Probable Domain Generalization
via Quantile Risk Minimization

Abstract

Domain generalization (DG) leverages labeled training data from multiple domains with the goal of generalizing to related test domains. To realize this goal, the standard formulation optimizes for worst-case performance by taking an adversarial perspective over the set of all possible domains. However, this worst-case problem is generally intractable and, with adversarial shifts extremely unlikely in practice, leads to overly-conservative solutions. In fact, a recent study found that no DG algorithm outperformed empirical risk minimization in terms of *average performance* over test domains. To address these shortcomings, we propose a probabilistic framework for DG, which we call *Probable Domain Generalization*, and advocate for predictors that perform well *with high probability* rather than in the worst-case or on-average. Our key idea is that distribution shifts seen during training should inform us of *probable* shifts at test time. To achieve this, we explicitly relate training and test domains as draws from the same underlying meta-distribution, and propose a new optimization problem—*Quantile Risk Minimization* (QRM)—which requires that predictors generalize with high probability. We then prove that, given sufficiently many domains and samples, the empirical version(EQRM) produces predictors that generalize to new domains with the desired probability. We also show that EQRM recovers the causal predictor as the desired probability of generalization approaches one. In our experiments, we introduce a new evaluation protocol for DG, which underscores the importance of multiple test domains for evaluating the quantile performance of DG algorithms, and we show that our algorithms outperform strong DG baselines on real and synthetic data.

1 Introduction

Despite remarkable successes in recent years [1–3], machine learning systems often fail calamitously when presented with *out-of-distribution* (OOD) data [4–7]. In fact, evidence of state-of-the-art systems failing in the face of distribution shift is mounting rapidly—be it due to spurious correlations [8–10], changing sub-populations [11–13], changes in location or time [14–16], or other naturally-occurring variations [17–22]. These OOD failures are particularly concerning in safety-critical applications such as medical imaging [23–27] and autonomous driving [28–30], where they represent one of the most significant barriers to real-world deployment [31–34].

Domain generalization (DG) seeks to improve a system's OOD performance by leveraging datasets from multiple environments or *domains* at training time, each collected under different experimental conditions [35–37]. The goal is to build a predictor which exploits invariances across the training domains in the hope that these invariances also hold in related but distinct test domains [37–40]. To realize this goal, the standard DG formulation optimizes for worst-case performance by taking an adversarial perspective over the set of all possible domains. However, this worst-case problem is generally intractable [41–43] and, with adversarial shifts unlikely in practice, leads to overly-conservative solutions—making large accuracy sacrifices in pursuit of adversarial robustness. In fact, a recent study [37] found that, across several popular datasets, no DG algorithm outperformed *empirical risk minimization* (ERM) when compared in terms of *average performance* over test domains.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

(a)   (b)   (c)

Figure 1: **Overview of Probable Domain Generalization and quantile risk.** (a) Domains differ due to changes in (or interventions on) an underlying system of variables, e.g. hospitals with different equipment, procedures, etc. Hammers depict interventions. (b) Training and test domains are randomly sampled from the same underlying meta-distribution over domains $\mathbb{Q}(e)$. (c) Estimated risk distribution over training domains for a fixed predictor. The α-quantile is an approximate upper bound on the test risk which holds with probability $\approx \alpha$.

In this work, we argue that DG should both seek and evaluate predictors based on their ability to perform well *with high probability*, rather than in the worst-case or on-average. To this end, we propose a probabilistic framework for DG, which we call *Probable Domain Generalization*, wherein the key idea is that distribution shifts seen during training should inform us of *probable* shifts at test time. To realize this, we explicitly relate training and test domains as draws from the same underlying meta-distribution, then propose *Quantile Risk Minimization* (QRM)—a new optimization problem which requires that predictors *generalize with high probability* (§ 3). The goal of QRM is to minimize the quantiles of a predictor's risk distribution over training domains, leveraging the fact that the α-quantile of this distribution is an upper bound on the test risk which holds with probability α (see Fig. 1c).

Next, we prove that, given sufficiently many domains and samples, the empirical analogue (EQRM, § 4) produces predictors which generalize to new domains with probability α. We also show that, under similar assumptions to previous work [44], EQRM recovers the causal predictor as $\alpha \to 1$. Thus, α is an interpretable conservativeness-hyperparameter, with $\alpha = 1$ corresponding to the adversarial setting. In our experiments (§ 6), we introduce a new evaluation protocol for DG which focuses on a predictor's performance at different test quantiles, and demonstrate that: (i) the performance of DG algorithms was likely never "lost" [37], but rather invisible through the lens of average performance; and (ii) our quantile-minimizing algorithms outperform state-of-the-art baselines on real and synthetic datasets.

Contributions. To summarize our main contributions, we:

- Propose *Quantile Risk Minimization* for learning predictors that generalize with probability α (§ 3).
- Prove that, given sufficiently many domains and samples, empirical QRM (EQRM) does indeed produce predictors that generalize to new domains with probability α (Thm. 4.1).
- Prove that EQRM recovers the causal predictor as $\alpha \to 1$ (similar assumptions to [44], Prop. A.4).
- Demonstrate empirically that: (i) effective comparison of DG algorithms requires evaluating quantile performance; and (ii) our algorithms outperform strong baselines on real and synthetic data (§ 6).

2   Background

Domain generalization (DG). In DG, predictors are trained on data drawn from multiple related training distributions or *domains* and then evaluated on related but unseen test domains. For example, in the Camelyon17 dataset [45], the task is to predict if a given image of cells contains cancerous tumors, and domains correspond to different hospitals in which these images were captured. More formally, we consider datasets $D^e = \{(x_i^e, y_i^e)\}_{i=1}^{n_e}$ collected from m different training domains or *environments* $\mathcal{E}_{\text{tr}} := \{e_1, \ldots, e_m\}$, with each dataset D^e containing data pairs (x_i^e, y_i^e) sampled i.i.d. from $\mathbb{P}(X^e, Y^e)$. Then, given a suitable function class $\mathcal{F}$ and loss function ℓ, the goal of DG is to learn a predictor $f \in \mathcal{F}$ that generalizes to data drawn from a larger set of all possible environments $\mathcal{E}_{\text{all}} \supset \mathcal{E}_{\text{tr}}$. By letting $\mathcal{R}^e(f)$ denote the statistical risk of f in domain e, this can be formalized as the following minimax optimization problem:

$$\min_{f \in \mathcal{F}} \max_{e \in \mathcal{E}_{\text{all}}} \mathcal{R}^e(f) \qquad \text{where} \qquad \mathcal{R}^e(f) := \mathbb{E}_{\mathbb{P}(X^e, Y^e)}[\ell(f(X^e), Y^e)]. \tag{DG}$$

As we only have access to data from a finite subset of $\mathcal{E}_{\text{all}}$ during training, solving (DG) is not just challenging but in fact impossible [40, 46, 47] without restrictions on how the domains may differ.

