A Constrained Bayesian Approach to Out-of-Distribution Prediction

Ziyu Wang[*1] **Binjie Yuan**[*1] **Jiaxun Lu**[2] **Bowen Ding**[3] **Yunfeng Shao**[2] **Qibin Wu**[3] **Jun Zhu**[#1]

[1]Dept. of Comp. Sci. & Tech., BNRist Center, Tsinghua-Huawei Joint Center for AI, THBI Lab, Tsinghua University
[2]Huawei Noah's Ark Lab
[3]Huawei Technologies Co., Ltd.

Abstract

Consider the problem of out-of-distribution prediction given data from multiple environments. While a sufficiently diverse collection of training environments will facilitate the identification of an invariant predictor, with an optimal generalization performance, many applications only provide us with a limited number of environments. It is thus necessary to consider adapting to distribution shift using a handful of labeled test samples. We propose a constrained Bayesian approach for this task, which restricts to models with a worst-group training loss above a prespecified threshold. Our method avoids a pathology of the standard Bayesian posterior, which occurs when spurious correlations improve in-distribution prediction. We also show that on certain high-dimensional linear problems, constrained modeling improves the sample efficiency of adaptation. Synthetic and real-world experiments demonstrate the robust performance of our approach.

1 INTRODUCTION

A crucial challenge in machine learning applications is to make predictions in a novel environment, with a data distribution different from those of the training environments [Quinonero-Candela et al., 2008, Blanchard et al., 2011]. In such scenarios, there often exist *spurious features* [Sagawa et al., 2020a] that exhibit environment-specific correlation structures to the target variable, which can be drastically different between training and test data. For example, in aggregated medical imaging datasets, factors such as radiographic positioning or projection often appear predictive about the diagnostic outcome, but only because both factors are correlated with the data source; machine learning models may thus learn "shortcuts" based on such features, leading to poor generalization [DeGrave et al., 2021].

A diverse body of literature is dedicated to this issue of out-of-distribution (OOD) prediction, with different assumptions introduced on the forms of distribution shift and the information available to aid generalization. We are primarliy interested in scenarios where a causally *invariant predictor* [Bühlmann, 2018, Rojas-Carulla et al., 2018] exists, and is reasonably performant across environments. In such cases, its recovery can be possible given a sufficiently diverse collection of training environments [Peters et al., 2016, Arjovsky et al., 2019]. It is often possible to have training data from multiple environments, such as in medical applications where hospitals form distinct environments; we can then apply algorithms such as invariant risk minimization [IRM, Arjovsky et al., 2019] and group distributionally-robust optimization [GDRO, Sagawa et al., 2020a].

Unfortunately, identification of the invariant predictor may require an excessive number of training environments: for d-dimensional linear models this may amount to $\mathcal{O}(d)$ environments [Rosenfeld et al., 2021]. Thus, in a large proportion of practical applications, we will find ourselves in an underidentified regime with an insufficient number of training environments, in which case the benefits of existing methods are far less clear. Indeed, Gulrajani and Lopez-Paz [2020] showed that across a number of benchmarks with a smaller number of environments, methods such as IRM and GDRO consistently fail to outperform an ERM baseline, even though the latter does not account for distribution shift. The challenge of underspecification can be fundamental; as we demonstrate in Lemma 1, in seemingly benign scenarios with $o(d)$ environments, there may exist spurious features that are indistinguishable from invariant features, by *any* statistical procedure working on finite data.

In light of the practical need for OOD prediction given few environments and the inherent difficulties of generalization, it is thus necessary to take a step back and consider *adapting* a learned model to a target environment, using a handful

[*]Equal contribution. [#]Corresponding author.

Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023), PMLR 216:2248–2258.

of labeled samples. Such samples are often available, for example if our deployment process involves first testing the model in the target environment; in such cases, we can simply set out a few samples for adaptation.

It is natural to consider a Bayesian approach given the underidentified nature of our problem, as is also advocated by Lee et al. [2022] in a connected but different setting. The Bayesian formulation may also appear desirable due to the interpretation as sequential belief updates — the posterior given training data "naturally" serves as a prior during adaptation. Unfortunately, the Bayesian approach can be inherently flawed for our purpose, as long as there exists a non-negligible gap between the in-distribution performance of the invariant and non-invariant models. As we shall discuss in Section 3.1, this gap will get amplified by a scaling of evidence that is required in (generalized) Bayesian modeling, and cause the posterior to remain concentrated on non-invariant models until a very large number of adaptation samples have been seen. This is concerning due to the prevelance of such *performance gaps*; indeed, they are part of the reason for the failure of domain generalization algorithms [e.g., Rosenfeld et al., 2021, Sagawa et al., 2020b].

In this work, we attempt to address this issue by proposing a principled approach for the adaptation task. We assume the knowledge of a lower bound of the invariant predictor's performance. Such knowledge is often possible, given our implicit assumption that the invariant predictor has an acceptable performance. We then use the training environments to define *constraints*: we restrict to the subset of models that do not perform significantly worse than the lower bound, across all training environments. This ensures the invariant predictor presents in the constraint set with high probability, and is weighed similarly to the non-invariant predictors, even though the latter would have induced a much better likelihood on training data. Consequently, efficient adaptation can be achieved.

Our method can be justified in many ways, by considering its behavior in the presence of performance gaps as sketched above, or by relating it to a relaxed formulation of GDRO. We complement these justifications with an asymptotic analysis, showing that in certain asymptotics for high-dimensional linear models, *adaptation with constrained models may achieve a vanishing estimation error with a relatively small number of training environments, whereas using neither training nor adaptation data alone cannot guarantee convergence.* This result improves the understanding of OOD learning, by showing that a smaller number of training environments can still be useful.

We evaluate the proposed method through synthetic and real-world experiments. On several image classification tasks where off-the-shelf domain generalization algorithms struggle to improve over ERM, our method delivers significant improvement, with only a handful of adaptation samples. Moreover, among all the adaptation algorithms evaluated, our method is the only one with reliable performance across all settings; in contrast, the baseline procedures fail intermittently, in different settings, which can be attributed to their less principled nature.

The rest of this paper is structured as follows: in Sec. 2 we review the setup of OOD generalization and discuss its hardness. Sec. 3 discusses the pitfall of standard Bayesian modeling and introduces our method, which is further justified in Sec. 4 through asymptotic analyses. We review related work in Sec. 5, present empirical evaluations in Sec. 6, and provide concluding remarks in Sec. 7.

2 OOD GENERALIZATION AND ITS HARDNESS

Notations We adopt the following notations in the paper: $[n] := \{1, \dots, n\}$. $\asymp, \lesssim, \gtrsim$ denote (in)equality up to constants. $c_1, \dots,$ denote universal constants. For finite-dimensional vectors, $\|\cdot\|_2$ denotes the Euclidean norm.

