difference principle maximizes the welfare of the worst-off group and provides incentives for groups to maintain the status quo [67]. Similarly, Hashimoto et al. [45] studied negative feedback loops generated by user retention—they use a more conservative notion of worst-case loss than ours—as poor performance on a currently underrepresented user group can have long-term consequences.

Our diagnostics complement the recent benchmarking approaches under distribution shift [83, 69, 77, 53, 74, 81, 60]. Since good performance on a particular distribution shift may not translate to other shifts, we evaluate model performance over worst-case subpopulations (1) defined over Z.

2 Approach

Accuracy of typical models drops on certain covariate regions [58]. To evaluate model performance across covariates, standard analyses consider pre-defined, fixed demographic groups [62]. While a natural approach, identifying disadvantaged subgroups a priori is often challenging as they are determined by *intersections* of multiple demographic variables. To illustrate such complex intersectionality, consider a drug dosage prediction problem for Warfarin [25], a common anti-coagulant (blood thinner). Taking the best prediction model for the optimal dosage on this dataset Consortium [25] based on genetic, demographic and clinical factors, we present the squared error on the square root of dosage (as considered by Consortium [25]). In Figure 1, when age and race are considered *simultaneously* instead of *separately*, subpopulation performance vary significantly across intersected subgroups. In particular, *worse-case* subpopulation performance deteriorates drastically.

Our worst-case subpopulation performance (2) automatically accounts for arbitrary intersectionality present in the attributes Z. It is agnostic to demographic groupings and allows considering infinitely many subpopulations that represent at least α-fraction of the training population P. By allowing the modeler to select arbitrary protected attributes Z, we are able to consider potentially complex subpopulations. For example, Z can even be defined with respect to a natural language description

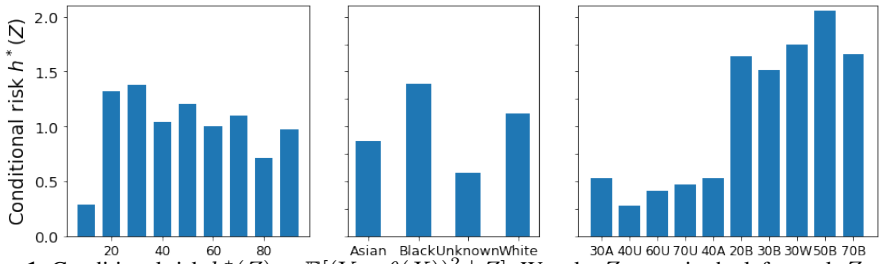

Figure 1. Conditional risk $h^\star(Z) = \mathbb{E}[(Y - \theta(X))^2 \mid Z]$. We take Z = *age* in the left panel, Z = *race* in the center panel, and Z = (*age*, *race*) in the right panel.

of the input X. The choice of Z—and subsequent worst-case subpopulation performance (2) of the conditional risk $h^\star(Z) = \mathbb{E}[\ell(\theta(X); Y) \mid Z]$—interpolates between the most conservative notion of subpopulations (when $Z = (X, Y)$) and simple counterparts defined over a single variable.

The choice of the subpopulation size α should be informed by domain knowledge—desired robustness of the system—and the dataset size relative to the complexity of Z. When it is challenging to commit to a specific subpopulation size, it may be natural to postulate a *maximum level of acceptable loss* $\bar{\ell}$. To measure the robustness of a model, we define the smallest subpopulation size $\alpha^\star(\theta)$ for which the worst-case subpopulation performance is acceptable

$$\alpha^\star(\theta) := \inf\{\alpha : \mathsf{W}_\alpha(\theta) \leq \bar{\ell}\}. \tag{4}$$

This provides a *certificate of robustness*: if $\alpha^\star(\theta)$ is large, then θ is brittle against even majority subpopulations; if it is sufficiently small, then θ performs well on underrepresented subpopulations.

We derive estimators for the worst-case subpopulation performance (2) and the certificate of robustness (4), based on i.i.d. observations $(X_i, Y_i, Z_i)_{i=1}^n \sim P$. We assume our observations are independent from the data used to train the model $\theta(\cdot)$. The worst-case subpopulation performance (2) is unwieldy as it involves an infinite dimensional optimization problem over probabilities. Instead, we use its dual reformulation for tractable estimation. We denote $[\cdot]_+ = \max(\cdot, 0)$, and abuse notation by letting $\mathsf{W}_\alpha(h)$ be the worst-case subpopulation performance for $h(Z)$ (so that $\mathsf{W}_\alpha(\theta) = \mathsf{W}_\alpha(h^\star)$).

Lemma 1 (Shapiro et al. [78, Example 6.19] and Rockafellar and Uryasev [70]). *If* $\mathbb{E}[h(Z)_+] < \infty$,

$$\mathsf{W}_\alpha(h) := \sup_{Q_Z \in \mathcal{Q}_\alpha} \mathbb{E}_{Z \sim Q_Z}[h(Z)] = \inf_{\eta \in \mathbb{R}} \left\{ \frac{1}{\alpha} \mathbb{E}_P [h(Z) - \eta]_+ + \eta \right\}. \tag{5}$$

The dual optimum is attained at the $(1 - \alpha)$-quantile of the $h(Z)$ [71, Theorem 10]. The dual (5) hence shows $\mathsf{W}_\alpha(\theta)$ is a tail-average of $h^\star(Z)$, a popular risk measure known as the conditional value-at-risk (CVaR) in portfolio optimization [70].