Causality and invariance in DG. Causal works on DG [9, 40, 44, 47, 48] describe domain or environmental differences using the language of causality and the notion of *interventions* [49, 50]. In particular, they assume all environments share the same underlying *structural causal model* (SCM) [49], with different environments corresponding to different interventions (see Appendix A.1 for formal definitions and a simple example). Early work studied the problem of learning from multiple cause-effect datasets that share a functional mechanism but differ in noise distributions [38]. More generally, given (data from) multiple distributions, one can try to identify components which are stable, robust, or *invariant*, and find means to transfer them across problems [51–55]. Assuming that the mechanism of Y remains fixed or invariant[1] but all Xs may be intervened upon, recent works have shown that only the causal predictor: (i) has invariant predictive distributions [44], coefficients [9], or risks [40] across environments; and (ii) generalizes to arbitrary interventions on the Xs [9, 44, 48]. These works then exploit such insights by leveraging some form of invariance across environments to discover causal relationships which, through the invariant mechanism assumption, generalize to new environments.

3 Quantile Risk Minimization

In this section we introduce *Quantile Risk Minimization* (QRM) for achieving *probable* domain generalization. The core idea is to replace the worst-case perspective of (DG) with a probabilistic one. This approach is founded on a great deal of work in classical fields such as control theory [56, 57] and smoothed analysis [58], wherein approaches that yield high-probability guarantees are used in place of worst-case approaches in an effort to mitigate conservatism and computational limitations. This mitigation is of particular interest in domain generalization since generalizing to arbitrary domains is impossible [40, 46, 47]. Thus, motivated by this classical literature, our goal is to obtain predictors that are robust *with high probability* over domains drawn from $\mathcal{E}_{\text{all}}$, rather than in the worst-case.

A distribution over environments. We start by assuming the existence of a probability distribution $\mathbb{Q}(e)$ over the set of all environments $\mathcal{E}_{\text{all}}$. For instance, in the context of medical imaging, $\mathbb{Q}$ could represent a distribution over potential changes to a hospital's setup (see Fig. 1a) or simply a distribution over candidate hospitals. Given that such a distribution $\mathbb{Q}$ exists[2], we can think of the risk $\mathcal{R}^e(f)$ as a *random variable* for each $f \in \mathcal{F}$, where the randomness is engendered by the draw of $e \sim \mathbb{Q}$. This perspective gives rise to the following analogue of the optimization problem in (DG):

$$\min_{f \in \mathcal{F}} \operatorname*{ess\,sup}_{e \sim \mathbb{Q}} \mathcal{R}^e(f) \quad \text{where} \quad \operatorname*{ess\,sup}_{e \sim \mathbb{Q}} \mathcal{R}^e(f) = \inf \left\{ t \geq 0 : \Pr_{e \sim \mathbb{Q}} \left\{ \mathcal{R}^e(f) \leq t \right\} = 1 \right\} \quad (3.1)$$

Here, ess sup denotes the *essential-supremum* operator from measure theory, meaning that for each $f \in \mathcal{F}$, $\operatorname{ess\,sup}_{\mathbb{Q}} \mathcal{R}^e(f)$ is the least upper bound on $\mathcal{R}^e(f)$ that holds for almost every $e \sim \mathbb{Q}$. In this way, the ess sup in (3.1) is the measure-theoretic analogue to the max operator in (DG), the caveat being that the ess sup in (3.1) can neglect domains of $\mathbb{Q}$-measure zero.

High-probability generalization. Although the minimax problem in (3.1) explicitly incorporates the distribution $\mathbb{Q}$ over environments, this formulation is no less conservative than (DG). Indeed, in many cases of interest, (3.1) is equivalent to (DG); see Appendix E for details. Therefore, rather than considering the worst-case problem in (3.1), we propose the following generalization of (3.1) which requires that predictors generalize with probability α rather than in the worst-case:

$$\min_{f \in \mathcal{F}, t \in \mathbb{R}} t \qquad \text{subject to} \qquad \Pr_{e \sim \mathbb{Q}} \left\{ \mathcal{R}^e(f) \leq t \right\} \geq \alpha \qquad (3.2)$$

The optimization problem of (3.2) formally defines what we mean by *probable* domain generalization. In particular, we say that *a predictor f generalizes with risk t at level α* if f has risk at most t with probability at least α over domains sampled from $\mathbb{Q}$. In this way, the conservativeness parameter α controls the strictness of generalizing to unseen domains.

A distribution over risks. The optimization problem presented in (3.2) offers a principled formulation for generalizing to unseen distributional shifts governed by $\mathbb{Q}$. However, $\mathbb{Q}$ is often unknown in practice and its support $\mathcal{E}_{\text{all}}$ may be high-dimensional or challenging to define [22]. While many previous works have made progress by limiting the scope of possible shifts over domains [19, 59, 60], in practice, such structural assumptions are often difficult to justify and impossible to test. For this reason, we start our exposition of QRM by offering an alternative view of (3.2) which elucidates how a predictor's *risk distribution* plays a central role in achieving probable DG.

---

[1]Arjovsky et al. [9] allow the noise variance of Y to vary.

[2]As $\mathbb{Q}$ is often unknown, our analysis does not rely on leveraging an explicit expression for $\mathbb{Q}$.