Invariant models and OOD generalization Consider a prediction task with training data from m environments: $\mathcal{D}_{tr} := \{\{(x_i^e, y_i^e) \sim P_e : i \in [n_e]\} : e \in \mathcal{E}_{tr}\}$, where $|\mathcal{E}_{tr}| = m$. We are interested in an out-of-distribution test environment where the data comes from a different P_*. We assume the existence of an *invariant predictor* that only depends on the input x through some $\Phi_{inv}(x)$, such that $p_e(y \mid \Phi_{inv}(x)) \equiv p(y \mid \Phi_{inv}(x))$ is invariant across all environments. We also assume that $\Phi_{inv}(x)$ can be reasonably informative about y. Such $\Phi_{inv}(x)$ are named *invariant features*, in contrast to the *spurious features* $\Phi_{spu}(x)$ which induce different $p_e(y \mid \Phi_{spu}(x))$ across environments, and hinder generalization when they are included in a predictor.

A variety of approaches have been proposed for learning the invariant predictor. Of particular interest is the method of *group distributionally-robust optimization* (GDRO), which minimizes the worst-case risk across training environments:

$$\min_{f \in \mathcal{H}} \max_{e \in \mathcal{E}_{tr}} \hat{R}_e(f), \tag{GDRO}$$

and invariant risk minimization (IRM):

$$\begin{aligned} &\min_{f = w \circ \Phi \in \mathcal{H}} \sum_{e \in \mathcal{E}_{tr}} \hat{R}_e(w \circ \Phi), \\ &\text{subject to} \quad w \in \arg\min_{w'} \hat{R}_e(w' \circ \Phi), \ \ \forall e \in \mathcal{E}_{tr}. \end{aligned} \tag{IRM}$$

In the above, $\hat{R}_e(f) := \frac{1}{n_e} \sum_{i=1}^{n_e} \ell(f(x_i^e), y_i^e)$ denotes the empirical risk for an environment e, ℓ denotes a suitable loss function, and $\mathcal{H}$ is our hypothesis space. For IRM, Φ and w denote the learned invariant features and the optimal predictor atop them.

Hardness of OOD generalization It is intuitive that an invariant predictor may be recovered, given a large and diverse collection of training environments. For IRM and certain linear models with dimensionality d, this amounts to having $m \asymp d$ environments that are independent in a certain sense [Arjovsky et al., 2019]. Unfortunately, such requirements can be unrealistic for high-dimensional data, and/or nonlinear models, and with a smaller m the empirical performance of domain generalization algorithms can often be disappointing: Gulrajani and Lopez-Paz [2020] show that a wide range of methods may fail to match the performance of an empirical risk minimization (ERM) baseline.

Let us illustrate the hardness of invariant prediction using the following example, adapted from Rosenfeld et al. [2021]:

Example 1. Consider a classification problem with data generated as follows:

$$\bar{\beta}^e_{spu} \sim \mathcal{N}(0, \tau_s^2 I) \in \mathbb{R}^{d_{spu}},\ y^e_i \sim \mathrm{Unif}\{\pm 1\},$$
$$x^e_i = \begin{bmatrix} x^e_{i,inv} \\ x^e_{i,spu} \end{bmatrix} \sim \mathcal{N}\left(y^e_i \begin{bmatrix} \bar{\beta}_{inv} \\ \bar{\beta}^e_{spu} \end{bmatrix}, \begin{bmatrix} \sigma_i^2 I & 0 \\ 0 & \sigma_s^2 I \end{bmatrix} \right),$$

where $\tau_s, \sigma_s, \sigma_i > 0$, and $\bar{\beta}_{inv} \in \mathbb{R}^{d_{inv}}$ is fixed. When $m < d_{spu}/4$, the vectors $\{\bar{\beta}^e_{spu}\}$ are linearly independent with high probability [Wainwright, 2019, chapter 6]. Thus, by Theorems 5.1 and 5.3 in Rosenfeld et al. [2021], all of ERM, IRM and GDRO will learn a non-invariant predictor. We provide further insights through the following:

Lemma 1. *In the setting of Example 1, let*

$$x^e_{pe,i} := \alpha \textstyle\sum_{e\in\mathcal{E}_{tr}} (\bar{\beta}^e_{spu})^\top x^e_{i,spu}, \text{ with } \alpha \neq 0.$$

be a "purely environmental" feature. Then,

(i) A classifier based on x_{pe} alone will achieve a vanishing error, if $m \ll d_{spu} \min\{1, (\tau_s/\sigma_s)^2\}$.

(ii) For all $e \in \mathcal{E}_{tr}$, denote the marginal distribution of (y, x^e_{inv}, x^e_{pe}) by $p_{e,marg}$. Then, w.p. $\geq 1 - e^{-m/18}$ w.r.t. $\{\bar{\beta}^e_{spu}\}$ there exists some $\tilde{e} \in \mathcal{E}_{tr}$ s.t.

$$\mathrm{KL}\Big(\bigotimes_{e\in\mathcal{E}_{tr}} p_{e,marg} \,\Big\|\, p^{\otimes m}_{\tilde{e},marg} \Big) \leq \frac{256m}{\sigma_s^2 d_{spu}}. \quad (1)$$

Consequently, given a training sample with size

$$\max_{e\in\mathcal{E}_{tr}} n_e \ll \sigma_s^2 d_{spu}/m, \quad (2)$$

__no__ statistical test with a size of $o(1)$ could reject x_{pe} as a non-invariant feature w.p. $\geq o(1)$.

Proof. See the supplementary material. □

The above result highlights the hardness of OOD generalization in high dimensions. It shows in the $m \ll d_{spu}$ regime the existence of a spurious feature that has an arbitrarily high predictive power across $e \in \mathcal{E}_{tr}$, yet can be *indistinguishable from invariant features* given finite samples. In reality, the sample size threshold will be much higher than (2), since for features learned from finite data *it is only valid to test for approximate invariance*; see the supplementary material for a detailed discussion.[1] It should be noted that quantitatively similar results do not always hold, across all linear models: Chen et al. [2022] showed that under certain data generating processes, identification may become possible when $m = \mathcal{O}(\log d)$. Still, it remains concerning that such a pathology arises from a seemingly benign setting, with i.i.d. training environments and x_{pe} constructed by a simple averaging. Also note while past works have studied adaptation based on unlabeled test samples [Zhang et al., 2021], it would be ineffective on this setup, since the input has the same distribution across all environments.

We note that multiple mechanisms exist that may explain the hardness of OOD generalization, and the possible (in-distribution) *performance gap* between the invariant and non-invariant predictors: they may be inherent to the data distribution as demonstrated above, or they can arise from inappropriate model specifications, which may lead to the memorization of data [Sagawa et al., 2020b], undesirable margin-maximization behavior [Nagarajan et al., 2020, Wald et al., 2023], or simply a larger approximation error for the invariant predictors. We take an agnostic view to the cause, but stress the ubiquity of hard-to-learn problems: as exemplified by claim (ii) above, there are many scenarios where generalization to completely unseen environments is fundamentally difficult. Instead, we may have to take a step back, and seek additional information about the target environment.

