# Evaluating model performance under worst-case subpopulations

**Mike Li***
Decision, Risk, and Operations Division
Columbia Business School
New York, NY 10027
MLi24@gsb.columbia.edu

**Hongseok Namkoong**
Decision, Risk, and Operations Division
Columbia Business School
New York, NY 10027
namkoong@gsb.columbia.edu

**Shangzhou Xia**
Decision, Risk, and Operations Division
Columbia Business School
New York, NY 10027
SXia24@gsb.columbia.edu

## Abstract

The performance of ML models degrades when the training population is different from that seen under operation. Towards assessing distributional robustness, we study the worst-case performance of a model over *all* subpopulations of a given size, defined with respect to core attributes $Z$. This notion of robustness can consider arbitrary (continuous) attributes $Z$, and automatically accounts for complex intersectionality in disadvantaged groups. We develop a scalable yet principled two-stage estimation procedure that can evaluate the robustness of state-of-the-art models. We prove that our procedure enjoys several finite-sample convergence guarantees, including *dimension-free* convergence. Instead of overly conservative notions based on Rademacher complexities, our evaluation error depends on the dimension of $Z$ only through the out-of-sample error in estimating the performance conditional on $Z$. On real datasets, we demonstrate that our method certifies the robustness of a model and prevents deployment of unreliable models.

## 1 Introduction

The training population typically does not accurately represent what the model will encounter under operation. Model performance has been observed to substantially degrade under distribution shift [16, 28, 69, 80, 53] in speech recognition [52], automated essay scoring [4], and wildlife conservation [11]. Similar trends persist for state-of-the-art NLP and computer vision models [78, 74], even on new data constructed under a near-identical process [57, 66]. 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].

A particularly problematic form of distribution shift comes from embedded power structures in data collection. Data forms the infrastructure on which we build prediction models [30], and they inherit socioeconomic and political inequities against marginalized communities. For example, out of 10,000+ cancer clinical trials the National Cancer Institute funds, less than 5% of participants were non-white [21]. Typical models replicate and perpetuate such bias, and their performance drops significantly on underrepresented groups. Speech recognition systems work poorly for Blacks [52]

*Authors ordered alphabetically.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

and those with minority accents [3]. More generally, model performance degrades across demographic attributes such as race, gender, or age, in facial recognition, video captioning, language identification, and academic recommender systems [41, 46, 17, 72, 79, 19].

Model training typically relies on varied engineering practices. It is crucial to *rigorously certify* model robustness prior to deployment for these heuristic approaches to bear fruit and transform consequential applications. Ensuring that models perform uniformly well across subpopulations is simultaneously critical for reliability, fairness, satisfactory user experience, and long-term business goals. While a natural approach is to evaluate performance across a small set of groups, 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 varies across demographic characteristics and 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. This conservative notion of performance evaluates robustness to unanticipated distribution shifts in $Z$, and automatically accounts for complex intersectionality by virtue of being agnostic to demographic groupings. Formally, let $Z$ be a set of core attributes 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 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 $\theta$ 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 $P_Z$. Before deploying the model $\theta$, we wish to evaluate the worst-case subpopulation performance

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

The worst-case subpopulation performance (2) guarantees uniform performance over subpopulations (1) and has a clear interpretation that can be communicated to diverse stakeholders. The minority proportion $\alpha$ can often be chosen from first principles, e.g., we wish to guarantee uniformly good performance over subpopulations comprising at least $\alpha = 20\%$ of the collected data. Alternatively, 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*. The threshold $\alpha^\star$ provides a *certificate of robustness* on the model $\theta(\cdot)$, guaranteeing that all subpopulations large than $\alpha^\star$ enjoy good performance.

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

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

In Section 2, we propose a two-stage estimation approach where we compute an estimate $\widehat{h}_1(\cdot)$ of the conditional risk $\mu(\cdot)$. Then, we compute a plug-in estimate of the worst-case subpopulation performance under $\widehat{h}_1(\cdot)$ using a dual reformulation of the worst-case problem (2). We show several theoretical guarantees for our estimator 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 for the model class estimating the conditional risk (3).

In some applications, it may be natural to define $Z$ using images or natural languages describing the input and use deep networks to predict the conditional risk (3). As the complexity term $\mathfrak{Comp}_n(\mathcal{H})$ becomes prohibitively large in this case [10, 86], our second result (Section 3.2) shows data-dependent

*dimension-free* concentration of our two-stage estimator: our bound only depends on the complexity of the model class $\mathcal{H}$ through the out-of-sample error for estimating the conditional risk (3). This error can be made small using overparameterized deep networks, allowing us to estimate the conditional risk (3) using even the largest deep networks and still obtain a theoretically principled upper confidence bound on the worst-case subpopulation performance. Leveraging these guarantees, we develop principled procedures for estimating the certificates of robustness $\alpha^\star$ in Section 3.3.

In Section 4, we demonstrate the effectiveness of our procedure on real data. By evaluating model robustness under subpopulation shifts, our methods allow the selection of robust models before deployment as we illustrate using the recently proposed CLIP model [62].

**Related work.** The long line of works on distributionally robust optimization (DRO) aims to *train models* to perform well under distribution shifts. Previous approaches considered finite-dimensional worst-case regions such as constraint sets [29, 39, 5] and those based on notions of distances for probability measures such as $f$-divergences [12, 13, 56, 55, 60, 33, 32], Levy-Prokhorov [34], Wasserstein distances [35, 73, 15, 37, 14, 82], and integral probability metrics based on reproducing kernels [77, 87]. The distribution shifts considered in these approaches are often 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 proposed 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 pre-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 the worst-case conditional risk $\mathbb{E}[\ell(\theta(X); Y) \mid Z]$ do not scale to deep networks that can overfit to the training data [70]. By contrast, our evaluation perspective aims to take advantage of the rapid progress in deep learning. We build scalable evaluation methods that apply to arbitrary models, which allows leveraging state-of-the-art engineered approaches for training $\theta(\cdot)$. Our narrower focus on evaluation allows us to provide convergence rates that scale advantageously with the dimension of $Z$, compared to the nonparametric $O_p(n^{-1/d})$ rates for training [31]. Recently, Jeong and Namkoong [48] studied a similar notion of worst-case subpopulation performance in causal inference.

Our notion of worst-case subpopulation performance is also related to the by now vast literature on fairness in ML. We give a necessarily abridged discussion and refer readers to Barocas et al. [8] and Corbett-Davies and Goel [27] for a comprehensive treatment. A large body of work studies *equalizing* a notion of performance over fixed, pre-defined demographic groups for *classification tasks* [24, 36, 7, 43, 51, 84]. Kearns et al. [49, 50], Hébert-Johnson et al. [45] 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 [64, 65], where we consider the 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 [64]. Similarly, Hashimoto et al. [44] 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 approaches to benchmarking under distribution shifts [80, 66, 74, 53, 71, 78, 57] as our procedure does not require out-of-distribution data. Since good performance on a particular distribution shift does not necessitate robustness, we evaluate models using the worst-case subpopulation performance (2).

## 2 Approach

We begin by contrasting our approach to standard alternatives that consider pre-defined, fixed demographic groups [59]. 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 [26], a common anti-coagulant (blood thinner). Taking the best prediction model for the optimal dosage on this dataset based on genetic, demographic and clinical factors [26], we present the squared error on the root dosage. In Figure 1, when age and race are considered *simultaneously* instead of *separately*, subpopulation performance vary significantly across intersectional groups.

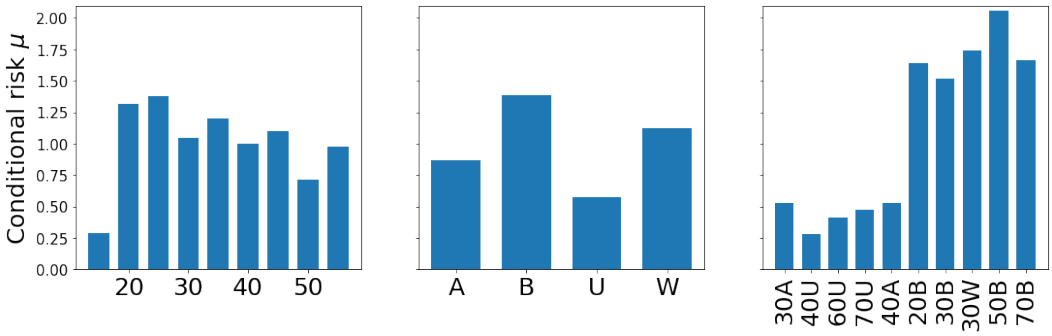


**Figure 1.** Conditional risk $\mu(Z) = \mathbb{E}[(Y - \theta(X))^2 \mid Z]$. Here $Z$ = *age* on the left panel, $Z$ = *race* in the center, and $Z$ = (*age*, *race*) on the right. A = Asian, B = Black, U = Unknown, W = White.

The worst-case subpopulation performance (2) automatically accounts for latent intersectionality. It is agnostic to demographic groupings and allows considering infinitely many subpopulations that represent at least $\alpha$-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 of the input $X$. The choice of $Z$—and subsequent worst-case subpopulation performance (2) of the conditional risk $\mu(Z) = \mathbb{E}[\ell(\theta(X); Y) \mid Z]$—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 $\alpha$ should be informed by domain knowledge—desired robustness of the system—and the dataset size relative to the complexity of $Z$. Often, proxy groups can be used for selecting $\alpha$. If we wish to ensure good performance over patients of all races aged 50 years or older, we can choose $\alpha$ to be the proportion of the least represented $(race, age \geq 50)$ group—this leads to $\alpha = 5\%$ in the Warfarin data. The corresponding worst-case subpopulation performance (2) guarantees good performance over all groups of similar size.

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 $\theta$ is brittle against even majority subpopulations; if it is sufficiently small, then $\theta$ performs well on underrepresented subpopulations.

We now 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 $\theta(\cdot)$.

**Dual reformulation** 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 $\mathsf{W}_\alpha(\theta) = \mathsf{W}_\alpha(\mu)$).

**Lemma 1** (Shapiro et al. [75, Example 6.19] and Rockafellar and Uryasev [67]). *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 $h(Z)$ [68, Theorem 10]. The dual (5) hence shows $\mathsf{W}_\alpha(\theta)$ is a tail-average of $\mu(Z)$, a popular risk measure known as the conditional value-at-risk (CVaR) in portfolio optimization [67].

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

1: INPUT: Subpopulation size $\alpha$, model class $\mathcal{H}$, samples $S_1$ and $S_2$
2: On $S_1$, solve $\widehat{h}_1 \in \operatorname{argmin}_{h \in \mathcal{H}} \frac{1}{|S_1|} \sum_{i \in S_1} (\ell(\theta(X_i); Y_i) - h(Z_i))^2$.
3: On $S_2$, compute the plug-in estimator $\widehat{\mathsf{W}}_\alpha(\widehat{h}_1) = \inf_\eta \left\{ \frac{1}{\alpha |S_2|} \sum_{i \in S_2} \left[ \widehat{h}_1(Z_i) - \eta \right]_+ + \eta \right\}$.