A key remaining challenge in estimating $\mathsf{W}_\alpha(\theta)$ is that the conditional risk (3) is never observed, and that we can only observe losses $\ell(\theta(X_i); Y_i)$. We follow a two-stage procedure (Algorithm 1), where we split the sample into two sets S_1 and S_2. On the first sample S_1, we fit an estimator $\widehat{h}(Z)$ of the conditional risk $h^\star(Z)$, using any model class $\mathcal{H}$ (class of mappings $\mathcal{Z} \to \mathbb{R}$), by solving an empirical approximation to the loss minimization problem

$$\underset{h \in \mathcal{H}}{\text{minimize}}\, \mathbb{E}\left[(\ell(\theta(X); Y) - h(Z))^2\right]. \tag{6}$$

For a sufficiently rich model class $\mathcal{H}$, the minimizer is given by $h^\star(Z) = \mathbb{E}[\ell(\theta(X); Y) \mid Z]$. The loss minimization formulation (6) allows the use of any ML estimator, as well as standard tools for model selection (e.g. cross validation). In the second stage, on S_2 we construct a plug-in estimator for the dual form (5), under the estimated conditional risk $\widehat{h}(\cdot)$. In practice, we switch the role of S_1 and S_2 and average the resulting estimates to leverage the entire sample. To estimate the threshold subpopulation size $\alpha^\star(\theta)$, we simply take the plug-in estimator

$$\widehat{\alpha}(\theta) := \inf\{\alpha : \widehat{\mathsf{W}}_\alpha(\widehat{h})(\theta) \leq \bar{\ell}\}. \tag{7}$$

Since $\alpha \mapsto \widehat{\mathsf{W}}_\alpha(\widehat{h})$ is decreasing, the threshold can be efficiently found by a simple bisection search.

3 Convergence guarantees

To *rigorously* verify the robustness of a model prior to deployment, we present convergence guarantees for our estimator (Algorithm 1). In Sections 3.1, we first give finite-sample convergence at the rate

Algorithm 1 Two-stage procedure for estimating worst-case subpopulation performance (2)

1: INPUT: Subpopulation size α, model class $\mathcal{H}$, samples S_1 and S_2
2: On S_1, solve $\widehat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} \frac{1}{|S_1|} \sum_{i \in S_1} \left(\ell(\theta(X_i); Y_i) - h(Z_i)\right)^2$.
3: On S_2, compute the plug-in estimator $\widehat{\mathsf{W}}_\alpha(\widehat{h}) = \inf_\eta \left\{ \frac{1}{\alpha |S_2|} \sum_{i \in S_2} \left[\widehat{h}(Z_i) - \eta\right]_+ + \eta \right\}$.

$O_p(\sqrt{\mathfrak{Comp}_n(\mathcal{H})/n})$, where $\mathfrak{Comp}_n(\mathcal{H})$ is the complexity of the model class $\mathcal{H}$ for estimating the conditional risk $h^\star(Z)$. We use a notion of localized Rademacher complexity [8] for this purpose.

In some situations, it may be appropriate to define subpopulations (Z) over features of an image, or natural language descriptions. For such high-dimensional variables Z and complex model classes $\mathcal{H}$ such as deep networks, the complexity measure $\mathfrak{Comp}_n$ is often prohibitively conservative and renders the resulting concentration guarantee meaningless. In Section 3.2, we provide a finite-sample, data-dependent convergence result that only depends on the size of $\mathcal{H}$ through the out-of-sample error for estimating $h^\star(\cdot)$ (Line 2, Algorithm 1). In practice, the out-of-sample error is directly optimized via hyperparameter tuning and model selection and is often very low for overparameterized models such as deep networks. This allows us to construct valid finite-sample upper confidence bounds for the worst-case subpopulation performance (2) even when Z is defined over high-dimensional spaces.

Finally, in Section 3.3, we provide convergence guarantees for our estimator (7) for the certificate of robustness (4). By building on previous guarantees, we are again able to obtain both types of results.

3.1 Concentration using localized Rademacher complexity

To characterize the finite-sample convergence behavior of our estimator $\widehat{\mathsf{W}}_\alpha(\theta)$, we begin by decomposing the error into two terms relating to the two stages in Algorithm 1. Recalling the notation in Eq. (5) (so that $\mathsf{W}_\alpha(h^\star) = \mathsf{W}_\alpha(\theta)$), we have

$$\widehat{\mathsf{W}}_\alpha(\widehat{h}) - \mathsf{W}_\alpha(h^\star) = \underbrace{\widehat{\mathsf{W}}_\alpha(\widehat{h}) - \mathsf{W}_\alpha(\widehat{h})}_{(a)\text{: second stage}} + \underbrace{\mathsf{W}_\alpha(\widehat{h}) - \mathsf{W}_\alpha(h^\star)}_{(b)\text{: first stage}}. \tag{8}$$

To bound term (a), we prove concentration guarantees for estimators of the dual formulation (5) (see Proposition 4 to come). To bound term (b), we use a localized notion of the Rademacher complexity.