3 ADAPTING TO ENVIRONMENT SHIFT WITH CONSTRAINED BAYESIAN MODELS

In many applications, it is possible to collect a handful of labeled samples from the test environment before deploying the model; for example, such samples may come "for free" if the deployment process involves first evaluating the model in the test environment. In light of the inherent difficulties of generalization to unseen environments, it is reasonable to study the use of such samples to adapt our model to the shifted environment.

3.1 WHY NOT (GENERALIZED) BAYES?

Before presenting our method, let us first consider a naïve alternative which employs as the prior for adaptation a (generalized/Gibbs) posterior from training data, which is

[1] It is also clear from the proof that if $\{\bar{\beta}^e_{spu}\}$ are exactly orthonormal, indistinguishability will hold for all finite n_e.

then updated with samples from the test environment. Let $\mathcal{D}_{ad} := \{(x_i^*, y_i^*) \sim P_* : i \in [n_*]\}$ denote the *adaptation samples*, and $\theta \in \Theta$ denote the parameters of a predictor f_θ. The updated posterior is then

$$p_{GB}(d\theta \mid \mathcal{D}_{tr}, \mathcal{D}_{ad}) \propto \pi(d\theta) e^{-\mathcal{L}(\theta;\mathcal{D}_{tr})} \prod_{i=1}^{n_*} e^{-\ell(y_i^*, f_\theta(x_i^*))}.$$

In the above, the "initial prior" π represents our subjective belief before seeing any data, $\mathcal{L}(\theta;\mathcal{D}_{tr})$ can be any **properly scaled** training objective, and $\ell(y_i^*, f_\theta(x_i^*))$ denotes an arbitrary loss. With $\ell(y_i^*, f_\theta(x_i^*)) \leftarrow -\log p(y_i^* \mid f_\theta(x_i^*))$, $\mathcal{L}(\theta;\mathcal{D}_{tr}) \leftarrow \sum_{e\in\mathcal{E}_{tr}} \sum_{i=1}^{n_e} \ell(y_i^e, f_\theta(x_i^e))$ we recover the standard Bayesian posterior, while using (GDRO) or (IRM) for $\mathcal{L}$, or using a different ℓ, will lead to different generalized posteriors [Zhang, 2006, Bissiri et al., 2016]. Note that this generalized posterior can also understood from a variational perspective with proper posterior regularization [Zhu et al., 2014].

Importantly, in (generalized) Bayesian modeling, the scale of $\mathcal{L}$ should be proportional to, or at least increasing w.r.t. the training sample size, as otherwise the "adaptation-time prior" $\pi_{ad,GB}(d\theta) \propto \pi(d\theta)e^{-\mathcal{L}(\theta;\mathcal{D}_{tr})}$ would be equivalent to the original π, rendering the training data useless. With an additive $\mathcal{L}$ such as in ERM or (IRM), the linear scaling is also desirable because it allows us to maintain a coherence property of sequential Bayesian updates [Bissiri et al., 2016].

It is precisely this necessary scaling that make the generalized Bayesian approach unsuitable for our adaptation goal. The problem is that in many OOD problems, there exists a small but non-negligible gap between the in-distribution performance of the invariant predictor and a non-invariant predictor, as we discussed at the end of Section 2; and such a gap gets amplified by the scaling of the objective:

Example 2. As a pedagogical example, consider a classification task with ℓ being the 0/1 loss, $n_e \equiv 10^5$, and a two-point prior π supported on the invariant predictor and a non-invariant predictor: $\pi = \mathrm{Unif}\{\theta_{non-inv}, \theta_{inv}\}$, where

$$R_*(\theta_{non-inv}) - R_*(\theta_{inv}) \geq 0.99,$$
$$\min_{e\in\mathcal{E}_{tr}} (\hat{R}_e(\theta_{inv}) - \hat{R}_e(\theta_{non-inv})) \geq 0.01.$$

(Note the shorthand notation $R_{(\cdot)}(\theta) := R_{(\cdot)}(f_\theta)$, and R_* denotes the population risk on the test environment.) Let $\mathcal{L}$ be scaled by n_e. Then we have $\frac{\pi_{ad,GB}(\{\theta_{inv}\})}{\pi_{ad,GB}(\{\theta_{non-inv}\})} = e^{10^5\times 0.01}$ for the adaptation-time prior, and the log posterior mass ratio is approximately $10^3 - 0.99n_*$. Therefore, even though $\theta_{non-inv}$ has catastrophic performance on the test environment, it would take more than 10^3 adaptation samples for p_{GB} to concentrate to the right parameter.[2]

While it is certainly possible to alleviate this issue with more heuristics, e.g., by switching to a smaller scaling, it is difficult to determine a sensible scheme that facilitates efficient adaptation; a slower scaling also discounts the training data "as a whole", making them less useful for the invariant features, and for identifying part of the spurious correlations that could have been identified from training data. The awkward situation reflects the inherently different roles of training and adaptation samples, which necessitates a different treatment for the distinct forms of evidence they provide.

3.2 CONSTRAINED BAYESIAN MODELING

In light of the pathological inefficiency of the generalized Bayesian approach, we propose an alternative which is to use the training environments to define constraints. Concretely, let $\rho \geq R_e(f_{inv})$ be a prespecified upper bound for the risk of the invariant predictor. We define our predictive distribution using the constrained posterior

$$p_C(d\theta \mid \mathcal{D}_{tr}, \mathcal{D}_{ad}) \propto \pi(d\theta)\mathbf{1}_{\{\theta\in\mathcal{C}_{tr}\}} \prod_{i=1}^{n_*} e^{-\ell(y_i^*, f_\theta(x_i^*))},$$
$$\text{where} \quad \mathcal{C}_{tr} := \left\{\theta : \max_{e\in\mathcal{E}_{tr}} \hat{R}^e(f_\theta) \leq \rho + \varepsilon_n\right\}$$

is the constraint set, and $\varepsilon_n \to 0$ covers the small sampling error $|\hat{R}_e(f_{inv}) - R_e(f_{inv})|$ so that we can have $\theta_{inv} \in \mathcal{C}_{tr}$ with high probability.[3] In many applications we have knowledge of a good choice for ρ, due to the implicit assumption that the invariant predictor has an acceptable performance; e.g., in classification problems where the performance gap between the invariant and non-invariant classifiers can be attributed to various types of label noise, we can often upper bound the noise level based on our domain knowledge. It is also possible to utilize less reliable sources of information about ρ, by viewing ρ as a model parameter and equipping it with a prior. Alternatively, we may simply set ρ to be larger than the risk of the ERM to achieve a better trade-off between in-distribution and OOD performance; this approach will be evaluated in section 6.2.