To begin, note that for each $f \in \mathcal{F}$, the distribution over domains $\mathbb{Q}$ naturally induces[3] a distribution $\mathbb{T}_f$ over the risks in each domain $\mathcal{R}^e(f)$. In this way, rather than considering the randomness of $\mathbb{Q}$ in the often-unknown and (potentially) high-dimensional space of possible distribution shifts or interventions (Figs. 1a and 1b), one can consider it in the real-valued space of risks (Fig. 1c). This is analogous to statistical learning theory, where the analysis of convergence of empirical risk minimizers (i.e., of functions) is substituted by that of a weaker form of convergence, namely that of scalar risk functionals—a crucial step for VC theory [61]. From this perspective, the statistics of $\mathbb{T}_f$ can be thought of as capturing the sensitivity of f to different environmental shifts, summarizing the effect of different intervention types, strengths, and frequencies. To this end, (3.2) can be equivalently rewritten in terms of the risk distribution $\mathbb{T}_f$ as follows:

$$\min_{f \in \mathcal{F}} F_{\mathbb{T}_f}^{-1}(\alpha) \quad \text{where} \quad F_{\mathbb{T}_f}^{-1}(\alpha) := \inf \left\{ t \in \mathbb{R} : \Pr_{R \sim \mathbb{T}_f} \{R \leq t\} \geq \alpha \right\}. \quad \text{(QRM)}$$

Here, $F_{\mathbb{T}_f}^{-1}(\alpha)$ denotes the inverse CDF (or quantile) function of the risk distribution $\mathbb{T}_f$. By means of this reformulation, we elucidate how solving (QRM) amounts to finding a predictor with minimal α-quantile risk. Also, by varying α, (QRM) recovers several notable objectives for DG as special cases.

Proposition 3.1. *For $\alpha = 1$, (QRM) is equivalent to the minimax problem of (3.1). Furthermore, if the mean and the median of $\mathbb{T}_f$ coincide for all $f \in \mathcal{F}$, then for $\alpha = 1/2$, (QRM) is equivalent to the average-case problem of risk minimization (RM, [61]) in which $\mathbb{T}_f$ is the uniform distribution*

$$\min_{f \in \mathcal{F}} \mathbb{E}_{R \sim \mathbb{T}_f} R = \min_{f \in \mathcal{F}} \mathbb{E}_{e \sim \mathbb{Q}} \mathcal{R}^e(f) \quad (3.3)$$

4 Algorithms for Quantile Risk Minimization

In this section we introduce two algorithms for empirically solving (QRM): empirical QRM (EQRM) and empirical superquantile RM (ESQRM).

4.1 Empirical QRM

Estimated risk distribution $\widehat{\mathbb{T}}_f$. In practice, given a predictor f and its empirical risks $\hat{\mathcal{R}}^{e_1}(f), \ldots, \hat{\mathcal{R}}^{e_m}(f)$ on the m training domains, we must form an *estimated* risk distribution $\widehat{\mathbb{T}}_f$. If we have prior knowledge about the form of $\mathbb{T}_f$ (e.g. Gaussian), we can simply estimate the parameters of this distribution (e.g. via the sample mean and variance). In the absence of such knowledge, we can use nonparametric methods like *kernel density estimation* (KDE, [62, 63]). Fig. 1c depicts the PDF and CDF for 10 training risks when using KDE with Gaussian kernels, while Fig. 4 of Appendix B compares the smoothed KDE CDF (black) to the unsmoothed empirical CDF (gray). As shown in Fig. 4, the KDE smoothing permits risk "extrapolation" [40] beyond our largest training risk, with the estimated α-quantile risk going to infinity as α goes to one (when kernels have full support). Importantly, different bandwidth-selection methods encode different assumptions about right-tail heaviness and thus about projected OOD risk. In Appendix B, we discuss the suitability of different data-dependent bandwidth-selection methods for our purposes. In practice, we use Gaussian kernels with either the Gaussian-optimal rule [64] or Silverman's rule-of-thumb [64] for bandwidth selection.

Empirical problem. Armed with a predictor's estimated risk distribution $\widehat{\mathbb{T}}_f$, we can approximately solve (QRM) using the following empirical analogue:

$$\widehat{\mathrm{QRM}}_\alpha(f) := \min_{f \in \mathcal{F}} F_{\widehat{\mathbb{T}}_f}^{-1}(\alpha) \quad (4.1)$$

Note that (4.1) depends only on known quantities, so we can compute and minimize it in practice. Also note that in finance, when concerned with a distribution over potential investment losses, the α-quantile value is known as the *value at risk*[4] (VaR) at level α and represents the maximum possible loss after excluding all worst outcomes whose combined probability is at most $1 - \alpha$ [65]. Appendix D.1 gives further algorithmic details, including pseudocode in Alg. 1.

---

[3]More formally, $\mathbb{T}_f$ can be defined as the push-forward measure of $\mathbb{Q}$ through the risk functional $\mathcal{R}^e(f)$; see Appendix E for a formal definition.

[4]The "risk" in value-at-risk simply refers to the fact that some amount/value of money is at risk of being lost.

Generalization bound. We now give a simplified version of our main generalization bound—Thm. G.1. The full version, stated and proved in Appendix G, provides specific finite-sample bounds on ϵ_1 and ϵ_2 below, depending on the hypothesis class $\mathcal{F}$, the empirical estimator $\widehat{\text{QRM}}_\alpha$, and the assumptions made on the possible risk profiles of hypotheses $f \in \mathcal{F}$.

Theorem 4.1 (Simplified form of Thm. G.1, uniform convergence). *Given m domains and n samples in each, then with high probability over the training data,*

$$\sup_{f \in \mathcal{F}} \left| F^{-1}_{\mathbb{T}_f}(\alpha - \epsilon_2) - F^{-1}_{\widehat{\mathbb{T}}_f}(\alpha) \right| \leq \epsilon_1, \tag{4.2}$$

where $\epsilon_1 \to 0$ as $n \to \infty$ and $\epsilon_2 \to 0$ as $m \to \infty$.

This theorem states that, for m, n sufficiently large, the empirical α-quantile risk $F^{-1}_{\widehat{\mathbb{T}}_f}(\alpha)$ is a good estimate of the population α-quantile risk $F^{-1}_{\mathbb{T}_f}(\alpha)$. In particular, a predictor f which minimizes the empirical problem in (4.1) should approximately minimize the population problem in (QRM), meaning f will have risk $\lesssim\!\!\approx F^{-1}_{\widehat{\mathbb{T}}_f}(\alpha)$ with probability $\gtrsim\!\!\approx \alpha$. That is, such an f will have risk below $F^{-1}_{\widehat{\mathbb{T}}_f}(\alpha)$ on approximately $100\alpha\%$ of test domains.

Recovering the causal predictor. We now give an informal version of the results in Appendix A.2, which show that QRM can recover the causal mechanism of Y as α goes to one.