**Two-stage procedure** A key remaining challenge in estimating $\mathsf{W}_\alpha(\theta)$ is that we can only observe losses $\ell(\theta(X_i); Y_i)$ and never observe the conditional risk $\mu(\cdot)$ (3). We propose 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 $\widehat{h}_1(Z)$ of the conditional risk $\mu(Z)$, using any model class $\mathcal{H}$ (class of mappings $\mathcal{Z} \to \mathbb{R}$), by solving an empirical approximation to the loss minimization problem

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

We denote by $h^\star$ a minimizer of (6); for a sufficiently rich model class $\mathcal{H}$, the minimizer is given by $\mu(Z) = \mathbb{E}[\ell(\theta(X); Y) \mid Z]$. 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 $\widehat{h}_1(\cdot)$. In practice, we switch the roles 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

$$\widehat{\alpha} := \inf\{\alpha : \widehat{\mathsf{W}}_\alpha(\widehat{h}_1) \le \bar{\ell}\}. \tag{7}$$

Since $\alpha \mapsto \widehat{\mathsf{W}}_\alpha(\widehat{h}_1)$ 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 Section 3.1, we first give finite-sample convergence at the rate $O_p(\sqrt{\mathfrak{Comp}_n(\mathcal{H})/n})$, where $\mathfrak{Comp}_n(\mathcal{H})$ is the localized Rademacher complexity [9] of the model class $\mathcal{H}$ for estimating the conditional risk $\mu(Z)$. 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 depends only on the out-of-sample error for estimating $\mu(\cdot)$. In particular, the out-of-sample error can grow smaller as $\mathcal{H}$ gets richer, and as a result of hyperparameter tuning and model selection, it is often very small 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 features and $\mathcal{H}$ represent deep networks. 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.

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 $h \in \mathcal{H}$.*

Throughout, we do not stipulate well-specification, meaning that we allow the conditional risk $\mu(\cdot) = \mathbb{E}[\ell(\theta(X); Y) \mid \cdot]$ not to be in the model class $\mathcal{H}$.

### 3.1 Concentration using the 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 $\mathsf{W}_\alpha(\mu) = \mathsf{W}_\alpha(\theta)$), we have

$$\mathsf{W}_\alpha(\mu) - \widehat{\mathsf{W}}_\alpha(\widehat{h}_1) = \underbrace{\mathsf{W}_\alpha(\mu) - \mathsf{W}_\alpha(\widehat{h}_1)}_{(a)\text{: first stage}} + \underbrace{\mathsf{W}_\alpha(\widehat{h}_1) - \widehat{\mathsf{W}}_\alpha(\widehat{h}_1)}_{(b)\text{: second stage}}. \tag{8}$$

To bound term $(b)$, we prove concentration guarantees for estimators of the dual (5) (see Proposition 4 in Appendix A.1). To bound term $(a)$, we use a localized notion of the Rademacher complexity.

Formally, for $\xi_1, \ldots, \xi_n \in \Xi$ and i.i.d. random signs $\varepsilon_i \in \{-1, 1\}$ (independent of $\xi_i$), 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* [9] 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 $\psi_n(r) \geq \mathbb{E}\left[\mathfrak{R}_n\left\{g \in \mathcal{G} : \mathbb{E}[g^2] \leq r\right\}\right]$. (The localized Rademacher complexity itself is sub-root.) The fixed point of $\psi_n$ characterizes generalization guarantees [9, 54].

Let $h^\star$ be the best model in the model class $\mathcal{H}$

$$h^\star := \operatorname*{argmin}_{h \in \mathcal{H}} \mathbb{E}[(\ell(\theta; X, Y) - h(Z))^2].$$

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

$$\psi_{|S_1|}(r) \geq 2\mathbb{E}\left[\mathfrak{R}_{|S_1|}\left\{h \in \mathcal{H} : \mathbb{E}[(h(Z) - h^\star(Z))^2] \leq rB^2/4\right\}\right]. \tag{9}$$

We define $r^\star_{|S_1|}$ as the fixed point of $\psi_{|S_1|}(r)$.

As we show shortly, we bound the estimation error of our procedure using the *square root* of the excess risk in the first-stage problem (6)

$$\mathbb{E}\left[\left(\ell(\theta; X, Y) - \widehat{h}_1(Z)\right)^2 \mid S_1\right] - \mathbb{E}\left[(\ell(\theta; X, Y) - h^\star(Z))^2\right]$$

By using a refined analysis offered by localized Rademacher complexities, we are able to use a fast rate of convergence of $O_p(\mathfrak{Comp}_n(\mathcal{H})/n)$ on the preceding excess risk. In turn, this provides the following $O_p(\sqrt{\mathfrak{Comp}_n(\mathcal{H})/n})$ bound on the estimation error as we prove in Appendix A.2. In the bound, we have made explicit the approximation error term $\|h^\star - \mu\|_{L^2}$. As the model class $\mathcal{H}$ grows richer, there is tension as the approximation error term will shrink, yet the localized Rademacher complexity of $\mathcal{H}$ will grow.

**Theorem 1.** *Let Assumption 1 hold. For some constant $C > 0$, with probability at least $1 - 2\delta$,*

$$\left|\mathsf{W}_\alpha(\theta) - \widehat{\mathsf{W}}_\alpha(\widehat{h}_1)\right| \leq \frac{CB}{\alpha}\left(\sqrt{r^\star_{|S_1|}} + \sqrt{\frac{\log(1/\delta)}{|S_1|}} + \sqrt{\frac{\log(2/\delta)}{|S_2|}}\right) + \frac{1}{\alpha}\|h^\star - \mu\|_{L_2}.$$

If we let $S_1$ be $(1 - 1/k)$-fraction of the data and $S_2$ be the remaining $1/k$-fraction for some 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 (3). For example, when $\mathcal{H}$ is a bounded VC-class [81], it is known that its fixed point satisfy [9, Corollary 3.7]

$$r^\star_{|S_1|} \asymp \log(|S_1|/\mathsf{VC}(\mathcal{H})) \cdot \mathsf{VC}(\mathcal{H})/|S_1|,$$

where $\mathsf{VC}(\cdot)$ is the VC-dimension.

### 3.2 Data-dependent dimension-free concentration

In some applications, it may be natural to model $Z$ as a high-dimensional variable. This may include large subsets of $(X, Y)$, or defining $Z$ using unstructured information such as images or natural languages. In these instances, we may wish to use deep networks as the model class $\mathcal{H}$ for estimating the conditional risk (3). We now provide an alternative concentration result that depends on the size of model class $\mathcal{H}$ only through the out-of-sample error in the first-stage problem (6). We denote for simplicity

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

for any function $h : \mathcal{Z} \to \mathbb{R}$ on any data set $S$. We prove the following result in Appendix A.3.

**Theorem 2.** *Let Assumption 1 hold. For some constant $C > 0$, with probability at least $1 - 3\delta$,*

$$|\mathsf{W}_\alpha(\theta) - \widehat{\mathsf{W}}_\alpha(\widehat{h}_1)| \leq \frac{1}{\alpha}\left(\sqrt{\left[\Delta_{S_2}(\widehat{h}_1) - \Delta_{S_2}(h^\star)\right]_+} + CB\left(\frac{\log(1/\delta)}{|S_2|}\right)^{1/4} + \|h^\star - \mu\|_{L^2}\right).$$

*Moreover, if the model class $\mathcal{H}$ is convex, then $\|h^\star - \mu\|_{L^2}$ can be replaced with $\|h^\star - \mu\|_{L^1}$.*

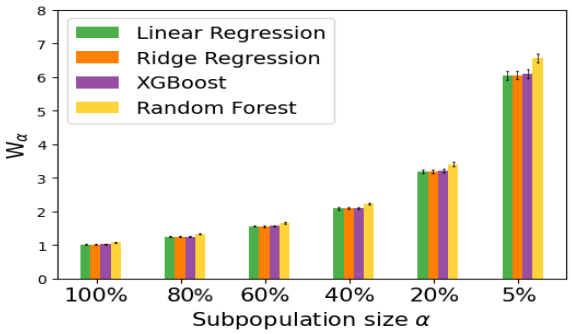


**Figure 2:** Worst-case subpopulation performance $\mathsf{W}_\alpha(\theta)$, where $\mathsf{W}_{1.0}(\theta) = \mathbb{E}[\ell(\theta(X); Y)]$.

Following convention in learning theory, we refer to our data-dependent concentration guarantee *dimension-free*. For overparameterized model classes $\mathcal{H}$ such as deep networks, the localized Rademacher complexity in Theorem 1 becomes prohibitively large [10, 86]. In contrast, the current result can still provide meaningful finite-sample bounds: model selection and hyperparameter tuning provides low out-of-sample performance in practice, and the difference $\Delta_{S_2}(\widehat{h}_1) - \Delta_{S_2}(h^\star)$ can be often made very small. Concretely, it is possible to calculate an upper bound on this term as $\Delta_{S_2}(h^\star)$ is lower bounded by $\min_{h\in\mathcal{H}} \Delta_{S_2}(h)$.

### 3.3 Certificate of robustness

Instead of estimating the worst-case subpopulation performance for a fixed subpopulation size $\alpha$, it may be natural to posit a level of acceptable performance (upper bound $\bar{\ell}$ on the loss) and study $\alpha^\star(\theta)$, the smallest subpopulation size (4) over which the model $\theta(\cdot)$ can guarantee acceptable performance. Our plug-in estimator $\widehat{\alpha}$ given in Eq. (7) enjoys similar concentration guarantees as those given in Sections 3.1, 3.2. The following theorem—whose proof we give in Appendix A.4—states that the true $\alpha^\star(\theta)$ is either close to our estimator $\widehat{\alpha}$ or it is sufficiently small, certifying the robustness of the model against subpopulation shifts.

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

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

*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 $S_2$.*

Our approach simultaneously provides localized Rademacher complexity bounds and dimension-free guarantees. Our bound becomes large as $\underline{\alpha} \to 0$ and we conjecture this to be a fundamental difficulty as the worst-case subpopulation performance (2) focuses on $\alpha$-faction of the data.

## 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 computer vision application based on satellite images [25] where natural distribution shifts were recently studied [53]. In both settings, we illustrate how our worst-case subpopulation approach raises awareness on brittle models without knowledge of out-of-distribution samples. Finally, to verify asymptotic convergence of our proposed two-stage estimator, we present a simultion experiment on a classification task in Appendix C. For all experiments, we use gradient boosted decision trees (package XGBoost [22]) to estimate the conditional risk $\mu(Z) = \mathbb{E}[\ell(\theta(X); Y) \mid Z]$.

### 4.1 Warfarin

Warfarin is one of the most widely used anticoagulant, often prescribed to prevent strokes [26]. 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 through trial and error by clinicians. The dataset comprises of 4,788 patients (after excluding missing data) alongside features representing demographics, genetic markers, medication history, pre-existing conditions, and reason for treatment. Consortium [26] found that 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 perform well on underrepresented groups [20, 63, 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 [26], 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. Moreover, 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 and interpretable, our diagnostic certifies its advantageous performance even on tail subpopulations. However, our diagnostic raises some concerns about poor subpopulation performance: on $\alpha = 5\%$ of the training population, all models suffer prediction error six times worse than the average-case performance.