When ρ is small, any $\theta \in \mathcal{C}_{tr}$ will correspond to an approximate optima for (GDRO). Thus, *the constrained posterior is a natural generalization of GDRO*, and will not perform significantly worse, which is useful if the training data turns out to be informative. In the underidentified regime, the constrained posterior allows for more efficient adpatation, by relaxing the optimization problem and modeling the uncertainty in training data. Comparing with the naïve Bayesian approach, the constrained posterior is based on an adaptation-time prior $\pi_{ad,C}(d\theta) \propto \pi(d\theta)\mathbf{1}_{\theta\in\mathcal{C}_{tr}}$ that does not introduce additional weighting to models in the constraint set; this allows us to avoid the pathological behavior of the former: returning to Example 2, we can see

[2] Note that while the example concerns generalized Bayesian posteriors, a similar pathology exists for the respective point estimators, due to the exponential concentration of the loss functions.

[3] For subgaussian loss we can choose $\varepsilon_{n,e} \propto n_e^{-1/2}\sqrt{\log m}$.

that the constrained posterior only requires $\mathcal{O}_p(1)$ samples to converge to the correct prediction.

3.3 ALGORITHM IMPLEMENTATION

We draw approximate samples from the constrained posterior using a simple algorithm that augments Langevin Monte Carlo (LMC) with line search: at each iteration, we choose within a prescribed range the largest step-size s.t. the LMC update could stay in the constraint set. The process is listed as Algorithm 1. We run multiple LMC chains in parallel, a nd use the obtained samples $\{\theta_K^{(j)} : j \in [J]\}$ to define the predictor $\tilde{p}_C(y^* \mid x^*) = \frac{1}{J}\sum_{j=1}^{J} p(y^* \mid f_{\theta_K^{(j)}}(x^*))$.

Algorithm 1 Approximate inference for the constrained posterior.

```
Require: Training and adaptation samples $(\mathcal{D}_{tr}, \mathcal{D}_{ad})$, loss
    $\ell$, prior $\pi(d\theta)$, $K, \rho, \varepsilon_n, \varepsilon_b > 0$, $\{\bar{\eta}_k : k \in [K]\}$
Ensure: Approximate sample $\theta_K \sim \tilde{p}_C \approx p_C$
initialize $\theta_0$ using e.g., ERM on $\mathcal{D}_{tr}$   ▷ proper choices
    for $(\rho, \varepsilon)$ will ensure $\theta_0 \in \mathcal{C}_{tr}$
for $k \leftarrow 1, \ldots, K$ do
    draw $z_k \sim \mathcal{N}(0, I)$
    $g_k \leftarrow \nabla_\theta \sum_{i=1}^{n_*} \ell(y_i^*, f_\theta(x_i^*))|_{\theta=\theta_{k-1}}$
    $\theta_k \leftarrow \theta_{k-1} - \eta_k g_k + \sqrt{2\eta_k} z_k$, where $\eta_k \in [0, \bar{\eta}_k]$
    is the largest number s.t. $\theta_k \in \mathcal{C}_{tr}$   ▷ $\eta_k$ is determined
    (up to an error of $\varepsilon_b$) using binary search
end for
return $\theta_K$
```

Intuitively, the algorithm can be viewed as simulating a reflected Langevin equation [Lions and Sznitman, 1984, Bubeck et al., 2018], which is the constrained counterpart to the standard Langevin dynamics. Note that refined numerical schemes exist if we can compute the boundary of $\mathcal{C}_{tr}$ efficiently [Bubeck et al., 2018, Sato et al., 2022], which is possible in settings like linear models with a convex ℓ. Alternative constraint sampling algorithms, such as Zhang et al. [2022], can also be utilized; we opt for algorithm 1 merely for its simplicity. If needed, we can improve its computational efficiency through standard means, by introducing preconditioning, stochastic gradients, or by replacing the training set with a uniformly random or curated subset [e.g., Bachem et al., 2017].

4 THEORETICAL ANALYSIS

We have motivated our method by connecting it to a relaxation of GDRO, and by considering its small-sample behavior in simple settings. We now provide further justifications, by showing that on a family of linear models, constrained modeling in general can improve the sample efficiency even when the adaptation sample size is large.

Analysis setup We consider a regression setup with data generated as follows:

$$\bar{\beta}_{spu}^e \sim \mathcal{N}(0, d_{spu}^{-1} I),\ \mathbf{x}_i^e = M \begin{bmatrix} \mathbf{x}_{inv,i}^e \\ \mathbf{x}_{spu,i}^e \end{bmatrix} \sim \mathcal{N}(0, I),$$
$$\mathbf{y}_i^e \sim \mathcal{N}(\bar{\beta}_{inv}^\top \mathbf{x}_{inv,i}^e + (\bar{\beta}_{spu}^e)^\top \mathbf{x}_{spu,i}^e, \sigma_y^2). \tag{3}$$

In the above, $e \in \mathcal{E}_{tr}$ indexes the training domain, $\bar{\beta}_{inv}$ is an arbitrary, fixed vector with norm $\mathcal{O}(1)$, $\mathbf{x}_{inv,i}^e \in \mathbb{R}^{d_{inv}}, \mathbf{x}_{spu,i}^e \in \mathbb{R}^{d_{spu}}$ are the invariant and spurious features, and M is a mixing matrix assumed to be invertible and well-conditioned. Test data $(\mathbf{x}^*, \mathbf{y}^*)$ is generated similarly using $\bar{\beta}_{spu}^*$ in place of $\bar{\beta}_{spu}^e$, which we assume is an *arbitrary, fixed* vector. We use the square loss $\ell(s,t) = (s-t)^2$, and assume access to infinite training samples for simplicity. The invariant predictor is parameterized by $\bar{\theta}_{inv} = M^{-\top}(\bar{\beta}_{inv}, 0)$.

As discussed in Section 2, on similar linear problems, identification of $\bar{\theta}_{inv}$ may require $m = \mathcal{O}(d)$ domains. To understand the necessity on this setup, observe that when $m \ll d$, the vectors $\{\bar{\beta}_{spu}^e : e \in \mathcal{E}_{tr}\}$ are approximately orthonormal [Wainwright, 2019, Ch. 6], and that when they are exactly orthonormal, ERM and GRO will both identify the parameter $\tilde{\theta} = M^{-\top}(\bar{\beta}_{inv}, \frac{1}{m}\sum_{e\in\mathcal{E}_{tr}} \bar{\beta}_{spu}^e)$ which leads to

$$R_e(\tilde{\theta}) \equiv \sigma^2 + \frac{m-1}{m} \le \sigma^2 + 1 \equiv R_e(\bar{\theta}_{inv}),\ \ \forall e \in \mathcal{E}_{tr}.$$

(IRM) learns the same $f_{\tilde{\theta}}$, which can fulfill its constraint using $\Phi(x) = (\bar{\beta}_{inv}^\top M^{-1}x, \frac{1}{m}\sum_{e\in\mathcal{E}_{tr}}(\bar{\beta}_{spu}^e)^\top M^{-1}x)$. By the arbitrariness of $\bar{\beta}_{spu}^*$, the predictor $f_{\tilde{\theta}}$ may incur an arbitrarily high error on new environments.