Formally, for $\xi_1, \ldots, \xi_n$ and i.i.d. random signs $\varepsilon_i \in \{-1, 1\}$ (independent of ξ_i), recall the standard notion of (empirical) Rademacher complexity of $\mathcal{G} \subseteq \{g : \Xi \to \mathbb{R}\}$

$$\mathfrak{R}_n(\mathcal{G}) := \mathbb{E}_\varepsilon \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(\xi_i) \right].$$

We say that a function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is *sub-root* [8] if it is nonnegative, nondecreasing, and $r \mapsto \psi(r)/\sqrt{r}$ is nonincreasing for $r > 0$. Any (non-constant) sub-root function is continuous, and has a unique positive fixed point. Let $\psi_n : \mathbb{R}_+ \to \mathbb{R}_+$ be a sub-root upper bound on the localized Rademacher complexity $\psi_n(r) \geq \mathbb{E}\left[\mathfrak{R}_n\left\{g \in \mathcal{G} \mid \mathbb{E}[g^2] \leq r\right\}\right]$. (The localized Rademacher complexity itself is sub-root.) The fixed point of ψ_n characterizes generalization guarantees [8, 54].

To state the main result of this subsection, we restrict attention to nonnegative and bounded losses.

Assumption 1. *There is a B such that $\ell(\theta(X); Y) \in [0, B]$, and $h(Z) \in [0, B]$ a.s. for all $h \in \mathcal{H}$.*

We consider well-specified model classes $\mathcal{H}$ for the first-stage problem (6).

Assumption 2. *The conditional risk is in the model class $\mathcal{H}$: $h^\star(Z) = \mathbb{E}[\ell(\theta(X); Y) \mid Z] \in \mathcal{H}$.*

Let $\psi_{|S_1|}(r)$ be a subroot upper bound on the localized Rademacher complexity around $h^\star$

$$\psi_{|S_1|}(r) \geq B\mathbb{E}\left[\mathfrak{R}_{|S_1|}\left\{h \in \mathcal{H} : B^2\mathbb{E}[(h(Z) - h^\star(Z))^2] \leq r\right\}\right]. \tag{9}$$

We prove the following result in Section A.2, where we define $r^\star_{|S_1|}$ as the fixed point of $\psi_{|S_1|}(r)$.

Theorem 1. *Let Assumptions 1, 2 hold. There is a numerical constant $C > 0$ such that*

$$\left|\mathsf{W}_\alpha(\theta) - \widehat{\mathsf{W}}_\alpha(\widehat{h})\right| \leq \frac{C}{\alpha} B \left(\sqrt{r^\star_{|S_1|}} + \sqrt{\frac{B \log(1/\delta)}{|S_1|}} + \sqrt{\frac{\log(2/\delta)}{|S_2|}} \right) \text{ with probability at least } 1 - 2\delta.$$

Choosing S_1 to be $(1-1/k)$-fraction of the data, and S_2 to be the remaining $1/k$-fraction for a fixed integer k (e.g. $k=5$), we have $|S_1| \asymp |S_2| \asymp n$. Thus, by controlling the fixed point $r^\star_{|S_1|}$ of the localized Rademacher complexity, we are able to provide convergence of our estimator $\widehat{\mathsf{W}}_\alpha(\widehat{h})$. For example, when $\mathcal{H}$ is a bounded VC-class [84], it is known that its fixed point satisfy $r^\star_{|S_1|} \asymp \frac{\mathsf{VC}(\mathcal{H})\log(|S_1|/\mathsf{VC}(\mathcal{H}))}{|S_1|}$ [8, Corollary 3.7], where $\mathsf{VC}(\cdot)$ is the VC-dimension.

Consider nonparametric models in a reproducing kernel Hilbert space (RKHS). For these large model classes, localized arguments (9) can still give fast concentration [59], while standard covering number bounds are prohibitively loose [55]. Let S be a RKHS with norm $\|\cdot\|_S$ and associated kernel (representer of evaluation) $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$. From Mercer's theorem [27], the integral operator $T_K : L^2(\mathcal{Z}, P) \to L^2(\mathcal{Z}, P), T_K(f)(z) = \int K(z, z')dP(z')$ is compact, and $K(z, z') = \sum_{j=1}^\infty \lambda_j \phi_j(z)\phi_j(z')$ where λ_j are sorted eigenvalues of T (decreasing) and ϕ_j form an orthonormal decomposition in $L^2(\mathcal{Z}, P)$. For example, the Gaussian kernel $K(z, z') = \exp(-\frac{1}{2}\|z - z'\|_2^2)$ has eigenvalues that decay exponentially $\lambda_j \lesssim e^{-j^2}$ [59]. Kernel operators for Sobolev spaces [13, 42] have eigenvalues $\lambda_j \lesssim j^{-2\beta}$ where $\beta > \frac{1}{2}$ is the smoothness level, e.g., first-order Sobolev kernel $K(z, z') = 1 + \min\{z, z'\}$ generates RKHS of Lipschitz functions ($\beta = 1$).