### 4.2 Functional Map of the World (FMoW)

Satellite images can impact economic and environmental policies globally by allowing large-scale measurements on poverty [1], population changes, deforestation, and economic growth [42]. An automated approach allows analyzing data from remote regions at a relatively low cost and provides continuous monitoring of land usage. Towards this goal, it is critical that the models perform reliably across time and space. We study this problem on the Functional Map of the World (FMoW) dataset [25], where the goal is to predict building / land use categories (62 classes) based on satellite images. Across different models, we observe that performance remains similar either temporally or spatially when each dimension is considered *separately*, but there is substantial variability across intersections of region and year. For a standard *DenseNet ERM* model [47, 53] that achieves near-state-of-the-art performance, we present these trends in Figure 3(a). In Figure 3(b), we observe substantial variability in classwise error rates; there is a varying level of difficulty across different classes. (We observed similar patterns for other models.)

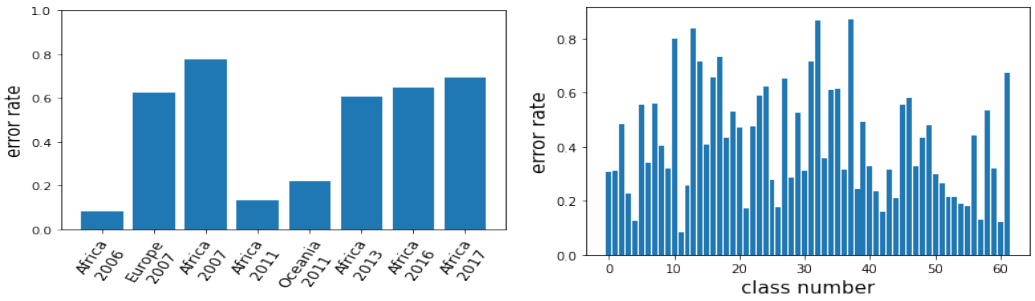


**Figure 3:** For *DenseNet ERM*, spatiotemproal intersectionality (left) and performance by class (right)

We take the perspective of an analyst evaluating prediction models for land usage, based on data collected during 2002-2013. The FMoW dataset provides fertile grounds for demonstrating our method as it includes natural distribution shifts [53], both spatial and temporal. In particular, we demonstrate model robustness on out-of-distribution samples collected in 2016-2018. On validation data collected during 2002-2013, we first evaluate model performance on subpopulations defined across metadata on a satellite image, which consists of (subsets of) {*longitude, latitude, cloud cover, region, year*} and the label $Y$. Then, we observe how our procedure selects models that perform well "in the future" without requiring out-of-distribution data.

We examine a range of different models trained on the FMoW-WILDS training set (collected in 2002-2013, $n$ =76,863) which fall into two broad categories. First, we consider models pre-trained on ImageNet and finetuned on the FMoW training set. These include *DenseNet* models trained using ERM and the recently proposed invariant risk minimization (IRM) framework [6]. We also study the Dual Path Network-68 (*DPN-68*) model with connection paths that enable feature reusage and feature exploration proposed by Chen et al. [23]. We use *DPN-68* trained on FMoW using ERM as reported in [58]. These models all achieve in distribution (ID) accuracy of $\sim 60\%$ on a heldout validation set ("ID val", collected in 2002-2013, $n$ =11,483).

Second, we consider models derived from the recently proposed CLIP model [62], which was trained on large and heterogeneous data sources comprising of 40M image-text pairs using natural language supervision and contrastive losses. The pre-training data for CLIP is 400 times bigger than ImageNet, and Radford et al. [62] have observed that zero-shot applications of CLIP exhibits substantial *relative robustness gains* over other state-of-the-art methods on natural distribution shifts of ImageNet.

However, on the FMoW in-distribution (2002–2013) validation set, zero-shot CLIP only achieves $19.3\%$ accuracy compared to the $60\%$ accuracy of ImageNet pre-trained models. We thus finetune it using satellite images in the FMoW training data. While finetuning substantially improves ID accuracy on FMoW to $70.2\%$, the relative robustness gains of the zero-shot CLIP model severely degrade. To address this problem, Wortsman et al. [85] proposed a weight-space ensembling method (*CLIP WiSE-FT*) where they average the network weights of the zero-shot CLIP model and its finetuned counterpart. These ensembled networks have been observed to exhibit large Pareto improvements in both in-distribution and out-of-distribution accuracy, including on the FMoW dataset.

Motivated by the observed robustness gains, we average the network weights $\theta_0$ of the *CLIP Zeroshot* model and that of *CLIP fine-tuned* $\theta_1$ to generate a new network $(1-\lambda)\theta_0 + \lambda\theta_1$, where $\lambda \in [0, 1]$ controls how much weight is given to the task-specific, fine-tuned model (domain expertise). We select $\lambda = 0.4$ so that the ensembled model (*CLIP WiSE-FT*) achieves similar performance as ImageNet pre-trained counterparts on the in-distribution validation data. To further make models comparable with respect to the cross entropy loss, we calibrate the *CLIP WiSE-FT* model by tuning the temperature parameter so that its average loss on the in-distribution validation set matches the worst average loss of ImageNet pre-trained models (*DenseNet ERM*). See Appendix B for detailed experimental settings and training specifications.

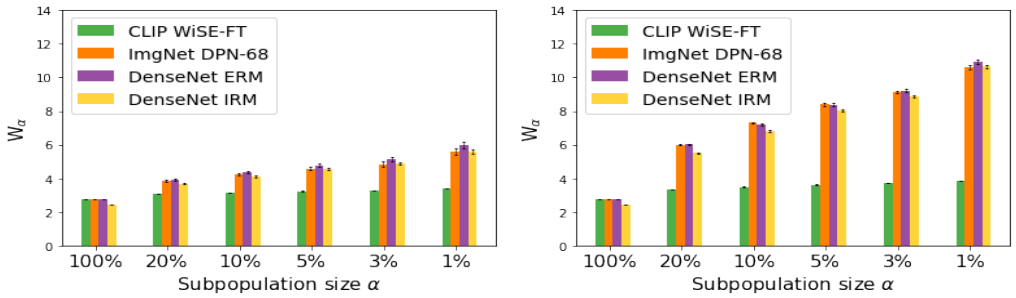


**Figure 4.** Left: $Z$ = (all metadata); Right: $Z$ = (all metadata, Y). Results are averaged over 50 random seeds with error bars corresponding to a 95% confidence interval over the random runs.

We compute estimators of $\mathsf{W}_\alpha$ (Algorithm 1) on the in-distribution validation data (ID val) using the standard cross entropy loss. In Figure 4, we summarize the estimated worst-case subpopulation performances defined over the entire *metadata*, across different subpopulation sizes $\alpha$. First, we note that all models have comparable in-distribution accuracy of $\sim 60\%$ and *DenseNet IRM* has the best average-case cross entropy loss. However, the worst-case subpopulation performance of the ImageNet pre-trained models is substantially worse compared to that of *CLIP WiSE-FT*. This gap grows larger as the subpopulation size $\alpha$ becomes increasingly small. Evaluations on worst-case subpopulations suggest that *CLIP WiSE-FT* exhibits robustness against subpopulation shifts; in contrast, average-case evaluations will select *DenseNet IRM*.

We observe a drastic performance deterioration on tail subpopulations. The inclusion of label information in $Z$ significantly deteriorates worst-case performance, raising concerns about the distributional robustness of all models including changes in the label distribution. In Table 1, we present model performances on the out-of-distribution ("future") data collected during 2016–2018.

| Models | ID, 2002–2013 Accuracy | Loss | OOD, 2016–2018 Accuracy | Loss | Africa Accuracy | Africa Loss |
|---|---|---|---|---|---|---|
| CLIP WiSE-FT | 0.61 | 2.78 | 0.56 | 2.84 | 0.38 | 3.08 |
| DenseNet ERM | 0.61 | 2.78 | 0.53 | 3.50 | 0.33 | 5.41 |
| DenseNet IRM | 0.59 | 2.44 | 0.51 | 2.94 | 0.31 | 4.46 |
| DPN-68 | 0.61 | 2.75 | 0.53 | 3.55 | 0.31 | 5.61 |

**Table 1.** Model performance on ID val and OOD test sets. All models suffer a performance drop on the OOD test set in both accuracy and average loss. The performance degradation is particularly significant on images from Africa. On the OOD data, *CLIP WiSE-FT* outperforms other models both in average accuracy/loss and worst-region accuracy/loss.

All models suffer a significant performance drop under temporal distribution shift, particularly on images collected in Africa where predictive accuracy drops by up to 20 percentage points. *CLIP WiSE-FT* exhibits the most robustness under spatiotemporal shift than any other model, as presaged by evaluations of worst-case subpopulation performance in Figure 4.

A key advantage of our method is the flexibility in the choice of $Z$; the modeler can define granular or coarse subpopulations based on this choice. As defining subpopulations over all metadata can be conservative, we present additional results under $Z =$(*region*, *year*) and $Z =$(*region*, *year*, *label* $Y$) in Appendix B. Instead of incorporating labels as a category, it may be more informative to use the *semantic meaning of each class label*. We generate natural language description of the labels by concatenating each class label with engineered prompts, and pass it to the CLIP text encoder [62] to generate a feature representation for the label. In Appendix B.3, we present evaluation results where we take the feature vector in place of the label $Y$ when defining $Z$.

## 5 Discussion

To ensure models perform reliably under operation, we need to *rigorously* certify their performance under distribution shift prior to deployment. We study the *worst-case subpopulation performance* of a model, a natural notion of model robustness that is easy to communicate with users, regulators, and business leaders. 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 further inform data collection and model improvement by suggesting data collection efforts and model fixes on regions of $\mathcal{Z}$ with high conditional risk (3).

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 in defining a tailored notion of subpopulation is a promising research direction towards operationalizing the concepts put forth in this work. As an example, Srivastava et al. [76] 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. A notable limitation of our approach is that we do not explicitly consider the power differential that often exists between those who deploy the prediction system and those for whom it gets used on. 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] Anonymous. Distributionally robust neural networks. In *Submitted to International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=ryxGuJrFvS`. under review.

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

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

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

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

[10] 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.

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

[12] 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.

[13] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *Mathematical Programming, Series A*, 167(2):235–292, 2018.

[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] M. S. Chen, 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.

[22] 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.

[23] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng. Dual path networks. *arXiv:1707.01629 [cs.CV]*, 2017.

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

[25] 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.

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

[27] 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.

[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*, 49(3):1378–1406, 2021.

[33] J. C. Duchi, P. W. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 2021.

[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] 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.

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

[44] 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.

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

[46] 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.

[47] 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.

[48] S. Jeong and H. Namkoong. Robust causal inference under covariate shift via worst-case subpopulation treatment effect. In *Proceedings of the Thirty Third Annual Conference on Computational Learning Theory*, 2020.

[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] 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`.

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

[57] 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.

[58] 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.

[59] 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.

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

[61] 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.

[62] 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. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[63] 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.

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

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

[66] 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.

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

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

[69] 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.

[70] 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.

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

[72] 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.

[73] 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.

[74] 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.

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

[76] 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.

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

[78] 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.

[79] 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.

[80] 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.

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

[82] 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.