Proposition 4.2 (QRM can recover the causal predictor). *Assume: (i) data is generated from a linear SEM $Y = \beta X + N$, with coefficient $\beta \in \mathbb{R}^d$ and noise distribution N constant across environments, $\mathbb{E}[N] = 0$, and N independent of X; (ii) $\mathcal{F}$ is the class of linear predictors, indexed by $\hat{\beta} \in \mathbb{R}^d$; (iii) all causes of Y are observed; (iv) the loss ℓ is squared-error; and (v) the distribution of $\mathbb{E}_{X \sim e}[X]$ (over environments $e \sim \mathbb{Q}$) has a density with respect to the Lebesgue measure on $\mathbb{R}^d$. Then, for both Gaussian estimators of $\mathbb{T}_f$ and kernel-density estimators of $\mathbb{T}_f$ with certain bandwidth-selection methods, for m sufficiently large, the QRM predictor $\hat{\beta}_\alpha \to \beta$ as $\alpha \to 1$, with probability 1.*

In Appendix A.2 we prove Proposition 4.2 and discuss both bandwidth-selection methods and connections to similar results like that of Krueger et al. [40, Thm 1]. In § 6.1, we demonstrate that this result holds empirically. Finally, in Appendix A.2.1, we relate our α parameter to the β parameter of VREx [40, Eq. 8] for the special case of Gaussian estimators of $\mathbb{T}_f$.

4.2 Empirical SQRM

Although the formulation presented in (4.1) is intuitive and interpretable, the quantile function is notoriously difficult to optimize [66]. Indeed, in many cases of interest, F^{-1} can be nonconvex and discontinuous, which induces an unfavorable optimization landscape [67, 68]. To circumvent this, we propose an alternative scheme in which we optimize the *superquantile* (SQ)—also known as the *conditional value at risk* (CVaR) or *expected tail loss*—at level α, which is defined[5] as follows:

$$\text{SQ}_\alpha(R; \mathbb{T}_f) := \mathbb{E}_{R \sim \mathbb{T}_f}\left[R \mid R \geq F^{-1}_{\mathbb{T}_f}(\alpha) \right] \tag{4.3}$$

SQ_α can be seen as the conditional expectation of a random variable R subject to R being larger than the α-quantile $F^{-1}(\alpha)$. In our case, where R represents the statistical risk on a randomly-sampled environment, SQ_α can be seen as the expected risk in the worst $100 \cdot (1 - \alpha)\%$ of cases/environments. Importantly, unlike the quantile, the superquantile has many desirable mathematical properties: (i) it is continuous in α; (ii) it is jointly convex in R and α; and (iii) it is the tightest convex surrogate for the inverse CDF [69, § 2]. Therefore, we also consider the following optimization problem:

$$\min_{f \in \mathcal{F}} \text{SQ}_\alpha(R; \mathbb{T}_f) \tag{SQRM}$$

We call the empirical analogue, which uses the empirical density rather than $\mathbb{T}_f$, empirical SQRM. Appendix D.1 gives further algorithmic details, including pseudocode in Alg. 2.

---

[5]This definition assumes that $\mathbb{T}_f$ is continuous; we present a more general definition in Appendix F.

Connections to existing problems. It is straightforward to show that (SQRM) recovers the RM problem in (3.3) and the minimax problem in (3.2) as special cases with $\alpha = 0$ and $\alpha = 1$ respectively. Moreover, in Appendix F, we exploit the well-known duality properties of CVaR to show an alternative view of (SQRM) as a distributionally-robust optimization problem, admitting close connections to previous work [60]; see Propositions F.1 and F.2 for details.

5 Related work

Robust optimization in DG. Throughout this paper, we follow an established line of work (see e.g., [9, 40, 70]) which formulates the DG problem through the lens of robust optimization [71]. To this end, various algorithms have been proposed for solving constrained [59] and distributionally robust [60] variants of the worst-case problem in (DG). Indeed, this robust formulation has a firm foundation in the broader machine learning literature, with notable works in adversarial robustness [72–76] and fair learning [77, 78] employing similar formulations. Unlike these past works, we consider a robust but non-adversarial formulation for DG, where predictors are trained to generalize with high probability rather than in the worst case. Moreover, the majority of this literature—both within and outside of DG—relies on specific structural assumptions (e.g. covariate shift) on the types of possible interventions or perturbations. In contrast, QRM and SQRM make an assumption on the environment-generating process (we further discuss this important difference in § 7).

Other approaches to DG. Outside of robust optimization, many algorithms have been proposed for the DG setting which draw on insights from a diverse array of fields, including approaches based on tools from meta-learning [39, 79–82], kernel methods [83, 84], and information theory [85, 86]. Also prominent are works that design regularizers to generalize OOD [87–89] and works that seek domain-invariant representations [90–93]. However, the majority of these works do not explicitly relate the environments seen during training and test time, which is a key feature of our approach. Moreover, many of these works employ hyperparameters which are difficult to interpret, which has no doubt contributed to the well-established model-selection problem in DG [37]. In contrast, in our framework, α can be clearly interpreted in terms of the quantiles of the risk distribution $\mathbb{T}_f$.

High-probability generalization. As noted in § 3, relaxing worst-case problems in favor of probabilistic ones has a long history in control theory [41, 56, 57, 94, 95], operations research [96], and smoothed analysis [58]. Recently, this paradigm has been applied to several areas of machine learning, including perturbation-based robustness [97, 98], fairness [99], active learning [100], and reinforcement learning [101, 102]. However, it has not yet been applied to domain generalization.

Quantile minimization. In financial optimization, VaR and CVaR [65, 67, 103] are central to the literature surrounding portfolio risk management, with numerous applications spanning banking regulations and insurance policies [66, 104]. In statistical learning theory, several recent papers have derived uniform convergence guarantees in terms of alternative risk functionals besides expected risk [100, 105–107]. These results focus on CVaR and closely related functionals that can be written in terms of expectations over the loss distribution. In contrast, our uniform convergence guarantee (Theorem G.1) shows uniform convergence of VaR, which *cannot* be written as an such expectation; this necessitates stronger conditions to obtain uniform convergence, which ultimately suggest regularizing the estimated risk distribution (e.g., by kernel smoothing).

Invariant prediction and causality. As discussed in § 2, under the invariant mechanism assumption [50], recent works have leveraged different forms of invariance across environments to discover causal relationships which, by assumption, generalize to new environments [9, 40, 44, 48, 108–110]. More recently, noting that causal parameters are often too conservative and thus that performing well out-of-distribution often requires leveraging some non-causal relationships, Rothenhäusler et al. [111] proposed *Anchor regression* for interpolating between causal and predictive parameters using an interpretable parameter. However, the assumed shift types are even more restrictive than the aforementioned methods, despite nonlinear extensions [47].