For simplicity, let $\mathcal{H}$ be a unit $\|\cdot\|_S$-ball in the RKHS S. In the following result, we denote by $\lesssim$ inequality up to some numerical constant. We give its proof in Section A.3 for completeness.

Lemma 2. *Let Assumptions 1, 2 hold, and assume $\lambda_1 \geq 1/|S_1|$. For a numerical constant $C > 0$,*

$$\psi_{|S_1|}(r) := C\left(\sqrt{\frac{r}{|S_1|}} + \left(\frac{1}{|S_1|}\sum_{j=1}^\infty \min\{\lambda_j, r\}\right)^{\frac{1}{2}}\right)$$

is a subroot upper bound on the localized Rademacher complexity (9) for all $r \geq 1/|S_1|$. If $\lambda_j \lesssim e^{-j^2}$, we have $r^\star_{|S_1|} \lesssim \frac{\sqrt{\log|S_1|}}{|S_1|}$, and if $\lambda_j \lesssim j^{-2\beta}$ for some $\beta > 1/2$, then $r^\star_{|S_1|} \lesssim |S_1|^{-\frac{2\beta}{2\beta+1}}$.

3.2 Data-dependent dimension-free concentration

In some applications, it may be natural to model Z as a high-dimensional variable. This could include large subsets of (X, Y), or defining Z using unstructured information such as images or natural languages. In these instances, even the RKHS-based nonparametric model classes considered in the previous subsection may not be sufficiently expressive. Instead, we may wish to use deep networks to estimate the conditional risk (3) in the problem (6). The localized Rademacher result in Theorem 1, however, becomes prohitively large for neural networks with many layers [9, 88].

We now show alternative concentration guarantees that only depend on the size of model class $\mathcal{H}$ through the out-of-sample error of the learned model $\widehat{h}$ on S_2

$$\Delta_n := \frac{1}{S_2}\sum_{i \in S_2}(\ell(\theta(X_i); Y_i) - \widehat{h}(Z_i))^2. \tag{10}$$

In particular, the term Δ_n is easy to compute. Since architecture selection and hyperparameter tuning optimizes out-of-sample performance, the term Δ_n is often very small in practice. Following convention in learning theory, we refer to our data-dependent concentration guarantee *dimension-free*. Our result allows obtaining finite-sample upper confidence bounds for the worst-case subpopulation performance (2), even when using the largest deep networks to estimate the conditional risk $h^\star(Z)$.

See Section A.4 for the proof of the following data-dependent, dimension-free concentration result.

Theorem 2. *Let Assumptions 1, 2 hold. For a numerical constant $C > 0$,*

$$\left|\mathsf{W}_\alpha(\theta) - \widehat{\mathsf{W}}_\alpha(\widehat{h})\right| \leq \frac{1}{\alpha}\left(\sqrt{\Delta_n} + CB\left(\frac{\log(1/\delta)}{|S_2|}\right)^{1/4}\right) \quad \textit{with probability at least } 1 - 2\delta.$$

Choosing $|S_1| \asymp |S_2| \asymp n$ again, we have the dimension-free convergence rate $O_p(\sqrt{\Delta_n} + n^{-1/4})$.

3.3 Certificate of robustness

Instead of estimating the worst-case subpopulation performance for a posited subpopulation size α, it may be natural to posit a level of acceptable performance (upper bound $\bar{\ell}$ on the loss) and

study $\alpha^\star$, the smallest subpopulation size (4) over which the model $\theta(\cdot)$ can guarantee acceptable performance. Building on our concentration guarantees for $\widehat{\mathsf{W}}_\alpha(\widehat{h})$ in Sections 3.1, 3.2, our plug-in estimator $\widehat{\alpha}(\theta)$ (given in Eq. (7)) of $\alpha^\star$ enjoys similar guarantees as those given in previous sections. The following theorem, proven in Section A.5, states that the true α^* is close to our estimator $\widehat{\alpha}$, or it is sufficiently small, certifying the robustness of the model against subpopulation shifts. Our approach allows simultaeneously proving localized Rademacher complexity bounds and data-dependent dimension-free guarantees. To ease notation, we drop the dependence on the model θ.

Theorem 3. *Let Assumption 1 hold, let $U(\delta) > 0$ be such that for any fixed $\alpha \in (0, 1]$, $|\widehat{\mathsf{W}}_\alpha(\widehat{h}) - \mathsf{W}_\alpha(\theta)| \leq U(\delta)/\alpha$ with probability at least $1 - \delta$. Then given any $\underline{\alpha} \in (0, 1]$, either $\alpha^\star < \underline{\alpha}$, or*

$$\left|\frac{\alpha^\star}{\widehat{\alpha}} - 1\right| \leq \frac{U(\delta)}{\widehat{\mathbb{E}}\left[\widehat{h}(Z) - \widehat{P}^{-1}_{1-\underline{\alpha}\wedge\widehat{\alpha}}\left(\widehat{h}(Z)\right)\right]_+}$$

with probability at least $1 - \delta$, where $\widehat{\mathbb{E}}$ and $\widehat{P}^{-1}_{1-\alpha}$ denote the expectation and the $(1 - \alpha)$-quantile under the empirical probability measure induced by $(Z_i : i \in S_2)$.