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

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

[85] M. Wortsman, G. Ilharco, M. Li, J. W. Kim, H. Hajishirzi, A. Farhadi, H. Namkoong, and L. Schmidt. Robust fine-tuning of zero-shot models. *arXiv:2109.01903 [cs.CV]*, 2021.

[86] 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.

[87] 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]

# A Proof of finite-sample concentration results

Our results are based on a general concentration guarantee for estimating the dual reformulation (5) for any given $h(Z)$. We give this result in Appendix A.1, and build on it in subsequent proofs of key results. In the following, we use $\lesssim$ to denote inequality up to a numerical constant that may change line by line.

## A.1 Concentration bounds for worst-case subpopulation performance

Since $\ell(\widehat{y}; y) \geq 0$ for losses used in most machine learning problems, we assume that $\mathcal{H}$ consists of nonnegative functions. To show exponential concentration guarantees, we consider sub-Gaussian conditional risk models $h(Z)$. Note the concentration results here are more general than needed for the purpose of proving the main results, because any random variable bounded in $[0, B]$ is inherently sub-Gaussian with paramater $B^2/4$.

**Definition 1.** *A function $h : \mathcal{Z} \to \mathbb{R}$ with $\mathbb{E}|h(Z)| < \infty$ is sub-Gaussian with parameter $\sigma^2$ if*

$$\mathbb{E}\left[\exp\left(\lambda(h(Z) - \mathbb{E}[h(Z)])\right)\right] \leq \exp\left(\frac{\sigma^2\lambda^2}{2}\right) \text{ for all } \lambda \in \mathbb{R}.$$

The sub-Gaussian assumption can be relaxed to sub-exponential random variables, with minor and standard modifications to subsequent results. We omit these results for brevity.

Define a dual plug-in estimator for the worst-case subpopulation performance of $h(Z)$ on $S_2$

$$\widehat{\mathsf{W}}_\alpha(h) = \inf_\eta \left\{ \frac{1}{\alpha|S_2|} \sum_{i \in S_2} \left[\widehat{h}_1(Z_i) - \eta\right]_+ + \eta \right\}. \tag{11}$$

The following result shows that for any sub-Gaussian $h$ that is bounded from below, the plug-in estimator (3) converges at the rate $O_p(|S_2|^{-1/2})$.

**Proposition 4.** *There is a universal constant $C > 0$ such that for all $h \geq 0$ that is sub-Gaussian with parameter $\sigma^2$,*

$$|\widehat{\mathsf{W}}_\alpha(h) - \mathsf{W}_\alpha(h)| \leq \frac{C\sigma}{\alpha}\sqrt{\frac{\log(2/\delta)}{|S_2|}} \text{ with probability at least } 1 - \delta.$$

We prove the proposition in the rest of the subsection. By a judicious application of the empirical process theory, our bounds—which apply to nonnegative random variables—are simpler than existing concentration guarantees for conditional value-at-risk [18, 61].

Our starting point is the following claim, which bounds $|\widehat{\mathsf{W}}_\alpha(h) - \mathsf{W}_\alpha(h)|$ in terms of the suprema of empirical process on $\{z \mapsto [h(z) - \eta]_+ : \eta \geq 0\}$.

**Claim 2.**

$$\left|\widehat{\mathsf{W}}_\alpha(h) - \mathsf{W}_\alpha(h)\right| \leq \frac{1}{\alpha}\sup_{\eta \geq 0}\left|\frac{1}{|S_2|}\sum_{i \in S_2}[h(Z_i) - \eta]_+ - \mathbb{E}\left[h(Z) - \eta\right]_+\right| \tag{12}$$

The crux of this claim is that $\eta$ does not range over $\mathbb{R}$, but rather has a lower bound; the value 0 can be replaced with any almost sure lower bound on $h(Z)$. Deferring the proof of Claim 2 to the end of the subsection, we proceed by bounding the suprema of the empirical process in the preceding display.

We begin by introducing requisite concepts in empirical process theory, which we use in the rest of the proof; we refer readers to van der Vaart and Wellner [81] for a comprehensive treatment. Recall the definition of Orlicz norms, which allows controlling the tail behavior of random variables.

**Definition 2** (Orlicz norms)**.** *Let $\psi : \mathbb{R} \to \mathbb{R}$ be a non-decreasing, convex function with $\psi(0) = 0$. For any random variable $W$, its Orlicz norm $\|W\|_\psi$ is*

$$\|W\|_\psi := \inf\left\{t > 0 : \mathbb{E}\left[\psi\left(\frac{|W|}{t}\right)\right] \leq 1\right\}.$$

**Remark 1:** From Markov's inequality, we have

$$\mathbb{P}(|W| > t) \leq \mathbb{P}\left(\psi\left(\frac{|W|}{\|W\|_\psi}\right) \geq \psi\left(\frac{t}{\|W\|_\psi}\right)\right) \leq \psi\left(\frac{t}{\|W\|_\psi}\right)^{-1}.$$

For $\psi_p(s) = e^{s^p} - 1$, a similar argument yields

$$\mathbb{P}(|W| > t) \leq 2\exp\left(-t^p / \|W\|_{\psi_p}^p\right). \tag{13}$$

$\diamondsuit$

A sub-Gaussian random variable $h(Z)$ with parameter $\sigma^2$ has bounded Orlicz norm $\|h(Z)\|_{\psi_2} \leq 2\sigma$ (see, for example, Wainwright [83, Section 2.4] and van der Vaart and Wellner [81, Lemma 2.2.1]).

**Remark 2:** The converse also holds: for $W$ such that $\mathbb{P}(|W| > t)] \leq c_1 \exp(-c_2 t^p)$ for all $t$, and constants $c_1, c_2 > 0$ and $p \geq 1$, Fubini gives

$$\mathbb{E}\left[\exp\left(\frac{|W|^p}{t^p}\right) - 1\right] = \mathbb{E}\left[\int_0^{|W|^p} t^{-1/p}\exp(t^{-1/p}s)ds\right] = \int_0^\infty \mathbb{P}(|W|^p > s)t^{-1/p}\exp\left(t^{-1/p}s\right)ds.$$

Using the tail probability bound, the preceding display is bounded by

$$c_1\int_0^\infty \exp(-c_2 s)t^{-1/p}\exp(t^{-1/p}s)ds = \frac{c_1 t^{-1/p}}{c_2 - t^{-1/p}}.$$

So the Orlicz norm $\|W\|_{\psi_p}$ is bounded by $\left(\frac{1+c_1}{c_2}\right)^{1/p}$. $\diamondsuit$

In the following, we let $W$ be the right hand side of the bound (12), and control its Orlicz norm $\|W\|_{\psi_2}$ using Dudley's entropy integral [81]. We use the standard notion of the covering number. For a vector space $\mathcal{V}$, let $V \subset \mathcal{V}$ be a collection of vectors. Letting $\|\cdot\|$ be a norm on $\mathcal{V}$, a collection $\{v_1, \ldots, v_N\} \subset \mathcal{V}$ is an *$\epsilon$-cover* of $\mathcal{V}$ if for each $v \in \mathcal{V}$, there is a $v_i$ satisfying $\|v - v_i\| \leq \epsilon$. The *covering number* of $V$ with respect to $\|\cdot\|$ is

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

For a collection $\mathcal{F}$ of functions $f : \mathcal{Z} \to \mathbb{R}$, let $F$ be its envelope function such that $|f(z)| \leq F(z)$ for all $z \in \mathcal{Z}$. The following result controls the suprema of empirical processes using the (uniform) metric entropy. The result is based on involved chaining arguments [81, Section 2.14].

**Lemma 3** (van der Vaart and Wellner [81, Theorem 2.14.1 and 2.14.5])**.**

$$\begin{aligned}&\sqrt{|S_2|}\left\|\sup_{f\in\mathcal{F}}\left|\frac{1}{|S_2|}\sum_{i\in S_2} f(Z_i) - \mathbb{E}f(Z)\right|\right\|_{\psi_2} \\ &\lesssim \|F\|_{\psi_2} + \|F\|_{L^2(P)}\sup_Q\int_0^1\sqrt{1 + \log N(\epsilon\|F\|_{L^2(Q)}, \mathcal{F}, L^2(Q))}d\epsilon,\end{aligned}$$

*where the supremum is over all discrete probability measures $Q$ such that $\|F\|_{L^2(Q)} > 0$.*

Evidently, $F(z) = [h(z)]_+ = h(z)$ is an envelope function for the following class of functions

$$\mathcal{F} = \{z \mapsto [h(z) - \eta]_+ : \eta \geq 0\}.$$

Using the tail probability bound (13), we conclude

$$\begin{aligned}&\sup_{\eta\geq 0}\left|\frac{1}{|S_2|}\sum_{i\in S_2}[h(Z_i) - \eta]_+ - \mathbb{E}[h(Z) - \eta]_+\right| \\ &\lesssim \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{|S_2|}}\left(\|F\|_{\psi_2} + \|F\|_2\sup_Q\int_0^1\sqrt{1 + \log N(\epsilon\|F\|_{L^2(Q)}, \mathcal{F}, L^2(Q))}d\epsilon\right),\end{aligned}$$

with probability at least $1 - \delta$.

Since we have $\|F\|_{L^2(P)} \leq \|F\|_{\psi_2} \lesssim \sigma$, it now suffices to show that the above uniform metric entropy is bounded by a universal constant. We use the standard notion of VC-dimension [81, Chapter 2.6, page 135].

**Lemma 4** (van der Vaart and Wellner [81, Theorem 2.6.7]). *Let $\mathsf{VC}(\mathcal{F})$ be the VC-dimension of the colelction of subsets $\{(z,t) : t < f(x)\}$ for $f \in \mathcal{F}$. For any probability measure $Q$ such that $\|F\|_{L^2(Q)} > 0$ and $0 < \epsilon < 1$, we have*

$$N(\epsilon \|F\|_{L^2(Q)}, \mathcal{F}, L^2(Q)) \lesssim \mathsf{VC}(\mathcal{F})(16e)^{\mathsf{VC}(\mathcal{F})} \left(\frac{1}{\epsilon}\right)^{2(\mathsf{VC}(\mathcal{F})-1)}.$$

Translations of a monotone function on $\mathbb{R}$ has VC-dimension 2.

**Lemma 5** (van der Vaart and Wellner [81, Theorem 2.6.16]). *The class of functions $\mathcal{F}' = \{z \mapsto [h(z) - \eta]_+ : \eta \in \mathbb{R}\}$ has VC-dimension $\mathsf{VC}(\mathcal{F}') = 2$.*

From Lemmas 4 and 5, we conclude that for the function class $\mathcal{F} = \{z \mapsto [h(z) - \eta]_+ : \eta \geq 0\}$, the uniform metric entropy

$$\sup_Q \int_0^1 \sqrt{1 + \log N(\epsilon \|F\|_{L^2(Q)}, \mathcal{F}, L^2(Q))} d\epsilon$$

is bounded by a universal constant. This gives our desired result.