Improved convergence of a constrained estimator We now present our analysis. For technical simplicity, we study a constrained *point estimator*:

$$\hat{\theta} := \mathop{\arg\min}_{\theta\in\mathcal{C}_{tr}\cap\Theta} \sum_{i=1}^{n_*} \ell(y_i^*, \theta^\top x_i^*), \text{where } \Theta := \{\theta : \|\theta\|_2 \le U\}$$

parameterizes our hypothesis space, and $U > \|M\|^{-1}$ is a constant. We then have the following:

Proposition 2. *Suppose the data is generated as as above,* $\bar{\beta}_{spu}^* \in \mathbb{R}^{d_{spu}}$ be arbitrary, *and $\hat{\theta}$ is defined as above. Let $\bar{f}^*$ be the Bayes predictor on the test domain. Then there exist universal constants $c_1, c_2, c_3 > 0$ s.t. when $n_* \ge 3d$ we have, with probability $\ge 1 - n_*^{-9}$,*

$$\begin{aligned} R_*(f_{\hat{\theta}}) - R_*(\bar{f}^*) &\le c_1 \inf_{\theta'\in\Theta\cap\mathcal{C}_{tr}} (R_*(f_{\theta'}) - R_*(\bar{f}^*)) + \epsilon_{n_*}^2 \\ &\le c_1(R_*(f_{\bar{\theta}_{inv}}) - R_*(\bar{f}^*)) + \epsilon_{n_*}^2, \end{aligned}$$

where $$\epsilon_n^2 = c_2 \frac{\sigma_y d_{inv} + \log n_*}{n_*} + c_3\sigma_y \min\left\{\sqrt{\frac{d_{spu}\log m}{n_* m}}, \frac{2^{-m/d_{spu}} d_{spu}}{n_*}\right\}.$$

Proof. The full proof is in supplementary material. Its main idea is that for any $\theta = M^{-\top}(\beta_i, \beta_s)$ with $\beta_s \neq 0$, we have

$$\mathbb{P}_{\mathcal{D}_{tr}}(\theta \in \mathcal{C}_{tr}) \leq \min\{e^{-md_{spu}\|\beta_s\|_2^2}, 2^{-m}\}.$$

This allows us to derive high-probability bounds on the reduced complexity of $\Theta \cap \mathcal{C}_{tr}$. □

Proposition 2 establishes an oracle inequality, which allows us to compare the performance of the constrained estimator with the invariant predictor. At the claimed probability, the unconstrained maxmimum likelihood estimate (MLE) that does not utilize training data achieves an estimation error of

$$\epsilon_n'^2 = c_4 \frac{\sigma_y(d_{inv} + d_{spu}) + \log n}{n}.$$

Therefore, we can see that the constrained formulation (at least) improves the efficiency in estimating the spurious component of the model. The improvement is most interesting when $d_{spu} \gg d_{inv}$; in particular, observe that

(i) When $n_* \asymp d_{spu}$, unconstrained MLE will fail to converge as we have $\epsilon_n' \asymp 1$. In contrast, the constrained estimator satisfies $\epsilon_n^2 = \tilde{\mathcal{O}}(m^{-1/2})$. This is useful on high-dimensional problems when we only have a moderate number of environments, i.e., when $1 \ll m \ll d_{spu}$: given the previous discussion on IRM and GDRO, we can see that *using neither the training or adaptation samples alone cannot guarantee convergence* in this regime, which demonstrates the efficacy of constrained modeling.

(ii) Even as $n_* \gg d_{spu}$ becomes larger, the training data still remains useful, as it improved the estimation error for the spurious component by a factor of $2^{-m/d_{spu}}$. When m/d_{spu} is small, the expansion $2^{-m/d_{spu}} d_{spu} \approx (1 - m/d_{spu} \log 2) d_{spu}$ shows that each environment roughly removes one "degree of freedom" from the adaptation process.

Our choice to analyze high-dimensional linear problems follows previous works in this area [e.g., Arjovsky et al., 2019, Sagawa et al., 2020b, Rosenfeld et al., 2021]. The linear setup is also justified by the observation that the last layer of DNN models often retain sufficient information about the invariant features [Kirichenko et al., 2022], even though our algorithm is not restricted to linear models. The regression setup is adapted from Arjovsky et al. [2019]; our assumption of i.i.d. training environment is stronger. However, our setup remains non-trivial, as existing domain generalization approaches still underperform the invariant predictor by a notable margin. (Also note that we did not impose any restrictions on the test environment.) It may be possible to demonstrate similar sample efficiency gains in other scenarios, but they need to be established on a case-by-case basis. Another limitation is that for simplicity, we did not analyze the efficiency gain in estimating the invariant component; numerical simulations will provide a more complete understanding on the benefits of our method.

5 RELATED WORK

Our work is motivated by the practical need of deploying machine learning models to OOD environments, given data from a small collection of training environments and assuming the presence of spurious correlations. Our setup is thus connected to, but different from, a few lines of works on spurious correlations and/or transfer learning; given the vast literature, we refer readers to Wilson and Cook [2020], Wang et al. [2022], Jiang et al. [2022] for a detailed review. Comparing with most works on spurious correlations, we do not assume the training data contains sufficient information for learning an invariant predictor, a common situation as discussed in Section 2. Comparing with the transfer learning literature, we have a specific focus on spurious correlations, as is also noted in Kirichenko et al. [2022].

The recent works of Kirichenko et al. [2022], Ye et al. [2022], Lee et al. [2022] operate in a similar underspecified setting and also utilize adaptation samples, but all of them assume a single training environment. We have demonstrated how environment annotations can be utilized to improve adaptation performance. Empirically, our method also outperforms the adaptation procedures in Kirichenko et al. [2022], Lee et al. [2022] in a multi-environment setup (section 6.2). Still, our general idea may also be interesting for single-environment problems, for which we may define constraints using alternative characterizations for the invariant predictor (e.g., as in IRM) to address the issue of possible performance gaps. It may be interesting for future work to combine the development in Ye et al. [2022], Lee et al. [2022] with our framework.

Lin et al. [2022], Lee et al. [2022] have investigated uncertainty modeling for OOD generalization and are broadly related to our work in this aspect, but both have a different focus: Lin et al. [2022] on finite-sample estimation error of the IRM objective, and Lee et al. [2022] on the computational cost of Bayesian inference. As such, neither work addresses the issue of potential performance gaps between the invariant and non-invariant predictors, which, as we have discussed in Section 3.1, requires a careful treatment. Finally, Wald et al. [2023] studied OOD learning in the presence of similar performance gaps, but focused on scenarios where the invariant predictor is *identifiable* by alternative strategies (e.g., by matching the class-conditional distributions of features across environments). As we discussed in Section 2, identifiability is not always a realistic assumption.

6 EXPERIMENTS

In this section we evaluate our method empirically, on synthetic data, benchmark datasets and a real-world application. Code for the experiments can be found at `https://gitee.com/mindspore/models/tree/master/research/cv/ConstrainedBayesian`.