4 Experiments

On two real datasets, we demonstrate that our diagnostic allows certifying model performance across subpopulations. We first study a drug dosage prediction problem, where our procedure ascertains the robustness of a linear regression model over substantially more expressive model classes. Then, we turn to a large-scale vision application based on sattelite images (FMoW [24]) where natural distribution shifts were recently considered [53]. We illustrate how our worst-case subpopulation approach is able to highlight the robustness of the recently proposed CLIP model [65] without knowledge of out-of-distribution samples. For all experiments, we use gradient boosted decision trees (using the package XGBoost [21]) to estimate the conditional risk $h^\star(Z) = \mathbb{E}[\ell(\theta(X); Y) \mid Z]$.

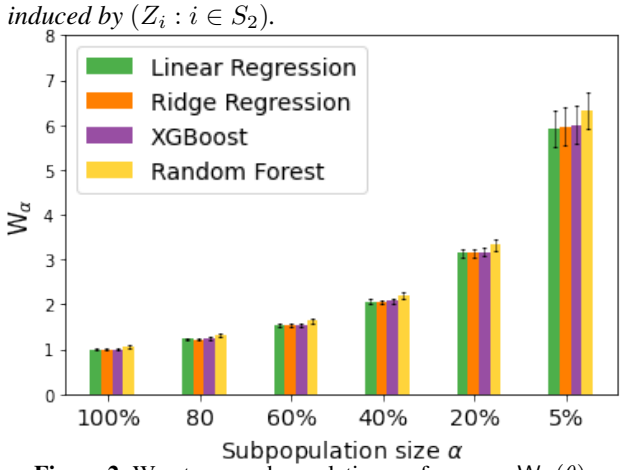

Figure 2. Worst-case subpopulation performance $\mathsf{W}_\alpha(\theta)$. When $\alpha = 1.0$, we have $\mathsf{W}_{1.0}(\theta) = \mathbb{E}[\ell(\theta(X); Y)]$.

Warfarin Warfarin is a widely prescribed anticoagulant, often prescribed to prevent strokes [25]. Its optimal dosage varies substantially across genetics, demographics, and existing conditions (up to ten times). We study a Pharmacogenetics and Pharmacogenomics Knowledge Base dataset constructed from optimal dosages found by trial and error by clinicians; the dataset consists of 5,700 patients (4,788 patients after excluding missing data), and features representing demographics, genetic markers, medication history, pre-existing conditions, and diagnosis (reason for treatment). Consortium [25] found a linear model outperforms a number of more complicated modeling approaches (e.g. kernel methods, neural networks, splines, boosting) for predicting the optimal dosage.

Such average-case performance needs to translate uniformly to different subpopulations; we need to ensure automated medical models benefit underrepresented groups [20, 66, 40, 2]. We wish to evaluate and compare the worst-case subpopulation performance of different models over $Z = X$, the entire feature vector. Following Consortium [25], we take the root-dosage as our outcome Y, and consider the squared loss $\ell(\theta(X); Y) = (Y - \theta(X))^2$. In Figure 2, we observe that the linear model closely matches the performance of more expressive models even over small subpopulations. The trend holds over a range of different subpopulation sizes (up to $\alpha = 5\%$). Our finding instills confidence in the linear regression model: in addition to being simple, interpretable, and easy-to-debug, our diagnostic certifies its advantageous performance even on tail subpopulations.

Functional Map of the World (FMoW) Satellite images can impact and direct economic and environmental policies globally, by measuring poverty [1] and evaluating economic growth [43], as well as population changes and deforestation. An automated approach allows analyzing data from

remote regions at relatively low cost, and continuous monitoring of land usage. Towards this goal, it is critical the models perform reliably across time and space. We study this problem on the Functional Map of the World (FMoW) dataset [24], where the goal is to predict building / land use categories (62 classes) based on a satellite image. Across different models, we observe that performance remains similar temporally or spatially when each dimension is considered *separately*, but have substantial variability—in particular, subpopulations with much worse performance—when region and year are considered *simultaneously*. We present such trends for a standard DenseNet ERM model [48, 53] (which achieves near-SOTA performance) in Figure 3a. We also observe substantial variability in performance over different classes. In Figure 3b, we plot the error rate per class DenseNet ERM model, demonstrating clear heterogeneity in the difficulty of different classes. For both results, we observed similar patterns on other models (see Section B).