**Proof of Claim 2** To show the bound (12), we use the dual reformulation for both $\mathsf{W}_\alpha(h)$ and its empirical approximation $\widehat{\mathsf{W}}_\alpha(h)$ on $S_2$. For any probability measure $P$, recall two different definitions of the quantile of $h(Z)$

$$P^{-1}_{1-\alpha}(h(Z)) := \inf\{t : \mathbb{P}_Z(h(Z) \leq t) \geq 1 - \alpha\}$$
$$P^{-1}_{1-\alpha,+}(h(Z)) := \inf\{t : \mathbb{P}_Z(h(Z) \leq t) > 1 - \alpha\}.$$

We call $P^{-1}_{1-\alpha,+}(h(Z))$ the upper $(1-\alpha)$-quantile. The two values characterize the optimal solution set of the dual problem (5); they are identical when $h(Z)$ has a positive density at $P^{-1}_{1-\alpha}(h(Z))$.

**Lemma 6** (Rockafellar and Uryasev [68, Theorem 10]). *For any probability measure $P$ such that $h(Z) \geq 0$ $P$-a.s. and $\mathbb{E}_P[h(Z)_+] < \infty$, we have*

$$[P^{-1}_{1-\alpha}(h(Z)), P^{-1}_{1-\alpha,+}(h(Z))] = \operatorname*{argmin}_{\eta \in \mathbb{R}} \left\{ \frac{1}{\alpha} \mathbb{E}_P [h(Z) - \eta]_+ + \eta \right\}.$$

Since $P$ was an arbitrary measure in Lemmas 1 and 6, identical results follow for the empirical distribution on $S_2$. Hence, we have

$$\begin{aligned}
\left|\widehat{\mathsf{W}}_\alpha(h) - \mathsf{W}_\alpha(h)\right| &= \left| \inf_{\eta \in \mathbb{R}} \left\{ \frac{1}{\alpha|S_2|} \sum_{i \in S_2} [h(Z_i) - \eta]_+ + \eta \right\} - \inf_{\eta \in \mathbb{R}} \left\{ \frac{1}{\alpha} \mathbb{E} [h(Z) - \eta]_+ + \eta \right\} \right| \\
&= \left| \inf_{\eta \geq 0} \left\{ \frac{1}{\alpha|S_2|} \sum_{i \in S_2} [h(Z_i) - \eta]_+ + \eta \right\} - \inf_{\eta \geq 0} \left\{ \frac{1}{\alpha} \mathbb{E} [h(Z) - \eta]_+ + \eta \right\} \right|
\end{aligned}$$

where we used Lemma 6 to restrict the feasible region in the last equality. The preceding display is then bounded by

$$\sup_{\eta \geq 0} \left| \frac{1}{\alpha|S_2|} \sum_{i \in S_2} [h(Z_i) - \eta]_+ + \eta - \frac{1}{\alpha} \mathbb{E} [h(Z) - \eta]_+ - \eta \right|.$$

### A.2 Proof of Theorem 1

We abuse notation and use $C$ for a numerical constant that may change line to line. From the decomposition (8), it suffices to bound term $(a)$ and term $(b)$ separately. Since $\widehat{h}_1(\cdot)$ is trained on a sample $S_1$ independent from $S_2$ used to estimate the worst-case subpopulation performance (Eq. (11)), we can directly apply Proposition 4 to bound term $(b)$. Recalling that any bounded random

variable random variable taking values in $[0, B]$ is sub-Gaussian with parameter $B^2/4$, Proposition 4 implies

$$\left|\widehat{\mathsf{W}}_\alpha(\widehat{h}_1) - \mathsf{W}_\alpha(\widehat{h}_1)\right| \leq \frac{CB}{\alpha}\sqrt{\frac{\log(2/\delta)}{|S_2|}} \text{ with probability at least } 1-\delta.$$

To bound term $(a)$ in the decomposition (8), we note

$$\begin{aligned}\left|\mathsf{W}_\alpha(\widehat{h}_1) - \mathsf{W}_\alpha(\mu)\right| &\leq \frac{1}{\alpha}\sup_\eta \left|\mathbb{E}\left[\left[\widehat{h}_1(Z)-\eta\right]_+ \mid S_1\right] - \mathbb{E}\left[\mu(Z)-\eta\right]_+\right| \\ &\leq \frac{1}{\alpha}\mathbb{E}\left[\left|\widehat{h}_1(Z)-\mu(Z)\right| \mid S_1\right] \\ &\leq \frac{1}{\alpha}\sqrt{\mathbb{E}\left[\left(\widehat{h}_1(Z)-\mu(Z)\right)^2 \mid S_1\right]} = \frac{1}{\alpha}\sqrt{\mathsf{err}(\mathcal{H}, S_1)},\end{aligned}$$

where the first inequality follows from the dual (5), the second inequality follows from the non-expansiveness of the function $[\cdot]_+$, the last inequality uses Holder inequality, and we define the generalization error for the first-stage estimation problem (6) based on $S_1$,

$$\begin{aligned}\mathsf{err}(\mathcal{H}, S_1) :=& \mathbb{E}\left[\left(\mu(Z)-\widehat{h}_1(Z)\right)^2 \mid S_1\right] \\ =& \mathbb{E}\left[(\ell(\theta(X);Y)-\widehat{h}_1(Z))^2 \mid S_1\right] - \mathbb{E}(\ell(\theta(X);Y)-\mu(Z))^2 \\ =& \mathbb{E}\left[(\ell(\theta(X);Y)-\widehat{h}_1(Z))^2 \mid S_1\right] - \mathbb{E}(\ell(\theta(X);Y)-h^\star(Z))^2 + \mathbb{E}\left(\mu(Z)-h^\star(Z)\right)^2\end{aligned}$$

We use the following concentration result based on the localized Rademacher complexity [9].

**Lemma 7** (Bartlett et al. [9, Corollary 5.3])**.** *Let Assumption 1 hold. Then, with probability at least* $1-\delta$*,*

$$\mathbb{E}\left[(\ell(\theta(X);Y)-\widehat{h}_1(Z))^2 \mid S_1\right] - \mathbb{E}(\ell(\theta(X);Y)-h^\star(Z))^2 \leq CB^2\left(r^\star_{|S_1|} + \frac{\log(1/\delta)}{|S_1|}\right).$$

Using $\sqrt{a+b+c} \leq \sqrt{a}+\sqrt{b}+\sqrt{c}$ for $a, b, c \geq 0$, we have the desired result.

### A.3 Proof of Theorem 2

Instead of the decomposition (8) we use for Theorem 1, we use an alterantive form

$$\widehat{\mathsf{W}}_\alpha(\widehat{h}_1) - \mathsf{W}_\alpha(\mu) = \underbrace{\widehat{\mathsf{W}}_\alpha(\widehat{h}_1) - \widehat{\mathsf{W}}_\alpha(\mu)}_{(a)\text{:first stage}} + \underbrace{\widehat{\mathsf{W}}_\alpha(\mu) - \mathsf{W}_\alpha(\mu)}_{(b)\text{:second stage}} \tag{14}$$

Term $(b)$ can be bounded using Proposition 4 as before. Without assuming $\mu \in \mathcal{H}$, recall that any bounded random variable taking values in $[0, B]$ is sub-Gaussian with parameter $B^2/4$, so Proposition 4 yields

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

It remains to bound term $(a)$. Our starting point is the bound

$$\begin{aligned}\left|\widehat{\mathsf{W}}_\alpha(\widehat{h}_1) - \widehat{\mathsf{W}}_\alpha(\mu)\right| &\leq \frac{1}{\alpha}\sup_\eta \left|\frac{1}{|S_2|}\sum_{i\in S_2}\left(\left[\widehat{h}_1(Z_i)-\eta\right]_+ - \left[\mu(Z_i)-\eta\right]_+\right)\right| \\ &\leq \frac{1}{\alpha|S_2|}\sum_{i\in S_2}|\widehat{h}_1(Z_i)-\mu(Z_i)| \leq \frac{1}{\alpha}\left(\frac{1}{|S_2|}\sum_{i\in S_2}\left(\widehat{h}_1(Z_i)-\mu(Z_i)\right)^2\right)^{1/2}.\end{aligned} \tag{15}$$

Denoting the residuals by $\zeta := \ell(\theta(X);Y) - \mu(Z)$ and $\zeta_i := \ell(\theta(X_i);Y_i) - \mu(Z_i)$ for all $i \in S_2$, we have the identity

$$\frac{1}{|S_2|}\sum_{i\in S_2}\left(\widehat{h}_1(Z_i) - \mu(Z_i)\right)^2 = \Delta_{S_2}(\widehat{h}_1) - \Delta_{S_2}(\mu) + \frac{2}{|S_2|}\sum_{i\in S_2}\zeta_i\left(\widehat{h}_1(Z_i) - \mu(Z_i)\right)$$
$$= \left[\Delta_{S_2}(\widehat{h}_1) - \Delta_{S_2}(h^\star)\right] + \left[\Delta_{S_2}(h^\star) - \Delta_{S_2}(\mu)\right] + \frac{2}{|S_2|}\sum_{i\in S_2}\zeta_i\left(\widehat{h}_1(Z_i) - \mu(Z_i)\right) \tag{16}$$

First notice by definition of $\mu$ that $\zeta$ has conditional mean $\mathbb{E}[\zeta \mid Z] = 0$. Hence, conditional on $S_1$, $\mathbb{E}\left[\zeta\left(\widehat{h}_1(Z) - \mu(Z)\right) \mid S_1\right] = 0$. Since $\zeta_i\left(\widehat{h}_1(Z_i) - \mu(Z_i)\right)$ are bounded in $[-B^2, B^2]$ and i.i.d. conditional on $S_1$, Hoeffding inequality [83, Ch. 2] yields

$$\frac{1}{|S_2|}\sum_{i\in S_2}\zeta_i\left(\widehat{h}_1(Z_i) - \mu(Z_i)\right) \le B^2\sqrt{\frac{2\log(1/\delta)}{|S_2|}} \text{ with probability at least } 1-\delta. \tag{17}$$

Similarly, Hoeffding inequality implies with probability at least $1-\delta$,

$$\Delta_{S_2}(h^\star) - \Delta_{S_2}(\mu) \le \mathbb{E}(\ell(\theta(X);Y) - h^\star(Z))^2 - \mathbb{E}(\ell(\theta(X);Y) - \mu(Z))^2 + 2B^2\sqrt{\frac{2\log(1/\delta)}{|S_2|}}. \tag{18}$$

Note the definition of the conditional risk $\mu(Z) = \mathbb{E}[\ell(\theta(X);Y) \mid Z]$ implies

$$\mathbb{E}(\ell(\theta(X);Y) - h^\star(Z))^2 - \mathbb{E}(\ell(\theta(X);Y) - \mu(Z))^2 = \mathbb{E}(h^\star(Z) - \mu(Z))^2 = \|h^\star - \mu\|_{L^2}^2 .$$

Hence, on the event where inequalities (17) and (18) hold (with probability at least $1-2\delta$),

$$\frac{1}{|S_2|}\sum_{i\in S_2}\left(\widehat{h}_1(Z_i) - \mu(Z_i)\right)^2 \le \Delta_{S_2}(\widehat{h}_1) - \Delta_{S_2}(h^\star) + \|h^\star - \mu\|_{L^2}^2 + 4B^2\sqrt{\frac{2\log(1/\delta)}{|S_2|}}.$$

Noticing $\sqrt{a+b+c} \le \sqrt{[a]_+} + \sqrt{b} + \sqrt{c}$ for $b, c \ge 0$, we obtain the first result.