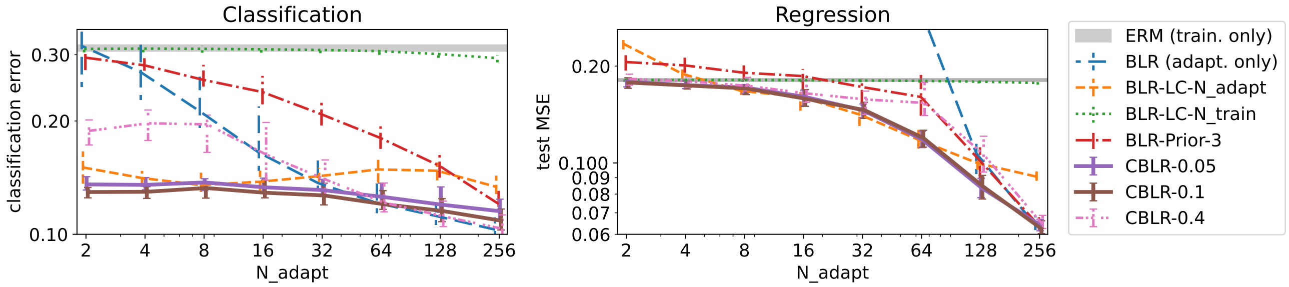

Figure 1: Synthetic experiment: test error vs. adaptation sample size for classification and regression. We report the median and (20th, 80th) percentile across 32 independently sampled adaptation sets. Plots are slightly shifted for visibility.

6.1 SYNTHETIC EXPERIMENTS

For the synthetic experiments, we consider the classification setup in Example 1 and the regression setup in Section 4.

Experiment setup The data generating processes are instantiated as follows: we set $\sigma_i = 7.5, \sigma_s = 3, \tau_s = 1$ for classification, and $\sigma_y = 0.5$ for regression. For both sets of experiments we use $m = 3, n_e = 6000, d_{inv} = 20, d_{spu} = 50$, $\bar{\beta}_{inv} \sim \mathcal{N}(0, 4\sigma_v^2 I)$, and $\bar{\beta}^*_{spu} := \frac{1}{2m}\sum_{e\in\mathcal{E}_{tr}} \bar{\beta}^e_{spu}$. The supplementary material includes additional experiments covering different parameters and choices of $\bar{\beta}^*_{spu}$.

We employ a Bayesian linear model with a Gaussian prior; the prior variance is set to match the norm of the empirical risk minimizer. We use a correctly specified likelihood, i.e., normal for regression and logistic for classification. For our method, We define the constraint set using the 0/1 loss for classification and the square loss for regression, and set $\rho + \epsilon_n := \max_{e\in\mathcal{E}_{tr}} \hat{R}_e(\bar{\theta}_{inv}) + \delta$, where $\delta \in \{0.05, 0.1, 0.4\}$ models our imprecise knowledge about $R_e(\bar{\theta}_{inv})$.

We compare our method (CBLR-δ) to the following:

- BLR: Bayesian inference using the same Gaussian prior, and only the adaptation samples for the likelihood.
- BLR-LC: the (generalized) Bayesian approach discussed in Section 3.1. The method involves scaling the empirical risk by a factor N; in the text we consider $N := n_e$, which corresponds to standard Bayesian modeling, and $N := n_*$, a heuristic that may allow for faster adaptation.
- BLR-Prior-α: another heuristic approach that replaces the prior mean with the empirical risk minimizer $\hat{\theta}_{ERM}$ on training data, and scales the prior variance by α^{-2}.

For all baselines, we run Metropolis-adjusted Langevin algorithm (MALA) using 10^4 iterations and 50 parallel chains. Based on the MALA acceptance rate, we set the step-size to $\bar{\eta}_{k,u} \equiv 0.016/n_*$. For our method we set the step-size upper bound to $\bar{\eta}_k := \bar{\eta}_{k,u}/4$ and use 4×10^4 iterations. The Markov chains are initialized at $\hat{\theta}_{ERM}$ for our method and BLR-Prior, and the minimizer of an interpolated empirical risk for BLR-LC. We also report the performance of $\hat{\theta}_{ERM}$ for reference.

Results and discussion The results are plotted in fig. 1. Our method demonstrates competitive performance across all choices of adaptation sample size, and is reasonably insensitive to the choice of the performance bound hyperparameter. All baselines have less reliable performance: BLR perform notably worse when n_* is small. Adaptation of the standard posterior (BLR-LC-N_train) is extremely slow since we have $n_* \ll n_e$, in line with our discussion in Section 3.1. With the heuristic scaling in BLR-LC-N_adapt, the performance becomes better at moderate sample sizes, but still not as good at smaller or larger n_*; the former is because the variance of the adaptation likelihood dominates, and the latter may be related to an asymptotic bias. BLR-Prior demonstrates generally worse performance with $\alpha = 3$. The supplementary material includes results for additional choices of α, which are omitted from fig. 1 for visibility. We find that a larger α improves the performance on regression, but at a significant cost for classification performance.

Importantly, *none of the baselines consistently match the performance of our method*, across both problems and all choices of n_*. Moreover, they involve hyperparameters that are difficult to determine *a priori*, in contrast to our method where the hyperparameter ρ has a clear interpretation. Also note that only our method has the appealing property of never underperforming the ERM baseline.

6.2 BENCHMARK DATASETS

We now turn to two datasets adapted from the DomainBed benchmark [Gulrajani and Lopez-Paz, 2020]: Colored MNIST [Arjovsky et al., 2019] and PACS [Li et al., 2017].

Background and setup The PACS dataset consists of 9991 images from 4 domains and 7 categories, with images from different domains having different stylistic features. The Colored MNIST dataset is defined as follows: let $(\bar{x}_{e,i}, \bar{y}_{e,i}) \in \mathbb{R}^{784} \times \{0, \ldots, 9\}$ be a MNIST sample, and sample $y^e_i := \mathbf{1}\{\bar{y}_{e,i} < 5\} \oplus \mathrm{Bern}(\alpha), c^e_i := y^e_i \oplus \mathrm{Bern}(\beta_e)$, where $\beta_e \in \{0.1, 0.2, 0.9\}$ depends on the environment, and $\oplus$ denotes the XOR operation. The input x^e_i is obtained by coloring $\bar{x}_{e,i}$ to green or red based on c_i^e. The original dataset Arjovsky et al. [2019] uses $\alpha := 0.25$, implying an accuracy of 75% for the invariant predictor. As we are interested in scenarios where the invariant predictor is more performant, *we use a modified value of* $\alpha := 0.1$. We note the different natures of the two datasets: by construction, on Colored MNIST there is an unavoidable trade-off between the in-distribution and test performance, whereas on PACS there may exist an invariant predictor with near-perfect accuracy, even though its recovery can still be hindered by the inductive bias of a neural network model.