6 Experiments

We now evaluate our algorithms on: synthetic datasets requiring OOD generalization; and real-world datasets from the WILDS [12] package, illustrating the practical benefits of our algorithms and also the importance of multiple test domains in DG benchmarks. Further results can be found in Appendix C.3, while full experimental details are given in Appendix D. Code will be made available.

Figure 2: **QRM on a toy linear regression dataset (A–D) and on ColoredMNIST (E–F). A:** Test risk at different quantiles or degrees of "OODness". Each α is optimal at the corresponding quantile. **B:** Estimated risk distributions $\widehat{\mathbb{T}}_f$ for different predictors f. As $\alpha \to 1$, $\widehat{\mathbb{T}}_{f_\alpha}$ approaches an invariant risk distribution centered on the risk of the causal predictor. See Fig. 5 for corresponding CDFs. **C:** Regression coefficients approach those of the causal predictor ($\beta_{\text{cause}} = 1, \beta_{\text{effect}} = 0$) as $\alpha \to 1$. **D:** Q-Q plot comparing the "true" risk quantiles (estimated with $m = 1000$) against "estimated" ones ($m < 1000$). Shown for QRM with $\alpha = 0.9$. **E:** Performance of different αs over increasingly OOD test domains. Dots show training-domain accuracies. **F:** KDE-estimated accuracy-CDFs depicting accuracy-robustness curves. Larger αs make lower accuracies (i.e. larger risks) less likely.

6.1 Synthetic datasets

Linear regression. We first consider a toy linear regression dataset based on the linear SCM of Ex. A.3. Here we have two features—one cause $X_1 = X_{\text{cause}} \in \text{Pa}(Y)$ and one effect $X_2 = X_{\text{effect}}$. By fixing $\sigma_1^2 = 1$ and $\sigma_Y^2 = 2$ across environments but sampling $\sigma_2 \sim \text{LogNormal}(0, 0.5)$, we create a dataset in which X_2 is more predictive of Y than X_1 but less stable. Importantly, as we know the true distribution over environments $\mathbb{Q}(e) = \text{LogNormal}(\sigma_2^e; 0, 0.5)$, we know the true risk quantiles. Fig. 2 depicts results for different α's with $m = 1000$ environments and $n = 100000$ samples in each, using the mean-squared-error loss. Here we see that: **A:** each α is optimal at the corresponding quantile, confirming that the empirical α-quantile risk is a good estimate of the population α-quantile risk; **B:** As $\alpha \to 1$, the estimated risk distribution of f_α approaches an invariant (or Dirac delta) distribution centered on the risk of the causal predictor; **C:** the regression coefficients approach those of the causal predictor as $\alpha \to 1$, trading predictive performance for robustness; and **D:** reducing the number of domains m reduces the accuracy of the estimated α-quantile risks. In Appendix C.1, we additionally: (i) depict the risk CDFs corresponding to plot **B** above, and discuss how they depict a predictor's risk-robustness curve (see Fig. 5); and (ii) compare QRM to existing algorithms on datasets in which σ_1^2, σ_2^2 or σ_Y^2 change over environments.

ColorMNIST. We next consider the `ColorMNIST` or `CMNIST` dataset of Arjovsky et al. [9], where the `MNIST` dataset is used to construct a binary classification task (0–4 or 5–9) in which digit color (red or green) is a highly-informative but spurious feature. In particular, the two training domains are constructed such that red digits have an 80% and 90% chance of belonging to class 0, while the single test domain is constructed such that red digits only have a 10% chance (labeled -90%). The goal is to learn an invariant predictor which uses only digit shape—a stable feature having a 75% chance of correctly determining the class in all 3 domains. Table 1

Table 1: `CMNIST` acc.

Alg.	-90%
ERM	28.7 ± 0.5
IRM	58.5 ± 3.3
V-REx	55.2 ± 4.0
GrpDRO	36.8 ± 2.8
QRM	$\mathbf{70.1 \pm 0.5}$
Oracle	73.1 ± 0.4

compares the performance of QRM to IRM [9], VREx [40] and GroupDRO [60] when using the standard hyperparameters and test-domain model selection of DomainBed [37]. Here we see that QRM outperforms existing methods and even approaches oracle performance (ERM trained on grayscale digits). Fig. 2 then depicts the behavior of QRM with different αs. In particular, we see

Table 2: QRM test risks on iWildCam.

Alg.	Mean risk	Quantile risk						
		0.0	0.25	0.50	0.75	0.90	0.99	1.0
ERM	1.31	0.015	0.42	0.76	2.25	2.73	4.99	5.25
$QRM_{0.25}$	2.03	0.024	0.46	2.70	3.01	3.48	5.03	5.26
$QRM_{0.50}$	1.11	**0.004**	0.24	0.68	1.71	2.15	4.04	4.11
$QRM_{0.75}$	1.05	0.009	**0.21**	0.68	1.50	2.35	4.88	5.45
$QRM_{0.90}$	**0.98**	0.047	0.28	**0.63**	**1.26**	**1.81**	4.11	4.48
$QRM_{0.99}$	0.99	0.12	0.35	0.64	1.30	2.00	**3.44**	**3.55**

Table 3: QRM test risks on OGB-MolPCBA.

Alg.	Mean risk	Quantile risk						
		0.0	0.25	0.50	0.75	0.90	0.99	1.0
ERM	**0.051**	0.0	0.004	0.017	0.060	0.13	0.49	16.04
$QRM_{0.25}$	0.054	0.0	0.003	0.016	0.059	0.13	0.48	15.46
$QRM_{0.50}$	0.052	0.0	0.003	0.015	0.059	0.13	0.48	11.33
$QRM_{0.75}$	0.052	0.0	0.003	0.015	0.059	0.13	0.47	12.15
$QRM_{0.90}$	0.052	0.0	0.003	0.015	0.059	0.12	0.47	10.81
$QRM_{0.99}$	0.053	0.0	0.003	**0.014**	**0.055**	**0.11**	**0.46**	**7.16**

Table 4: SQRM test risks on iWildCam.