### A.3.1 Proof of Theorem 2 with a convex model class

Now assume further that $\mathcal{H}$ is convex. We adopt an alternative three-element decomposition:

$$\mathsf{W}_\alpha(\mu) - \widehat{\mathsf{W}}_\alpha(\widehat{h}_1) = \underbrace{\widehat{\mathsf{W}}_\alpha(\widehat{h}_1) - \widehat{\mathsf{W}}_\alpha(h^\star)}_{(a)\text{: first stage}} + \underbrace{\widehat{\mathsf{W}}_\alpha(h^\star) - \mathsf{W}_\alpha(h^\star)}_{(b)\text{: second stage}} + \underbrace{\mathsf{W}_\alpha(h^\star) - \mathsf{W}_\alpha(\mu)}_{(c)\text{: approximation error}}. \tag{19}$$

The approximation error term $(c)$ can be bounded by

$$|\mathsf{W}_\alpha(h^\star) - \mathsf{W}_\alpha(\mu)| \le \frac{1}{\alpha}\sup_\eta\left|\mathbb{E}\left[h^\star(Z) - \eta\right]_+ - \mathbb{E}\left[\mu(Z) - \eta\right]_+\right|$$
$$\le \frac{1}{\alpha}\mathbb{E}\left|h^\star(Z) - \mu(Z)\right| = \frac{1}{\alpha}\left\|h^\star - \mu\right\|_{L^1}.$$

The second-stage error term $(b)$ can be bounded using Proposition 4 by

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

The first-stage error term $(a)$ can be bounded, similarly to Equation (15) by

$$\left|\widehat{\mathsf{W}}_\alpha(\widehat{h}_1) - \widehat{\mathsf{W}}_\alpha(h^\star)\right| \le \frac{1}{\alpha}\sup_\eta\left|\frac{1}{|S_2|}\sum_{i\in S_2}\left(\left[\widehat{h}_1(Z_i) - \eta\right]_+ - \left[h^\star(Z_i) - \eta\right]_+\right)\right|$$
$$\le \frac{1}{\alpha|S_2|}\sum_{i\in S_2}\left|\widehat{h}_1(Z_i) - h^\star(Z_i)\right| \le \frac{1}{\alpha}\left(\frac{1}{|S_2|}\sum_{i\in S_2}\left(\widehat{h}_1(Z_i) - h^\star(Z_i)\right)^2\right)^{1/2}. \tag{20}$$

Again, we have the identity

$$\frac{1}{|S_2|}\sum_{i\in S_2}(\widehat{h}_1(Z_i)-h^\star(Z_i))^2 = \Delta_{S_2}(\widehat{h}_1)-\Delta_{S_2}(h^\star)+\frac{2}{|S_2|}\sum_{i\in S_2}(\ell(\theta(X_i);Y_i)-h^\star(Z_i))(\widehat{h}_1(Z_i)-h^\star(Z_i)).$$

Since we assume the model class $\mathcal{H}$ is convex and $\widehat{h}_1 \in \mathcal{H}$, the first-order condition of $h^\star \in \arg\min_{h\in\mathcal{H}} \mathbb{E}(\ell(\theta(X);Y)-h(Z))^2$ gives

$$\mathbb{E}[(\ell(\theta(X);Y)-h^\star(Z))(\widehat{h}_1(Z)-h^\star(Z)) \mid Z, S_1] \le 0,$$

so Hoeffding inequality implies with probability at least $1-\delta$,

$$\frac{1}{|S_2|}\sum_{i\in S_2}(\ell(\theta(X_i);Y_i)-h^\star(Z_i))(\widehat{h}_1(Z_i)-h^\star(Z_i)) \le B^2\sqrt{\frac{2\log(1/\delta)}{|S_2|}}. \tag{21}$$

Hence,

$$\frac{1}{|S_2|}\sum_{i\in S_2}(\widehat{h}_1(Z_i)-h^\star(Z_i))^2 \le \Delta_{S_2}(\widehat{h}_1)-\Delta_{S_2}(h^\star)+2B^2\sqrt{\frac{2\log(1/\delta)}{|S_2|}} \text{ w.p. } \ge 1-\delta.$$

Using $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ for all $a,b\ge 0$, we obtain the desired result.

### A.4 Proof of Theorem 3

For ease of notation, we suppress any dependence on the prediction model $\theta(X)$ under evaluation. Consider any $\alpha_1,\alpha_2 \in (0,1]$ with $\alpha_1 < \alpha_2$. For convenience denote $\xi_1 := \widehat{P}^{-1}_{1-\alpha_1}(\widehat{h}(Z))$ and $\xi_2 := \widehat{P}^{-1}_{1-\alpha_2}(\widehat{h}(Z))$, so $\xi_1 \ge \xi_2$ and $\widehat{\mathsf{W}}_{\alpha_1}(\widehat{h}) \ge \widehat{\mathsf{W}}_{\alpha_2}(\widehat{h})$. Denote by $\widehat{\mathbb{P}}$ the empirical probability measure induced by $(Z_i : i \in S_2)$. Notice that

$$\begin{aligned}
\widehat{\mathbb{E}}[\widehat{h}(Z)-\xi_2]_+ - \widehat{\mathbb{E}}[\widehat{h}(Z)-\xi_1]_+ &= \widehat{\mathbb{E}}[\widehat{h}(Z)-\xi_2; \widehat{h}(Z)>\xi_2] - \widehat{\mathbb{E}}[\widehat{h}(Z)-\xi_1; \widehat{h}(Z)\ge\xi_1] \\
&= \underbrace{\widehat{\mathbb{E}}[\widehat{h}(Z)-\xi_1; \xi_2<\widehat{h}(Z)<\xi_1]}_{\le 0} + (\xi_1-\xi_2)\underbrace{\widehat{\mathbb{P}}(\widehat{h}(Z)>\xi_2)}_{\le\alpha_2} \\
&\le (\xi_1-\xi_2)\alpha_2.
\end{aligned}$$

Hence, by Lemma 6,

$$\widehat{\mathsf{W}}_{\alpha_1}(\widehat{h})-\widehat{\mathsf{W}}_{\alpha_2}(\widehat{h}) = \left(\frac{\widehat{\mathbb{E}}[\widehat{h}(Z)-\xi_1]_+}{\alpha_1}+\xi_1\right)-\left(\frac{\widehat{\mathbb{E}}[\widehat{h}(Z)-\xi_2]_+}{\alpha_2}+\xi_2\right) \ge \frac{\widehat{\mathbb{E}}[\widehat{h}(Z)-\xi_1]_+}{\alpha_1\alpha_2}(\alpha_2-\alpha_1),$$

meaning

$$|\alpha_1-\alpha_2| \le \frac{\alpha_1\alpha_2|\widehat{\mathsf{W}}_{\alpha_1}(\widehat{h})-\widehat{\mathsf{W}}_{\alpha_2}(\widehat{h})|}{\widehat{\mathbb{E}}[\widehat{h}(Z)-\widehat{P}^{-1}_{1-\alpha_1}(\widehat{h}(Z))]_+}.$$

Now suppose $\alpha^\star \ge \underline{\alpha}$. Notice the boundedness of $h(Z)$ implies $\mathsf{W}_\alpha$ and $\widehat{\mathsf{W}}_\alpha$ are continuous and nonincreasing in $\alpha$, so the definitions (4) and (7) imply $\mathsf{W}_{\alpha^\star}(\theta) = \bar{\ell} = \widehat{\mathsf{W}}_{\widehat{\alpha}}(\widehat{h})$. Plugging $\widehat{\alpha}$ and $\alpha^\star$ into the inequality above, we know with probability at least $1-\delta$,

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

## B Additional experiment details

In this section, we present additional experiments for the Functional Map of the World (FMoW) dataset. Due to the ever-changing nature of aerial images and the uneven availability of data from different regions, it is imperative that ML models maintain good performance under temporal (learn from the past and generalize to future) and spatial distribution shifts (learn from one region and generalize to another). Without having access to the out-of-distribution samples, our diagnostic raises awareness on brittleness of model performance against subpopulation shifts.

### B.1 Dataset Description

The original Functional Map of the World (FMoW) dataset by [25] consists of over 1 million images from over 200 countries. We use a variant, FMoW-WILDS, proposed by Koh et al. [53], which temporally groups observations to simulate distribution shift across time. Each data point includes an RGB satellite image $x$, and a corresponding label $y$ on the land / building use of the image (there are 62 different classes). FMoW-WILDS splits data into non-overlapping time periods: we train and validate models $\theta(\cdot)$ on data collected from years 2002-2013, and simulate distribution shift by looking at data collected during 2013-2018. Data collected during 2002-2013 ("in-distribution") is split into training ($n$ =76,863), validation ($n$ =19,915), and test ($n$ =11,327). Data collected during 2013-2018 ("out-of-distribution") is split into two sets: one consisting of observations from years 2013-2016 ($n$ =19,915), and another consisting of observations from years 2016-2018 ($n$ =22,108). All data splits contain images from a diverse array of geographic regions. We evaluate the worst-case subpopulation performance on in-distribution validation data, and study model performance under distribution shift on data after 2016.

### B.2 Models Evaluated

We consider *DenseNet* models as reported by Koh et al. [53], including the vanilla empirical risk minimization (ERM) model and models trained with robustness interventions (IRM [6] method; Koh et al. [53] notes that ERM's performance closely match or outperform "robust" counterparts even under distribution shift. We also evaluate ImageNet pre-trained *DPN*-68 model from Miller et al. [58]. As separate experiments, we also consider *ResNet*-18 and *VGG*-11 from Miller et al. [58], and the results are reported in B.6.

CLIP (Contrastive Language-Image Pre-training) is a newly proposed model pre-trained on 400M image-text pairs, and has been shown to exhibit strong zero-shot performance on out-of-distribution samples [62]. Although not specifically designed for classification tasks, CLIP can be used for classification by predicting the class whose encoded text is the closest to the encoded image. We consider the weight-space ensembled *CLIP WiSE* models proposed in [85] as it is observed that these models exhibit robust behavior on FMoW. *CLIP WiSE* models are constructed by linearly combining the model weights of *CLIP ViT-B16 Zeroshot model* and *CLIP ViT-B16 FMoW end-to-end finetuned* model.

To illustrate the usage of our method, we choose the *CLIP WiSE* model that has similar ID validation accuracy as the *DenseNet* Models. This turns out to be putting $60\%$ weight on *CLIP ViT-B16 Zeroshot model* and $40\%$ weight on *CLIP ViT-B16 FMoW end-to-end finetuned*. *DenseNet* Models have average ID validation loss $2.4 - 2.8$, but *CLIP WiSE* has average ID validation loss $1.6$. To ensure fair comparison, we calibrate the temperature parameter such that the average loss of *CLIP WiSE* matches the worst average loss of the models considered. We deliberately make *CLIP WiSE* no better than any *DenseNet* Models, in the hope that out metric will recover its robustness property.