We compare the proposed method with `BLR` and `BLR-LC` baselines in the preceding section, as well as the `DivDis` method by Lee et al. [2022]. `DivDis` builds a finite collection of candidate predictors based on a diversity criterion, and selects the predictor with the best performance on the adaptation samples. *All adaptation algorithms are applied to the last linear layer of a ConvNet model*, which is initialized at the ERM. Note this is not a limitation of our algorithm (or the baselines), but is adopted for simplicity; still, this strategy is also advocated by recent works such as Kirichenko et al. [2022], and the `BLR` baseline recovers the procedure in their work.

Contrary to many applications, on these datasets it is unclear whether there exists a near-perfect invariant predictor within our hypothesis space. In such scenarios, it is not necessarily reasonable to assume good prior knowledge of a performance lower bound for the (best) invariant predictor. Therefore, to ensure a realistic setup, we determine the lower bound hyperparameter in our method *based on the ERM*, using a possibly misspecified choice of $\rho + \epsilon_n := \max_{e \in \mathcal{E}_{tr}} \hat{R}_e(\hat{\theta}_{ERM}) + \delta$. We report the results for $\delta = 0.1$ in the main text and defer the results for alternative choices to the appendix.

We use the ConvNet architecture in Gulrajani and Lopez-Paz [2020] and follow the training protocol therein. The number of learnable parameters is thus $(1024 + 1) \times 2$ for Colored MNIST, and $(2048 + 1) \times 7$ for PACS. For the `DivDis` baseline, we implement the method on the same adaptation samples, using 50 predictor heads; we vary its hyperparameters $(\lambda_{mi}, \lambda_{reg}) \in \{0.1, 1.0, 10\}$ and report the configuration with the best *test* accuracy. The rest of the setup largely follows the preceding section and are deferred to the appendix.

Table 1: Average accuracy and a lower estimate of performance on the modified Colored MNIST dataset; the latter is defined as the 20th percentile of accuracy across 20 replications, for the worst train/test environment split. `CBLR` denotes the proposed method.

n_*	0 (ERM)	4	8	16	32
BLR	82.4 / 71.5	82.5 / 75.2	85.7 / 80.9	87.4 / 85.4	88.3 / 86.3
BLR-LC-N_{adapt}		85.6 / 80.5	86.3 / 82.9	86.9 / 83.9	87.2 / 84.8
BLR-LC-N_{train}		81.7 / 67.9	81.9 / 68.3	82.4 / 70.1	83.1 / 72.7
DivDis		82.3 / 70.7	82.3 / 70.6	82.3 / 70.7	82.3 / 70.7
CBLR		85.7 / 81.5	87.0 / 85.0	87.6 / 86.8	88.1 / 86.7

Table 2: Average accuracy and a lower estimate of performance on PACS. The latter is defined as in Table 1.

n_*	0 (ERM)	16	32	64	256
BLR	83.2 / 72.6	83.8 / 70.1	86.8 / 76.4	88.3 / 79.4	89.4 / 80.6
BLR-LC-N_{adapt}		86.8 / 77.8	86.6 / 76.4	87.2 / 77.6	87.1 / 77.2
BLR-LC-N_{train}		85.0 / 76.1	85.1 / 75.9	85.2 / 75.5	85.6 / 76.9
DivDis		85.0 / 77.6	84.8 / 77.1	84.8 / 76.8	85.0 / 76.9
CBLR		86.4 / 77.6	87.4 / 78.6	88.4 / 79.9	90.3 / 83.7

Results and discussion For space reasons, we only report aggregated results in the text, deferring full results to the supplementary material. Table 1-2 present the average accuracy across environments, as well as a pessimistic performance estimate that provides intuition on unfavorable scenarios; the latter can be particularly important for domain generalization applications [Eastwood et al., 2022]. As we can see, our method demonstrates excellent performance across all settings. In contrast, `BLR` has unstable performance at small sample sizes, and `BLR-LC` becomes less competitive as sample size increases, which is consistent with the synthetic experiments. `DivDis` is generally less competitive on the modified Colored MNIST dataset, and on PACS at larger sample sizes, indicating insufficient coverage of its candidate solution set. A possible reason is that `DivDis` does not account for the performance gap between the invariant and non-invariant predictors, which is notably larger on Colored MNIST. However, its performance may also be improved if a larger number of unlabeled test samples are available and can be selectively labeled, as is done in the experiments of Lee et al. [2022].

On both datasets, there is a rapid improvement over the ERM baseline after a handful of adaptation samples, which is important because *ERM is a strong baseline* on these benchmarks, having outperformed all algorithms tested in Gulrajani and Lopez-Paz [2020]. We note the slower improvement of worst-case performance for PACS is because one domain exhibits significant label shift, which is generally at odds with the assumption that invariant predictor exists [Arjovsky et al., 2019]. It is in principle possible to adapt our method to label shift scenarios, by redefining the constraint set to use a reweighted accuracy, but we will not explore this for simplicity.

6.3 REAL-WORLD EXPERIMENT

Finally, we illustrate our method on a real-world application of out-of-distribution prediction.

Background and setup The task concerns the classification of acoustic array data, which are spatio-temporal signals that can be viewed as images. The input consists of certain "primary signals" that induce approximately invariant conditionals, superimposed with environment-specific responses. The latter induce spurious correlations and are consistently picked up by ConvNet models. There are 4 classes; we have data from 4 environments, each with $n_e \sim 10^5$ samples. Domain knowledge suggests that on class-balanced data, an invariant classifier should have an error rate lower than 5%.

Past experiments suggest that the training data do not contain sufficient information to guarantee OOD generalization: in leave-one-domain-out evaluation, a generalization gap of up to 20% shows up, and off-the-shelf algorithms including IRM, GDRO and domain-adversarial training all fail to improve over ERM. Thus, test-time adaptation appears necessary.

The experiment setup largely follows the last subsection: we perform adaptation on the last linear layer, and conduct leave-one-domain-out evaluation with repeatedly sampled adaptation sets. We set $\rho = 0.05$. Due to the large training sample size, we subsample 10^3 samples from each environment in defining our constraint set, and set ε_n accordingly. (We find that the result is generally insensitive to $\rho + \epsilon_n \in [0.03, 0.1]$.) We compare with BLR and BLR-LC. For the latter, we experiment with scaling the training loss using a factor of $N \in \{1, 2, 4, 8\} \times 100$, and provide an optimistic estimate for its performance by setting N based on test performance.