Alg.	Superquantile risk						
	0.0	0.25	0.50	0.75	0.90	0.99	1.0
ERM	1.31	1.91	2.58	3.69	4.83	9.58	9.58
$SQRM_{0.25}$	1.18	1.53	2.10	2.99	4.09	6.03	6.03
$SQRM_{0.50}$	1.16	1.51	2.03	2.89	3.91	5.95	5.95
$SQRM_{0.75}$	**1.08**	**1.41**	**1.94**	**2.71**	**3.64**	**4.55**	**4.55**

Table 5: SQRM test risks on OGB-MolPCBA.

Alg.	Superquantile risk						
	0.0	0.25	0.50	0.75	0.90	0.99	1.0
ERM	**0.051**	0.081	0.12	0.20	0.36	1.30	16.04
$SQRM_{0.10}$	0.054	**0.071**	**0.10**	**0.17**	**0.29**	**0.90**	16.04
$SQRM_{0.20}$	0.061	0.081	0.12	0.20	0.36	1.28	8.78
$SQRM_{0.30}$	0.060	0.079	0.16	0.20	0.35	1.21	**7.03**
$SQRM_{0.40}$	0.060	0.079	0.11	0.20	0.35	1.20	7.70

Figure 3: **Test risk distributions on iWildCam and OGD-MolPCBA.** Each distribution shows that for larger values of α, QRM_α and $SQRM_\alpha$ tend to display superior tail performance relative to ERM.

that: **E:** increasing α leads to more consistent performance across domains, eventually forcing the model to ignore color and focus on invariant, shaped-based prediction; and **F:** a predictor's estimated accuracy-CDF depicts its accuracy-robustness curve, just as its estimated risk-CDF $F_{\widehat{\mathbb{T}}_f}$ depicts its risk-robustness curve. Note that $\alpha = 0.5$ gives the best worst-case risk over the two training domains—the preferred solution of DRO [60]—while $\alpha \to 1$ sacrifices risk for increased robustness or invariance.

6.2 Real-world datasets

We now evaluate our methods on the real-world or *in-the-wild* distribution shifts of WILDS [12]. We focus our evaluation on two large-scale classification datasets—iWildCam [112] and OGB-MolPCBA [113, 114]—both of which have numerous test domains, facilitating a comparison of the risk quantiles and distributions. Additional comparisons (e.g. using average accuracy) can be found in Appendix C.2. Our results demonstrate that, across two distinct data types (images and molecular graphs), QRM and SQRM offer superior quantile and superquantile performance respectively.

iWildCam. We first consider the iWildCam dataset, where the task is to classify animal species from camera-trap images, given 243 training domains and 48 test domains. In Table 2, we observe that QRM_α does indeed tend to optimize the α-risk quantile, with larger αs during training resulting in lower test-time risks at the corresponding quantiles. Similarly, Table 4 shows that $SQRM_\alpha$ effectively optimizes the superquantile function, with $\alpha = 0.75$ yielding particularly low values relative to ERM. In the two leftmost panes of Figure 3, we plot the (smoothed) test-time risk distributions. Here we see a clear trend: as α increases, the tails of the risk distribution tend to drop below ERM, which corroborates the superior quantile and superquantile performance reported in Tables 2 and 4.

Note that, in both Tables 2 and 4, QRM and SQRM tend to record lower *average* risks than ERM. This has several plausible explanations. First, the number of testing domains (48) is relatively small, which could result in a biased sample with respect to the training domains. Second, the test domains may not represent i.i.d. draws from $\mathbb{Q}$, with WILDS [12] test domains tending to be more challenging.

OGB-MolPCBA. We next consider the OGB-MolPCBA (or OGB) dataset, which is a molecular graph-classification benchmark containing 44,930 training domains and 43,793 test domains. In

contrast to our results on iWildCam, Table 3 and 5 show that ERM achieves the lowest average test risk on OGB, which supports the hypothesis that having access to a large sample of domains is necessary to generalize OOD. Both of these tables also indicate that QRM_α and $SQRM_\alpha$ achieve stronger quantile and superquantile performance relative to ERM. Of particular note is the fact that our methods significantly outperform ERM with respect to worst-case performance (columns/quantiles labeled 1.0); when QRM_α and $SQRM_\alpha$ are run with large values of α, we reduce the worst-case risk by more than a factor of two. In Figure 3, we again see that the risk distributions of QRM_α and $SQRM_\alpha$ have lighter tails than that of ERM.

The importance of multiple test domains. The analysis provided in Tables 2-5 and Figure 3 diverges from the standard evaluation protocol in the DG literature [12, 37]. Rather than evaluating an algorithm's performance *on average* across test domains, we seek to understand the distribution of its performance, in particular in the tails by means of the quantile and superquantile functions. This new evaluation protocol lays bare the importance of having access to multiple test domains, which is necessary for analyzing the tail behavior of a predictor's risk distribution. Indeed, as (DG) is not an average-case problem, solely reporting a predictor's test risk (or accuracy) averaged over multiple test domains can be misleading when assessing its ability to generalize OOD (as shown in Tables 2-5). This underscores the necessity of incorporating tail risk measures into a holistic evaluation protocol for DG, providing a more nuanced and complete picture of a predictor's ability to generalize OOD.

7 Discussion

On the assumption of i.i.d. environments. For α to approximate the probability of generalizing, training and test environments must be i.i.d.-sampled. While this is rarely true in practice—e.g. hospitals have shared funders, service providers, foundation dates, etc.—we can better satisfy this assumption by prescribing to a new data collection process in which we collect training environments which are representative of how the underlying system can change. For example: (i) randomly select 100 US hospitals; (ii) gather and label data from these hospitals; (iii) train our system with the desired α; (iv) deploy our system to all US hospitals where it will be successful with probability $\approx \alpha$. While this process may seem expensive, time-consuming and vulnerable (e.g. new hospitals), it offers a promising path to machine learning systems which *generalize with high probability*. Moreover, it is worth noting the alternative: prior works achieve generalization by assuming that only particular types of shifts can occur, e.g. covariate shifts [59, 115, 116], label shifts [116, 117], concept shifts [118], measurement shifts [19], mean shifts [111], shifts which leave the mechanism of Y invariant [9, 38, 40, 44], etc. In real-world settings, where the shift mechanisms are often unknown, such assumptions are both difficult to justify and impossible to test. Future work could look to relax the i.i.d. assumption by leveraging knowledge of environment dependencies (e.g. time).