### B.3 Flexibility of our metric

We implement Algorithm 1 by partitioning the ID validation data into two; we estimate $h^\star(Z)$ using XGBoost on one sample, and estimate $\mathsf{W}_\alpha(\cdot)$ at varying subpopulation size $\alpha$ on the other. By switching the role of each split, our final estimator averages two versions of $\widehat{\mathsf{W}}_\alpha(\widehat{h})$.

#### B.3.1 A less conservative Z

In Section 4, we report results when $Z$ is defined over all metadata consisting of (longitude, latitude, cloud cover, region, year), as well as the label $Y$. Defining subpopulations over such a wide range of variables may be overly conservative in some scenarios, and to illustrate the flexibility of our approach, we now showcase a more tailored definition of subpopulations. Since FMoW-WILDS is specifically designed for spatiotemporal shifts, a natural choice of $Z$ is to condition on (region, year). Motivated by our observation that some classes are harder to predict than others (Figure 3(b)), we also consider $Z$= (region, year, label $Y$). We plot our findings in Figure 5. If we simply define $Z$= (year, region), the corresponding worst-case subpopulation performance is less pessimistic. However, when we add labels to $Z$, we again see a drastic decrease in the worst-case subpopulation performance, and that *CLIP WiSE-FT* outperforms all other models by a significant amount. This is consistent with our

| # | Text Prompt |
|---|---|
| 1 | "CLASSNAME" |
| 2 | "a picture of a CLASSNAME." |
| 3 | "a photo of a CLASSNAME." |
| 4 | "an image of an CLASSNAME" |
| 5 | "an image of a CLASSNAME in asia." |
| 6 | "an image of a CLASSNAME in africa." |
| 7 | "an image of a CLASSNAME in the americas." |
| 8 | "an image of a CLASSNAME in europe." |
| 9 | "an image of a CLASSNAME in oceania." |
| 10 | "satellite photo of a CLASSNAME" |
| 11 | "satellite photo of an CLASSNAME" |
| 12 | "satellite photo of a CLASSNAME in asia." |
| 13 | "satellite photo of a CLASSNAME in africa." |
| 14 | "satellite photo of a CLASSNAME in the americas." |
| 15 | "satellite photo of a CLASSNAME in europe." |
| 16 | "satellite photo of a CLASSNAME in oceania." |
| 17 | "an image of a CLASSNAME" |

**Table 2:** Text prompts for CLIP text encoders

motivation in defining subpopulations over labels; our procedure automatically takes into account the interplay between class labels and spatiotemporal information.

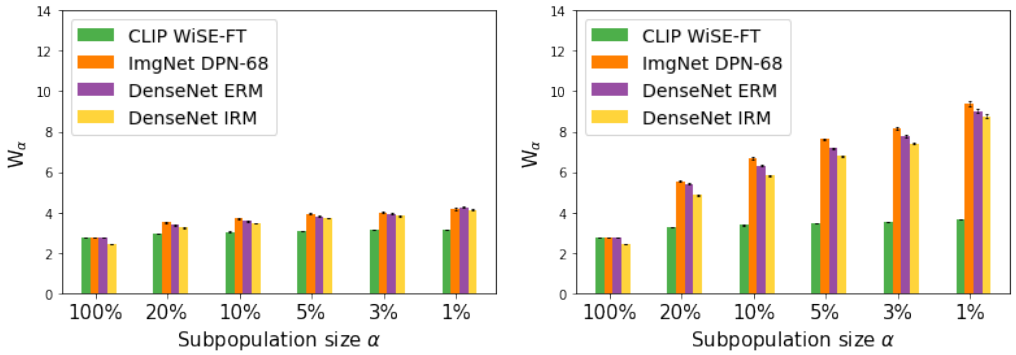


**Figure 5.** In the left panel $Z$ = (year, region); in the right panel $Z$ = (year, region, label $Y$). Here we take $Z$ to contain only spatial and temporal information, a less conservative counterpart to the experiment reported in the main text. We again see that introduction of labels in $Z$ drastically increase our metric, showing varying difficulty in learning different labels.

### B.3.2 Using semantics of the labels

Alternatively, we may wish to define subpopulations over rich natural language descriptions on the input $X$. To illustrate the flexibility of our procedure in such scenarios, we consider subpopulations defined over the semantic meaning of the class names: CLIP-encoded class names using the 17 prompts reported in Table 2. For comparison, we report the (estimated) worst-case subpopulation performance (2) when we take $Z$ = (all metadata, encoded labels) and (all metadata, label Y, encoded labels) in Figure 6. We observe that in this case, the semantics of the class names do not contribute to further deterioration in robustness, and the relative ordering across models remains unchanged.

## B.4 Analysis of spatiotemporal distribution shift

The significant performance drop in the Africa region on data collected from 2016-2018 was also observed in [53, 85]. In Figures 7-8, we plot the number of samples collected from Africa over data splits. In particular, we observe a large number of single-unit and multi-unit residential instances emerge in the OOD data. Data collection systems are often biased against the African continent—often as a result of remnants of colonialism—and addressing such bias is an important topic of future research.

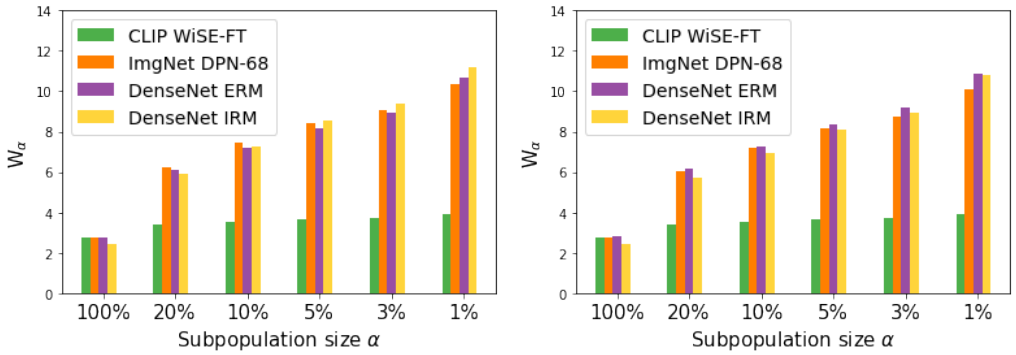


**Figure 6.** In the left panel, $Z$ = (all meta, encoded labels); in the right panel, $Z$ = (all meta, label Y, encoded labels). We see that in this case no significant difference is introduced when semantics of the class names are included.

### B.5 Estimator of model loss

One potential limitation of our approach is $\hat{h}$ does not always estimate the tail losses accurately, and this is important because our approach precisely is designed to counter ML models that perform poorly on tail subpopulations. Figure 9 plots a histogram of model losses and the estimated conditional risk $\hat{h}$ for *DenseNet ERM* and *CLIP WiSE*, where the y-axis is plotted on a log-scale. It is clear that *DenseNet ERM* has more extreme losses compared to the *CLIP WiSE* model, suggesting that at least part of the reason why *DenseNet ERM* suffers poor loss on subpopulations: it is overly confident when it's incorrect. While a direct comparison is not appropriate since the conditional risk $\mu(Z)$ represent *smoothed* losses, we observe that naive estimators of $\mu(\cdot)$ may consistently underestimate. In this particular instance, since the extent of underestimation is more severe for ImageNet pre-trained models, our experiments are fortuitously providing an even more conservative comparison between the two model classes, instilling confidence in the relative robustness of the *CLIP WiSE* model.

Alternatively, we can directly define the worst-case subpopulation performance (2) using the 0-1 loss. The discrete nature of the 0-1 loss pose some challenges in estimating $\mu(\cdot)$. While we chose to focus on the cross entropy loss that aligns with model training, we leave a thorough study of 0-1 loss to future work.

### B.6 Additional comparisons

We use the ensembled *CLIP WiSE* model constructed by averaging the network weights of *CLIP zero-shot* and *CLIP finetuned* models. So far, we used proportion $\lambda = 0.4$ to match the ID validation accuracy of *CLIP WiSE* to that of *DenseNet* models and *DPN*-68. In this subsection, we provide alternative choices:

1. $\lambda = 0.24$ to match ID accuracy of *ResNet*-18 of $47\%$
2. $\lambda = 0.27$ to match ID accuracy of *VGG*-11 of $51\%$.

Similar to *DPN*-68, *ResNet*-18 and *VGG*-11 are ImageNet pretrained models fine-tuned on FMoW as evaluated by Miller et al. [58]. We refer to the two *CLIP WiSE* models as *CLIP WiSE 24* and *CLIP WiSE 27* respectively, and report all model performances below. Again, we observe that our approach successfully picks out the more robust *CLIP WiSE* models, in contrast to the non-robust models chosen by ID accuracy or ID loss.

## C Simulation

In this section, we illustrate the asymptotic convergence of our two-stage estimator $\widehat{\mathsf{W}}_\alpha(\widehat{h}_1)$ of the worst-case subpopulation performance $\mathsf{W}_\alpha(\theta)$. We conduct a binary classification experiment similar to Duchi and Namkoong [32].