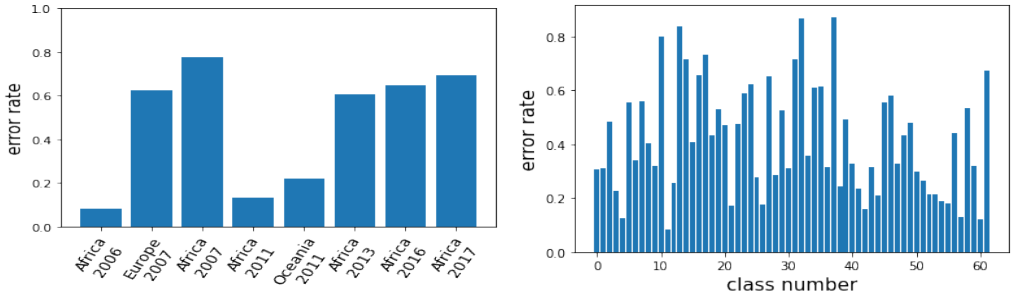

Figure 3: For DenseNet ERM, spatiotemproal intersectionality (left) and performance by class (right)

We take the perspective of an analyst evaluating prediction models for land usage, based on data collected during 2002-2013. Our worst-case subpopulation approach (Algorithm 1) evaluates and compares model performance across subpopulation shifts over space and time. The FMoW dataset provides fertile grounds for demonstrating our method as it includes natural distribution shifts [53]. We study how our procedure can select models that perform well "in the future" on out-of-distribution samples collected in 2013-2015 [53]. For each satellite image, we use meta information consisting of Z =(*longitude, latitude, cloud cover, region, year*), over which we define subpopulations (1). Since model performance also vary over different classes, we also consider an alternative definition of Z that includes the label $Z = (Y$, *longitude, latitude, cloud cover, region, year*).

We study a range of different models trained on the FMoW-WILDS training set (collected in 2002-2013, n =76,863). First, we consider 9 ImageNet pre-trained models fine-tuned on FMoW, which includes *DenseNet* models considered by Koh et al. [53] and *ResNet*-18 and *ResNet*-50 models studied by Miller et al. [61]. In addition to the ERM procedure that optimizes average-case performance, we also consider the recently proposed invariant risk minimization (IRM) framework [5] for the DenseNet architecture used by Koh et al. [53]. We also consider the recently proposed CLIP model [65], which were trained on multiple data sources using natural language supervision and contrastive losses (40M image-text pairs). On natural distribution shifts on ImageNet, Radford et al. [65] demonstrated that CLIP exhibits substantial robustness over other state-of-the-art methods. Motivated by this observed success, we take the pre-trained CLIP ViT-32B model, and look at the zero-shot model and its fine-tuned counterpart, which we call *CLIP Linear Probe*, following Radford et al. [65].

We compute estimators of W_α (Algorithm 1) on a heldout validation set (collected in 2002-2013, n =11,483), using the standard cross entropy loss on which all models except CLIP were trained on. On data collected in 2013-2015 (n =19,915), we observe our procedure selects models that exhibit substantial robustness in the future. In the main text, we report results for a subset of the models for clarity; see Section B for complete results and training specifications.

Figure 4 summarizes the worst-case subpopulation performances of models over the *metadata* (and labels in panel (b)), across different subpopulation sizes α. In both cases, DenseNet IRM has the best average performance ($\alpha = 100\%$). As α decreases, the worst-case subpopulation performances of the DenseNet and ImageNet models deteriorate faster than that of *CLIP Linear Probe*. At $\alpha = 1\%$, *CLIP Linear Probe* has the best worst-case subpopulation performance in comparison. The robustness of *CLIP Linear Probe* against subpopulation shifts is marked when subpopulations are defined also over classes in panel (b). In both cases, our worst-case subpopulation performance suggests that *CLIP Linear Probe* can maintain good performance even under distribution shift. Notably, this choice is different from that given by a comparison solely based on the average-case performance ($\alpha = 100\%$).

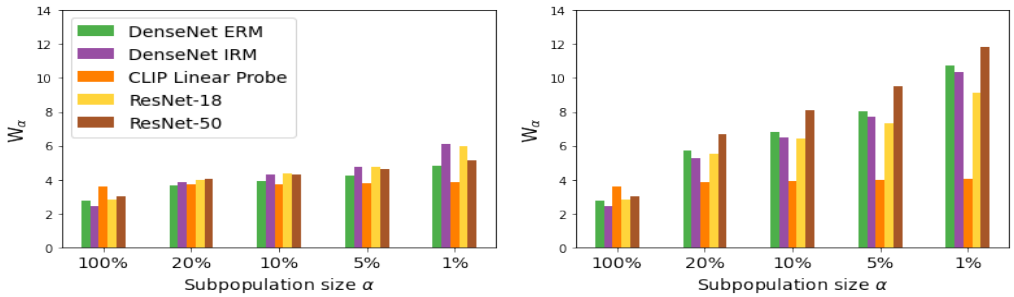

Figure 4. Worst-case subpopulation performance $\mathsf{W}_\alpha(\theta)$ defined over Z. In panel (a) (left) Z =(*metadata*), and in panel (b) (right) $Z = (Y, \textit{metadata})$, where metadata consists of (longitude, latitude, cloud cover, region, year). When $\alpha = 1.0$, we have $\mathsf{W}_{1.0}(\theta) = \mathbb{E}[\ell(\theta(X); Y)]$.

To illustrate the robustness of our chosen method on concrete groupings of subpopulation, We study performance of different models under spatiotemporal distribution shift in Table 1. We evaluate the models' worst case loss and prediction error on all subpopulation grouped by (region, year, label) that constitute at least 1% of out-of-distribution data. The chosen model *CLIP Linear Probe* yields the best worst-case prediction accuracy among all models. Moreover, for all models our estimator of the worst-case subpopulation performance (2) measure reported in Column $\mathsf{W}_{0.01}$ (taken from Figure 4b) gives an upper bound on the worst-case loss.