α **controls the risk-robustness trade-off.** α approximates the probability with which our predictor will generalize with risk below the associated α-quantile value. Thus, α represents an interpretable parameterization of the often-inherent risk-robustness trade-off. Such interpretability may aid practitioners meet application-specific requirements on performance and/or robustness.

The wider value of risk distributions. As demonstrated in § 6, a predictor's risk distribution has value beyond quantile-minimization—it estimates the probability associated with each level of risk. Thus, regardless of the algorithm used, risk distributions can be used to analyze trained predictors.

A new evaluation protocol for DG. We have argued that the goal of DG should be to learn predictors which perform well *with high probability*, rather than in the worst-case or on-average. Evaluating this newly-sought ability requires new evaluation protocols which provide a more complete picture of predictor performance, going beyond the average or worst-case. To this end, we have proposed using *multiple* test domains to evaluate predictor performance at different quantiles.

Conclusion. We have presented QRM for achieving *probable* domain generalization, motivated by the argument that DG should seek predictors which generalize with high probability rather than in the worst-case or on-average. In particular, by relating training and test domains as draws from the same underlying meta-distribution, we proposed to learn predictors with minimal α-quantile risk under the training domains. Theoretically, we proved that α does indeed approximate the probability of generalizing when using the empirical version EQRM, and that $EQRM_\alpha$ recovers the causal predictor as $\alpha \to 1$. Empirically, we demonstrated the importance of having *multiple* test domains to evaluate the quantile performance of DG algorithms, and that our algorithms perform well in practice.

References

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015.

[2] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587): 484–489, 2016.

[3] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

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

[5] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision*, pages 456–473, 2018.

[6] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[7] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

[8] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine*, 15(11), 2018.

[9] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

[10] Timothy Niven and Hung Yu Kao. Probing neural network comprehension of natural language arguments. In *Association for Computational Linguistics*, pages 4658–4664, 2020.

[11] Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. Breeds: Benchmarks for subpopulation shift. *arXiv:2008.04859*, 2020.

[12] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.

[13] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *World Wide Web Conference*, pages 491–500, 2019.

[14] Matthew C Hansen, Peter V Potapov, Rebecca Moore, Matt Hancher, Svetlana A Turubanova, Alexandra Tyukavina, David Thau, Stephen V Stehman, Scott J Goetz, Thomas R Loveland, et al. High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160): 850–853, 2013.

[15] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Computer Vision and Pattern Recognition*, pages 6172–6180, 2018.

[16] Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time? In *International Conference on Computer Vision*, pages 9661–9669, 2021.

[17] Samil Karahan, Merve Kilinc Yildirum, Kadir Kirtac, Ferhat Sukru Rende, Gultekin Butun, and Hazim Kemal Ekenel. How image degradations affect deep cnn-based face recognition? In *International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2016.

[18] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20:1–25, 2019.

[19] Cian Eastwood, Ian Mason, Chris Williams, and Bernhard Schölkopf. Source-free adaptation to measurement shift via bottom-up feature restoration. In *International Conference on Learning Representations*, 2021.

[20] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[21] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[22] Alexander Robey, Hamed Hassani, and George J Pappas. Model-based robust deep learning: Generalizing to natural, out-of-distribution data. *arXiv preprint arXiv:2005.10247*, 2020.

[23] Jorge Jovicich, Silvester Czanner, Xiao Han, David Salat, Andre van der Kouwe, Brian Quinn, Jenni Pacheco, Marilyn Albert, Ronald Killiany, Deborah Blacker, et al. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *Neuroimage*, 46(1):177–192, 2009.

[24] Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical Physics*, 45(3): 1150–1158, 2018.

[25] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101–544, 2019.

[26] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–12. Association for Computing Machinery, 2020.

[27] Christian Wachinger, Anna Rieckmann, Sebastian Pölsterl, Alzheimer's Disease Neuroimaging Initiative, et al. Detect and correct bias in multi-site neuroimaging datasets. *Medical Image Analysis*, 67:101879, 2021.

[28] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *International Conference on Intelligent Transportation Systems*, pages 3819–3824, 2018.

[29] Georg Volk, Stefan Müller, Alexander von Bernuth, Dennis Hospach, and Oliver Bringmann. Towards robust CNN-based object detection through augmentation with synthetic rain variations. In *International Conference on Intelligent Transportation Systems*, pages 285–292, 2019.

[30] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, and W. Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. In *Machine Learning for Autonomous Driving Workshop, NeurIPS 2019*, 2019.

[31] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[32] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[33] Gustav Mårtensson, Daniel Ferreira, Tobias Granberg, Lena Cavallin, Ketil Oppedal, Alessandro Padovani, Irena Rektorova, Laura Bonanni, Matteo Pardini, Milica G Kramberger, et al. The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. *Medical Image Analysis*, 66:101714, 2020.

[34] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 2020.

[35] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

[36] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

[37] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[38] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris M Mooij. On causal and anticausal learning. In *ICML*, 2012.

[39] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, pages 3490–3497, 2018.

[40] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *International Conference on Machine Learning*, volume 139, pages 5815–5826, 2021.

[41] Roberto Tempo, Giuseppe Calafiore, and Fabrizio Dabbene. *Randomized algorithms for analysis and control of uncertain systems: with applications*. Springer, 2013.

[42] Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, pages 2512–2530. PMLR, 2019.

[43] Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. Pac-learning in the presence of adversaries. *Advances in Neural Information Processing Systems*, 31, 2018.

[44] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

[45] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 38(2): 550–560, 2018.

[46] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79 (1):151–175, 2010.

[47] Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[48] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

[49] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[50] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[51] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827. PMLR, 2013.

[52] E. Bareinboim and J. Pearl. Transportability from multiple environments with limited experiments: Completeness results. In *Advances in Neural Information Processing Systems 27*, pages 280–288, 2014.

[53] Kun Zhang, Mingming Gong, and Bernhard Schölkopf. Multi-source domain adaptation: A causal view. In *Twenty-ninth AAAI Conference on Artificial Intelligence*, 2015.

[54] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International Conference on Machine Learning*, pages 2839–2848. PMLR, 2016.

[55] B. Huang, K. Zhang, J. Zhang, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. Behind distribution shift: Mining driving forces of changes and causal arrows. In *IEEE 17th International Conference on Data Mining (ICDM 2017)*, pages 913–918, 2017.

[56] Marco C Campi and Simone Garatti. The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, 19(3):1211–1230, 2008.

[57] Federico Alessandro Ramponi. Consistency of the scenario approach. *SIAM Journal on Optimization*, 28(1):135–162, 2018.