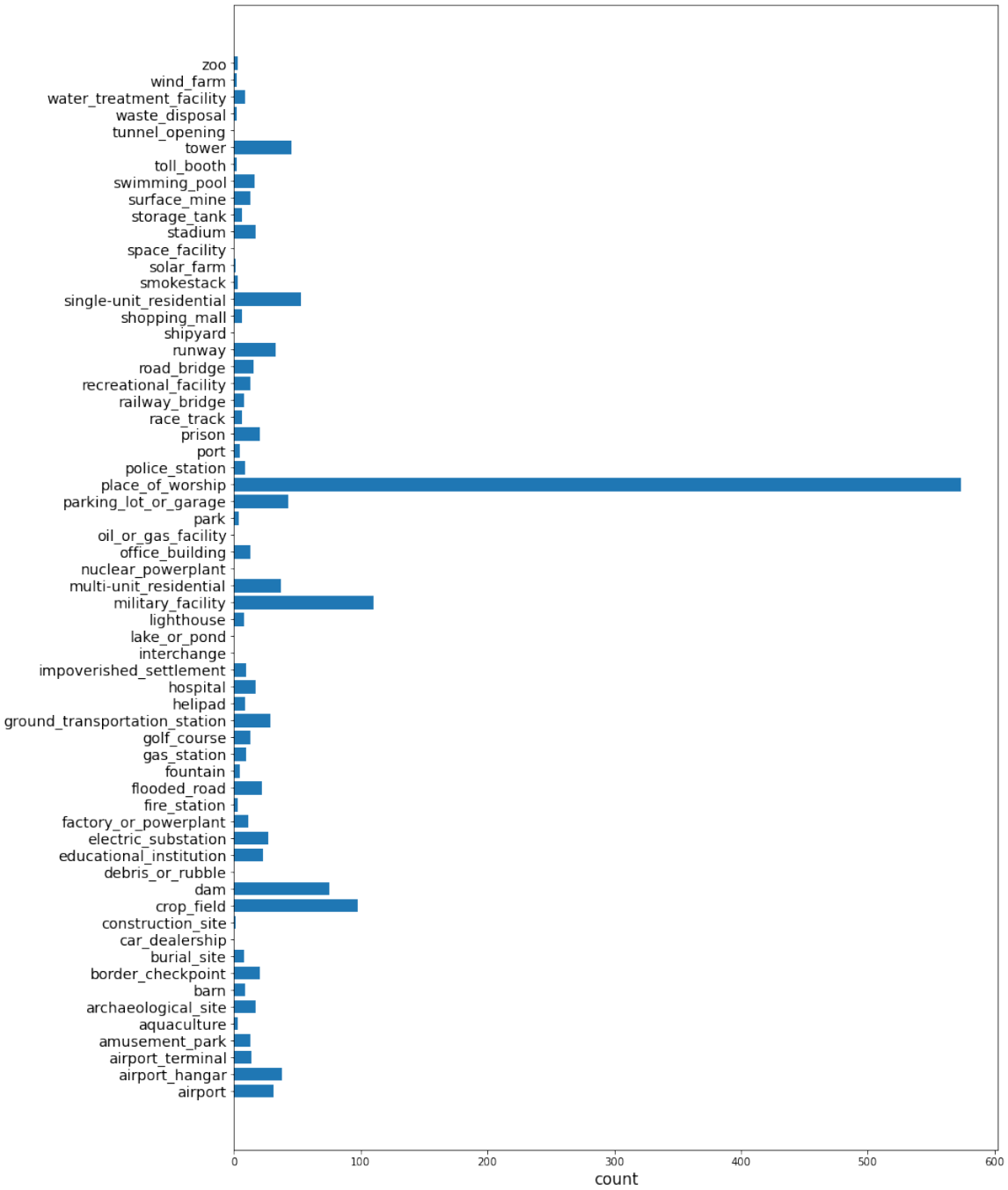


**Figure 7:** Instances by class, ID 2002-2013, Africa

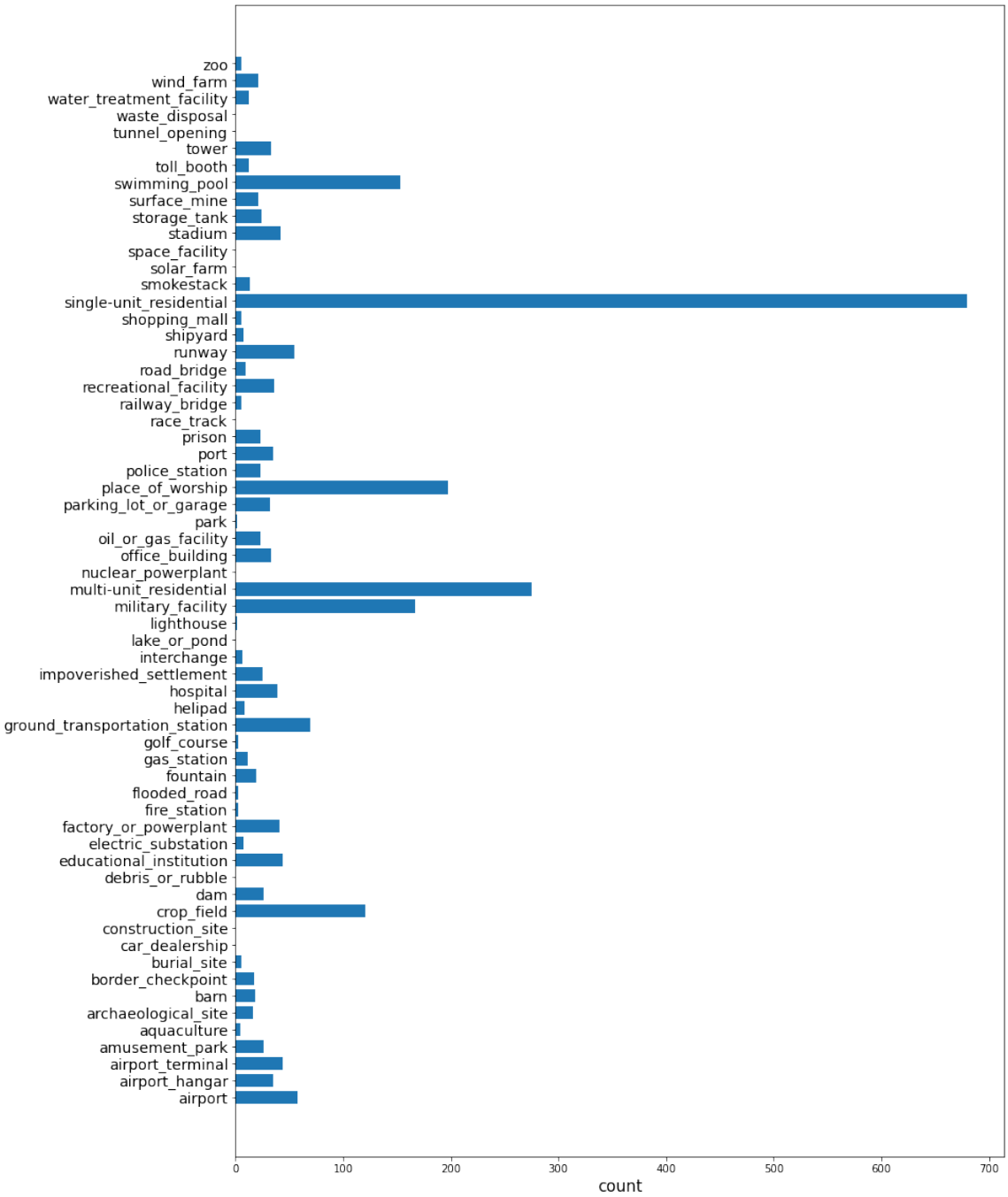


**Figure 8:** Instances by class, test 2016-2018, Africa

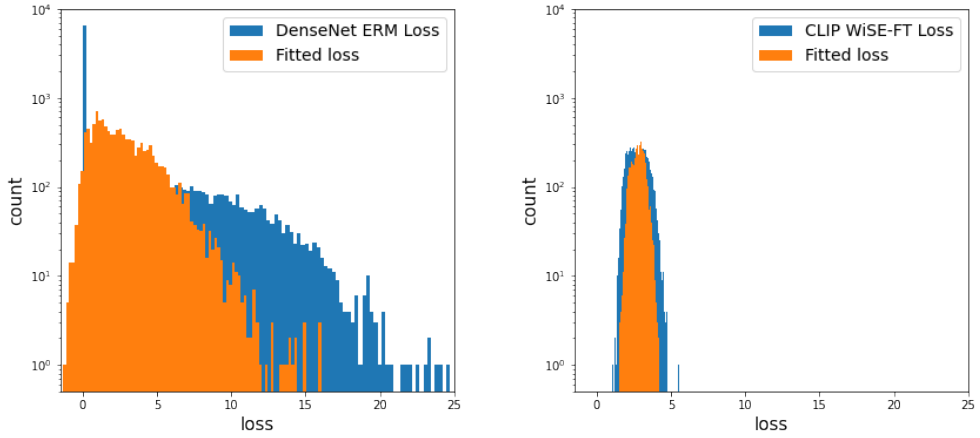


**Figure 9.** Histograms of model losses and fitted losses $\hat{h}$. Y-axis count is plotted in **log-scale**. For *DenseNet ERM* model, fitted $\hat{h}$ underestimates the extreme losses (right-tail).

| | ID, 2002-2013 | | | OOD, 2016-2018 | |
|---|---|---|---|---|---|
| Model | Accuracy | Loss | $\mathsf{W}_{0.10}$ | Accuracy | Loss |
| CLIP WiSE 24 | 0.47 | 2.84 | 3.47 | 0.45 | 2.85 |
| ResNet-18 | 0.48 | 2.84 | 5.05 | 0.40 | 3.36 |
| CLIP WiSE 27 | 0.51 | 3.07 | 3.54 | 0.48 | 3.08 |
| VGG-11 | 0.51 | 3.06 | 6.07 | 0.45 | 3.68 |

**Table 3.** Additional experiments showcasing our approach successfully identifies more robust models.

Formally, we randomly generate and fix two vectors $\theta, \theta_0^\star \in \mathbb{R}^d$ on the unit sphere. The data-generating distribution is given by $X \overset{iid}{\sim} \mathsf{N}(\gamma, \Sigma)$ and

$$Y \mid X = \begin{cases} \text{sgn}(X^\top \theta_0^\star) & X^1 \leq z_{0.95} = 1.645 \\ -\text{sgn}(X^\top \theta_0^\star) & \text{otherwise.} \end{cases}$$

In this data-generating distribution, there is a drastic difference between subpopulations generated by $X^1 \leq z_{0.95}$ and $X^1 > z_{0.95}$; typical prediction models will perform poorly on the latter rare group. The loss function is taken to be the hinge loss $\ell(\theta; x, y) = [1 - y \cdot \theta^\top x]_+$, where $y \in \{\pm 1\}$. We take the first covariate $X^1$ as our protected attribute $Z$. Let $d = 5$, $\Sigma = \mathsf{I}_5$, $\gamma = 0$.

We fix $\alpha = 0.3$. To analyze the asymptotic convergence of our two-stage estimator, for sample size ranging in 1,000 to 256,000 doubling each time, we run 40 repeated experiments of the estimation procedure on simulated data. We split each sample evenly into $S_1$ and $S_2$ and using gradient boosted trees in the package XGBoost [22] to estimate the conditional risk. On a log-scale, we report the mean estimate across random runs in Figure 10 alongside error bars. To compute the true worst-case subpopulation performance $\mathsf{W}_\alpha(\mu)$ of the conditional risk $\mu(X^1)$, we first run a Monte Carlo simulation for 150,000 copies of $X^1 \sim \mathsf{N}(0, 1)$. For each sampled $X^1$, we generate 100,000 copies of $(X^2, X^3, X^4, X^5) \sim \mathsf{N}(\mathbf{0}, \mathsf{I}_4)$ independent of $X^1$ and compute the mean loss among them to approximate the conditional risk $\mu(X^1)$. Finally, we approximate $\mathsf{W}_\alpha(\mu)$ using the empirical distribution of $\mu(\cdot)$, obtaining $6.47 \times 10^{-1}$. We observe convergence toward the true value as sample size $n$ grows, verifying the consistency of our two-stage estimator $\widehat{\mathsf{W}}_\alpha(\widehat{h}_1)$.

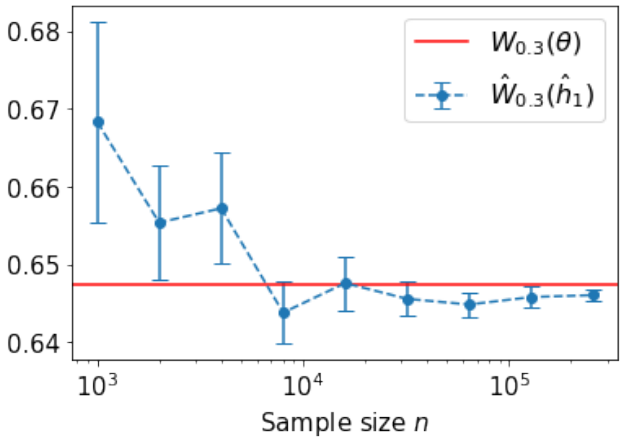


**Figure 10:** $\widehat{\mathsf{W}}_\alpha(\widehat{h}_1)$ and $\mathsf{W}_\alpha(\theta)$ from simulation experiments with $\alpha = 0.3$