Model	Worst Subpopulation Grouping	Worst 0-1 Error	Worst Loss	$\mathsf{W}_{0.01}$
DenseNet ERM	2015, Americas, place of worship	0.55	1.90	10.74
DenseNet IRM	2015, Americas, place of worship	0.63	3.19	10.35
CLIP Linear Probe	2015, Americas, recreational facility	0.54	3.55	4.07
ResNet-18	2015, Americas, place of worship	0.73	3.93	9.16
ResNet-50	2015, Americas, place of worship	0.55	3.21	11.85

Table 1: Model performance across *worst-case* subgroups larger than 1% of the data

A key advantage of our method is the flexibility in choosing Z. Since defining subpopulations over all the metadata can be conservative, we present additional results under Z =(*region*, *year*) in Section B, as well as its counterpart including the label Y. To incorporate richer information in the labels, we repeat previous experiments using a natural language description of the labels. We pass each class label with 17 engineered prompts to the CLIP text encoder [65], and take the averaged feature vector in place of the label Y in the previous definitions of Z. As we showcase in Section B, this allows taking advantage of the *semantic meaning of each class label*, as learned by the CLIP model.

5 Discussion

Our approach allows flexible modeling of subpopulations over an arbitrary variable Z and automatically accounts for complex intersectionality. We develop scalable estimation procedures for the worst-case subpopulation performance (2) and the certificate of robustness (4) of a model. Our convergence guarantees apply even when we use high-dimensional inputs (e.g. natural language) to define Z. Our diagnostic may pose implications for data collection, and inform further model improvement: focusing data collection efforts on concentrated and interpretable regions of $\mathcal{Z}$ where the conditional risk (3) is high, or methods that can "fixing" the model on such regions.

The worst-case performance (2) over mixture components as subpopulations (1) provides a strong guarantee over arbitrary subpopulations, but it may be overly conservative in cases when there is a natural geometry in $Z \in \mathcal{Z}$. Incorporating such problem-specific structures to define tailored notion of subpopulations is a promising research direction towards operationalizing the concepts put forth in this work. As an example, Srivastava et al. [79] recently studied similar notions of worst-case performance defined over human annotations.

We focus on the narrow question of evaluating model robustness under distribution shift; our evaluation perspective is thus inherently limited. Data collection systems inherit socioeconomic inequities, and reinforce existing political power structures. This affects *all* aspects of the ML development pipeline, and our diagnostic is no panacea. Systems must be deployed with considered analysis of its adverse impacts, and we advocate for a holistic approach towards addressing its varied implications.

References

[1] B. Abelson, R. Kush, and J. Sun. Targeting direct cash transfers to the extremely poor. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.

[2] American Medical Association. AMA passes first policy recommendations on augmented intelligence., 2018. URL www.ama-assn.org/ama-passes-first-policy-recommendations-augmented-intelligence.

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

[4] E. Amorim, M. Cançado, and A. Veloso. Automated essay scoring in the presence of biased ratings. In *Association for Computational Linguistics (ACL)*, pages 229–237, 2018.

[5] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. In *Proceedings of the Eighth International Conference on Learning Representations*, 2020.

[6] S. Barocas and A. D. Selbst. Big data's disparate impact. *104 California Law Review*, 3: 671–732, 2016.

[7] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019.

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

[9] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6241–6250, 2017.

[10] S. Beery, E. Cole, and A. Gjoka. The iwildcam 2020 competition dataset. *arXiv:2004.10340 [cs.CV]*, 2020.

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

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

[13] M. Birman and M. Solomjak. Piecewise-polynomial approximations of functions of the classes W_p^α. *Sbornik: Mathematics*, 2(3):295–317, 1967.

[14] J. Blanchet, Y. Kang, F. Zhang, and K. Murthy. Data-driven optimal transport cost selection for distributionally robust optimizatio. *arXiv:1705.07152 [stat.ML]*, 2017.

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

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

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

[18] D. B. Brown. Large deviations bounds for estimating conditional value-at-risk. *Operations Research Letters*, 35(6):722–730, 2007.

[19] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[20] D. S. Char, N. H. Shah, and D. Magnus. Implementing machine learning in health care—addressing ethical challenges. *New England Journal of Medicine*, 378(11):981, 2018.

[21] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[22] M. S. Chen Jr, P. N. Lara, J. H. Dang, D. A. Paterniti, and K. Kelly. Twenty years post-nih revitalization act: enhancing minority participation in clinical trials (empact): laying the groundwork for improving minority clinical trial accrual: renewing the case for enhancing minority participation in cancer clinical trials. *Cancer*, 120:1091–1096, 2014.

[23] A. Chouldechova. A study of bias in recidivism prediciton instruments. *Big Data*, pages 153–163, 2017.

[24] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018.

[25] I. W. P. Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753–764, 2009.

[26] S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv:1808.00023 [cs.CY]*, 2018.