[58] Daniel A Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004.

[59] Alexander Robey, George J. Pappas, and Hamed Hassani. Model-based domain generalization. In *Advances in Neural Information Processing Systems*, 2021.

[60] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[61] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[62] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837, 1956.

[63] Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[64] Bernard W Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26. CRC Press, 1986.

[65] Darrell Duffie and Jun Pan. An overview of value at risk. *Journal of derivatives*, 4(3):7–49, 1997.

[66] David Wozabal. Value-at-risk optimization using the difference of convex algorithm. *OR spectrum*, 34(4):861–883, 2012.

[67] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[68] Sergey Sarykalin, Gaia Serraino, and Stan Uryasev. Value-at-risk vs. conditional value-at-risk in risk management and optimization. In *State-of-the-art decision-making tools in the information-intensive age*, pages 270–294. Informs, 2008.

[69] Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2007.

[70] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34, 2021.

[71] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. Robust optimization. In *Robust optimization*. Princeton university press, 2009.

[72] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[73] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[74] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[75] Alexander Robey, Luiz Chamon, George J Pappas, Hamed Hassani, and Alejandro Ribeiro. Adversarial robustness with semi-infinite constrained learning. *Advances in Neural Information Processing Systems*, 34:6198–6215, 2021.

[76] Jia-Jie Zhu, Christina Kouridi, Yassine Nemmour, and Bernhard Schölkopf. Adversarially robust kernel smoothing. *arXiv preprint arXiv:2102.08474*, 2021.

[77] Natalia L Martinez, Martin A Bertran, Afroditi Papadaki, Miguel Rodrigues, and Guillermo Sapiro. Blind pareto fairness and subgroup robustness. In *International Conference on Machine Learning*, pages 7492–7501. PMLR, 2021.

[78] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 66–76, 2021.

[79] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018.

[80] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32, 2019.

[81] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9624–9633, 2021.

[82] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. *arXiv preprint arXiv:2007.02931*, 2020.

[83] Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. Adaptive methods for real-world domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14340–14349, 2021.

[84] Aniket Anand Deshmukh, Yunwen Lei, Srinagesh Sharma, Urun Dogan, James W Cutler, and Clayton Scott. A generalization error bound for multi-class domain generalization. *arXiv preprint arXiv:1905.10392*, 2019.

[85] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34, 2021.

[86] Yihang Chen, Grigorios Chrysos, and Volkan Cevher. A rate-distortion approach to domain generalization. 2021.

[87] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. *Advances in Neural Information Processing Systems*, 33:16096–16107, 2020.

[88] Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in Neural Information Processing Systems*, 33:3118–3129, 2020.

[89] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9619–9628, 2021.

[90] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[91] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018.

[92] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, pages 124–140. Springer, 2020.

[93] Masanori Koyama and Shoichiro Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. 2020.

[94] Lars Lindemann, Nikolai Matni, and George J Pappas. Stl robustness risk over discrete-time stochastic processes. *arXiv preprint arXiv:2104.01503*, 2021.

[95] Lars Lindemann, Alena Rodionova, and George Pappas. Temporal robustness of stochastic signals. In *25th ACM International Conference on Hybrid Systems: Computation and Control*, pages 1–11, 2022.

[96] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

[97] Alexander Robey, Luiz FO Chamon, George J Pappas, and Hamed Hassani. Probabilistically robust learning: Balancing average-and worst-case performance. *arXiv preprint arXiv:2202.01136*, 2022.

[98] Leslie Rice, Anna Bair, Huan Zhang, and J Zico Kolter. Robustness between the worst and average case. *Advances in Neural Information Processing Systems*, 34, 2021.

[99] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. *arXiv preprint arXiv:2007.01162*, 2020.

[100] Sebastian Curi, Kfir Y Levy, Stefanie Jegelka, and Andreas Krause. Adaptive sampling for stochastic risk-averse learning. *Advances in Neural Information Processing Systems*, 33: 1036–1047, 2020.

[101] Santiago Paternain, Miguel Calvo-Fullana, Luiz FO Chamon, and Alejandro Ribeiro. Safe policies for reinforcement learning via primal-dual methods. *IEEE Transactions on Automatic Control*, 2022.

[102] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

[103] Pavlo Krokhmal, Jonas Palmquist, and Stanislav Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of risk*, 4:43–68, 2002.

[104] Philippe Jorion. *Value at risk: the new benchmark for controlling market risk*. Irwin Professional Pub., 1997.

[105] Jaeho Lee, Sejun Park, and Jinwoo Shin. Learning bounds for risk-sensitive learning. *Advances in Neural Information Processing Systems*, 33:13867–13879, 2020.

[106] Justin Khim, Liu Leqi, Adarsh Prasad, and Pradeep Ravikumar. Uniform convergence of rank-weighted learning. In *International Conference on Machine Learning*, pages 5254–5263. PMLR, 2020.

[107] John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[108] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

[109] Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

[110] Juan L Gamella and Christina Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. *Advances in Neural Information Processing Systems*, 33:15464–15475, 2020.

[111] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021.

[112] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021.

[113] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

[114] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[115] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. MIT Press, 2008.

[116] Amos J Storkey. When training and test sets are different: characterising learning transfer. In *Dataset Shift in Machine Learning*, pages 3–28. MIT Press, 2009.

[117] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pages 3122–3130, 2018.

[118] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45:521–530, 2012.

[119] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, pages 583–588, 1997.

[120] Ichiro Takeuchi, Quoc V. Le, Timothy D. Sears, and Alexander J. Smola. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7(45):1231–1264, 2006.

[121] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Gert Lanckriet, and Bernhard Schölkopf. Kernel choice and classifiability for RKHS embeddings of probability distributions. *Advances in Neural Information Processing Systems*, 22, 2009.

[122] Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pages 322–348. PMLR, 2020.

[123] Núria Armengol Urpí, Sebastian Curi, and Andreas Krause. Risk-averse offline reinforcement learning. *arXiv preprint arXiv:2102.05371*, 2021.

[124] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[125] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[126] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer School on Machine Learning*, pages 169–207. Springer, 2003.

[127] Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

[128] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer, 2004.

[129] Ronald A DeVore and George G Lorentz. *Constructive approximation*, volume 303. Springer Science & Business Media, 1993.

[130] Keith Ball et al. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31(1-58):26, 1997. URL http://library.msri.org/books/Book31/files/ball.pdf.