Table 3: Results for Section 6.3. For each train/test domain split, we report the mean and standard deviation of test accuracy across 20 trials. LB denotes an estimate of performance in unfavorable scenarios, defined as in Table 1.

e_*	n_*	(ERM)	20	80	320
1	BLR	81.9	83.5 ±1.3	91.3 ±0.7	94.1 ±0.3
	BLR-LC		86.9 ±1.7	91.3 ±0.9	93.6 ±0.3
	CBLR		86.7 ±1.4	91.7 ±0.5	93.9 ±0.2
2	BLR	83.1	85.6 ±2.5	91.1 ±1.3	93.3 ±0.1
	BLR-LC		89.2 ±0.2	92.4 ±0.6	93.8 ±0.1
	CBLR		89.2 ±0.1	92.1 ±0.1	93.9 ±0.1
3	BLR	92.6	91.9 ±1.3	92.8 ±1.1	95.3 ±0.1
	BLR-LC		93.2 ±0.1	94.8 ±0.1	95.5 ±0.1
	CBLR		94.0 ±0.1	95.0 ±0.1	95.7 ±0.1
4	BLR	82.5	88.9 ±1.8	92.8 ±1.1	96.0 ±0.3
	BLR-LC		87.8 ±1.1	92.6 ±0.8	96.3 ±0.5
	CBLR		86.2 ±1.3	92.4 ±1.3	96.3 ±0.2
LB	BLR	81.9	82.4	90.0	93.2
	BLR-LC		85.4	90.5	93.4
	CBLR		85.1	91.3	93.7

Results and discussion The results are shown in Table 3. As we can see, test-time adaptation delivers significant improvements over the ERM baseline, which has not been possible in the past experiments with domain generalization algorithms. Our method and BLR-LC has similar performance, whereas BLR has less stable performance at smaller sample sizes. Our method can be preferable, because its hyperparameter can be easily determined using domain knowledge in a principled way.

7 CONCLUSION

In this work we study the problem of adaptation to distribution shift, given a small collection of training environments and a handful of test samples. We reveal a pathological behavior of the standard Bayesian posterior and address it with a constrained Bayesian formulation. We prove that constrained modeling may lead to sample efficiency gains in certain settings, and demonstrate the robust performance of our method on synthetic, benchmark and real-world tasks.

Our work addresses OOD prediction in the underidentified regime, which can be inherently challenging. It is thus necessary to introduce additional information or assumptions. We note our core assumptions: the existence of an invariant predictor, some knowledge about its performance, and access to adaptation samples. While these assumptions are satisfied in many problems and can be relaxed to various extents, there are inevitably scenarios where alternative assumptions are more appropriate. It would be interesting future work to study adaptation and uncertainty modeling in such settings.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2020AAA0106302), NSFC Projects (Nos. 62061136001, 62076145, 62076147, U19B2034, U1811461, U19A2081, 61972224), Tsinghua Institute for Guo Qiang, and the High Performance Computing Center, Tsinghua University. J.Z is also supported by the New Cornerstone Science Foundation through the XPLORER PRIZE.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Olivier Bachem, Mario Lucic, and Andreas Krause. Practical coreset constructions for machine learning. *arXiv preprint arXiv:1703.06476*, 2017.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016. ISSN 1467-9868. doi: 10.1111/rssb.12158. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12158. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12158.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011.

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. *Discrete & Computational Geometry*, 59:757–783, 2018.

Peter Bühlmann. Invariance, Causality and Robustness, December 2018. URL http://arxiv.org/abs/1812.08233. arXiv:1812.08233 [stat].

Yining Chen, Elan Rosenfeld, Mark Sellke, Tengyu Ma, and Andrej Risteski. Iterative feature matching: Toward provable domain generalization with logarithmic environments. *Advances in Neural Information Processing Systems*, 35:1725–1736, 2022.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, May 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00338-7. URL https://www.nature.com/articles/s42256-021-00338-7.

Cian Eastwood, Alexander Robey, Shashank Singh, Julius Von Kügelgen, Hamed Hassani, George J Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. *arXiv preprint arXiv:2207.09944*, 2022.

Ishaan Gulrajani and David Lopez-Paz. In Search of Lost Domain Generalization. *arXiv:2007.01434 [cs, stat]*, July 2020. URL http://arxiv.org/abs/2007.01434. arXiv: 2007.01434.

Junguang Jiang, Yang Shu, Jianmin Wang, and Mingsheng Long. Transferability in Deep Learning: A Survey, January 2022. URL http://arxiv.org/abs/2201.05867. arXiv:2201.05867 [cs] version: 1.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations, April 2022. URL http://arxiv.org/abs/2204.02937. arXiv:2204.02937 [cs, stat].

Yoonho Lee, Huaxiu Yao, and Chelsea Finn. Diversify and Disambiguate: Learning From Underspecified Data, June 2022. URL http://arxiv.org/abs/2202.03418. arXiv:2202.03418 [cs, stat].

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

Yong Lin, Hanze Dong, Hao Wang, and Tong Zhang. Bayesian invariant risk minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16021–16030, 2022.

Pierre-Louis Lions and Alain-Sol Sznitman. Stochastic differential equations with reflecting boundary conditions. *Communications on pure and applied Mathematics*, 37 (4):511–537, 1984.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

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.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant Models for Causal Transfer Learning, September 2018. URL http://arxiv.org/abs/1507.05333. arXiv:1507.05333 [stat].

Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The Risks of Invariant Risk Minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BbNIbVPJ-42.

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:1911.08731 [cs, stat]*, April 2020a. URL http://arxiv.org/abs/1911.08731. arXiv: 1911.08731.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An Investigation of Why Overparameterization Exacerbates Spurious Correlations. *arXiv:2005.04345 [cs, stat]*, August 2020b. URL http://arxiv.org/abs/2005.04345. arXiv: 2005.04345.

Kanji Sato, Akiko Takeda, Reiichiro Kawai, and Taiji Suzuki. Convergence error analysis of reflected gradient langevin dynamics for globally optimizing non-convex constrained problems. *arXiv preprint arXiv:2203.10215*, 2022.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

Yoav Wald, Gal Yona, Uri Shalit, and Yair Carmon. Malign overfitting: Interpolation and invariance are fundamentally at odds. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=dQNL7Zsta3.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. Generalizing to Unseen Domains: A Survey on Domain Generalization, May 2022. URL http://arxiv.org/abs/2103.03097. arXiv:2103.03097 [cs].

Garrett Wilson and Diane J. Cook. A Survey of Unsupervised Deep Domain Adaptation. *ACM Transactions on Intelligent Systems and Technology*, 11(5):1–46, October 2020. ISSN 2157-6904, 2157-6912. doi: 10.1145/3400066. URL https://dl.acm.org/doi/10.1145/3400066.

Haotian Ye, James Zou, and Linjun Zhang. Freeze then Train: Towards Provable Representation Learning under Spurious Correlations and Feature Noise, October 2022. URL http://arxiv.org/abs/2210.11075. arXiv:2210.11075 [cs].

Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34: 23664–23678, 2021.

Ruqi Zhang, Qiang Liu, and Xin T Tong. Sampling in constrained domains with orthogonal-space variational gradient descent. *arXiv preprint arXiv:2210.06447*, 2022.

Tong Zhang. From ε-entropy to kl-entropy: Analysis of minimum information complexity density estimation. *Annals of statistics*, 34(5):2180–2210, 2006.

Jun Zhu, Ning Chen, and Eric Xing. Bayesian inference with posterior regularization and applications to infinite latent svms. *Journal of Machine Learning Research*, 15: 1799–1847, 2014.