[27] N. Cristianini and J. Shawe-Taylor. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

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

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

[30] E. Denton, A. Hanna, R. Amironesei, A. Smart, H. Nicole, and M. K. Scheuerman. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv:2007.07399 [cs.CY]*, 2020.

[31] J. Duchi, T. Hashimoto, and H. Namkoong. Distributionally robust losses for latent covariate mixtures. *arXiv:2007.13982 [stat.ML]*, 2020.

[32] J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *Annals of Statistics*, 2021.

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

[34] E. Erdoğan and G. Iyengar. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, 107(1-2):37–61, 2006.

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

[36] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

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

[38] K. Goel, N. Rajani, J. Vig, S. Tan, J. Wu, S. Zheng, C. Xiong, M. Bansal, and C. Ré. Robustness gym: Unifying the nlp evaluation landscape. *arXiv:2101.04840 [cs.CL]*, 2021.

[39] J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Operations research*, 58(4-part-1):902–917, 2010.

[40] S. N. Goodman, S. Goel, and M. R. Cullen. Machine learning, health disparities, and causal reasoning. *Annals of Internal Medicine*, 2018.

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

[42] C. Gu. *Smoothing spline ANOVA models*. Springer, 2002.

[43] S. Han, D. Ahn, S. Park, J. Yang, S. Lee, J. Kim, H. Yang, S. Park, and M. Cha. Learning to score economic development from satellite imagery. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 2970–2979, 2020.

[44] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29*, 2016.

[45] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[46] Ú. Hébert-Johnson, M. P. Kim, O. Reingold, and G. N. Rothblum. Calibration for the (computationally-identifiable) masses. *arXiv:1711.08513 [cs.LG]*, 2017.

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

[48] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the International Conference on Computer Vision*, pages 4700–4708, 2017.

[49] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv:1711.05144 [cs.LG]*, 2018.

[50] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 100–109. ACM, 2019.

[51] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science (ITCS)*, 2016.

[52] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[53] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, S. Beery, et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv:2012.07421 [cs.LG]*, 2020.

[54] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6):2593–2656, 2006.

[55] T. Kühn. Covering numbers of Gaussian reproducing kernel Hilbert spaces. *Journal of Complexity*, 27(5):489–499, 2011.

[56] H. Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019. URL http://arXiv.org/abs/1605.09349.

[57] H. Lam and E. Zhou. Quantifying input uncertainty in stochastic optimization. In *Proceedings of the 2015 Winter Simulation Conference*. IEEE, 2015.

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

[59] S. Mendelson. On the performance of kernel classes. *Journal of Machine Learning Research*, 4 (Oct):759–771, 2003.

[60] J. Miller, K. Krauth, B. Recht, and L. Schmidt. The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning*, pages 6905–6916. PMLR, 2020.

[61] J. Miller, R. Taori, A. Raghunathan, S. Sagawa, P. W. Koh, V. Shankar, P. Liang, Y. Carmon, and L. Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[62] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019.

[63] T. Miyato, S.-i. Maeda, M. Koyama, K. Nakae, and S. Ishii. Distributional smoothing with virtual adversarial training. *arXiv:1507.00677 [stat.ML]*, 2015.

[64] L. Prashanth, K. Jagannathan, and R. K. Kolla. Concentration bounds for cvar estimation: The cases of light-tailed and heavy-tailed distributions. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[65] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. *arXiv:2103.00020 [cs.LG]*, 2021.

[66] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 2018.

[67] J. Rawls. *Justice as fairness: A restatement*. Harvard University Press, 2001.

[68] J. Rawls. *A theory of justice*. Harvard university press, 2009.

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

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

[71] R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

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

[73] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *Proceedings of the Seventh International Conference on Learning Representations*, 2019.

[74] S. Santurkar, D. Tsipras, and A. Madry. Breeds: Benchmarks for subpopulation shift. *arXiv:2008.04859 [cs.CV]*, 2020.

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

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

[77] V. Shankar, A. Dave, R. Roelofs, D. Ramanan, B. Recht, and L. Schmidt. Do image classifiers generalize across time? *arXiv:1906.02168 [cs.LG]*, 2019.

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

[79] M. Srivastava, T. Hashimoto, and P. Liang. Robustness to spurious correlations via human annotations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9109–9119. PMLR, 2020.

[80] M. Staib and S. Jegelka. Distributionally robust optimization and generalization in kernel methods. In *Advances in Neural Information Processing Systems*, pages 9131–9141, 2019.

[81] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems 20*, 2020.

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

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

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

[85] R. Volpi, H. Namkoong, O. Sener, J. Duchi, V. Murino, and S. Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems 31*, 2018.

[86] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[87] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR, 2017.

[88] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *Proceedings of the Fifth International Conference on Learning Representations*, 2017.

[89] J.-J. Zhu, W. Jitkrittum, M. Diehl, and B. Schölkopf. Kernel distributionally robust optimization: Generalized duality theorem and stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 280–288. PMLR, 2021.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 3
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Appendix B
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix B
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figure 2
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] Our diagnostic procedure does not require large computing resource.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Reference list
   (b) Did you mention the license of the assets? [Yes] See FMoW citation.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] This is only relevant to us when we took pretrained ImageNet models fine-tuned on FMoW. We obtained permission to do so.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]