
Mind the GAP: Improving Robustness to Subpopulation Shifts with Group-Aware Priors

Tim G. J. Rudner
New York University

Ya Shi Zhang
New York University

Andrew Gordon Wilson
New York University

Julia Kempe
New York University
Meta FAIR

Abstract

Machine learning models often perform poorly under subpopulation shifts in the data distribution. Developing methods that allow machine learning models to better generalize to such shifts is crucial for safe deployment in real-world settings. In this paper, we develop a family of group-aware prior (GAP) distributions over neural network parameters that explicitly favor models that generalize well under subpopulation shifts. We design a simple group-aware prior that only requires access to a small set of data with group information and demonstrate that training with this prior yields state-of-the-art performance—even when only retraining the final layer of a previously trained non-robust model. Group aware-priors are conceptually simple, complementary to existing approaches, such as attribute pseudo labeling and data reweighting, and open up promising new avenues for harnessing Bayesian inference to enable robustness to subpopulation shifts.

1 INTRODUCTION

Distribution shifts, frequently occurring in real-world data, have long plagued machine learning models [Quiñonero Candela, 2009]. Empirical risk minimization [ERM; Vapnik, 1998]—the minimization of average training loss—is known to generalize poorly under distribution shifts. In particular, *subpopulation shifts*—due to attribute and class biases—can cause significantly increased test error on certain population groups, even if the average test error remains low [Hashimoto et al., 2018]. In many applications, high

accuracy on certain subpopulations/groups is essential, and failure to generalize under subpopulation shifts can have severe and harmful consequences [Barocas and Selbst, 2016, Dastin, 2018].

In this paper, we focus on achieving *group robustness*, increasing the worst-case test performance across groups represented in the data, which is crucial for building equitable and effective machine learning systems. A variety of approaches exist to tackle this problem, including methods that explicitly optimize for worst-case group performance [Sagawa et al., 2020], approaches that identify neural network features that lead to improved group robustness [Kirichenko et al., 2023], and several techniques that rely on training helper models used to identify shifts and reweighting the data [e.g., Liu et al., 2021, Nam et al., 2022].

Unlike previous work, we approach group robustness from a Bayesian perspective and present a general approach to designing data-driven priors that favor models with high group robustness. Under such priors, performing Bayesian inference will lead to posterior distributions over neural network parameters that allow the model to fit the training data while also respecting the soft constraints imposed by the prior distribution.

To demonstrate the usefulness of such priors, we construct an example of a simple data-driven *group-aware prior* (GAP) distribution over the parameters of a neural network designed to place high probability density on parameter values that induce predictive models that generalize well under subpopulation shifts. While exact Bayesian inference with non-standard priors can be challenging, we build on the approach presented in Rudner et al. [2023] to find the most likely parameters under the posterior distribution implied by the prior and the data. We illustrate the process of training a model with a group-aware prior in Figure 1.

Data-driven group-aware prior distributions allow for probabilistically principled learning, are modular in their ability to be applied to any likelihood function, and are simple to implement. In our empirical eval-

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

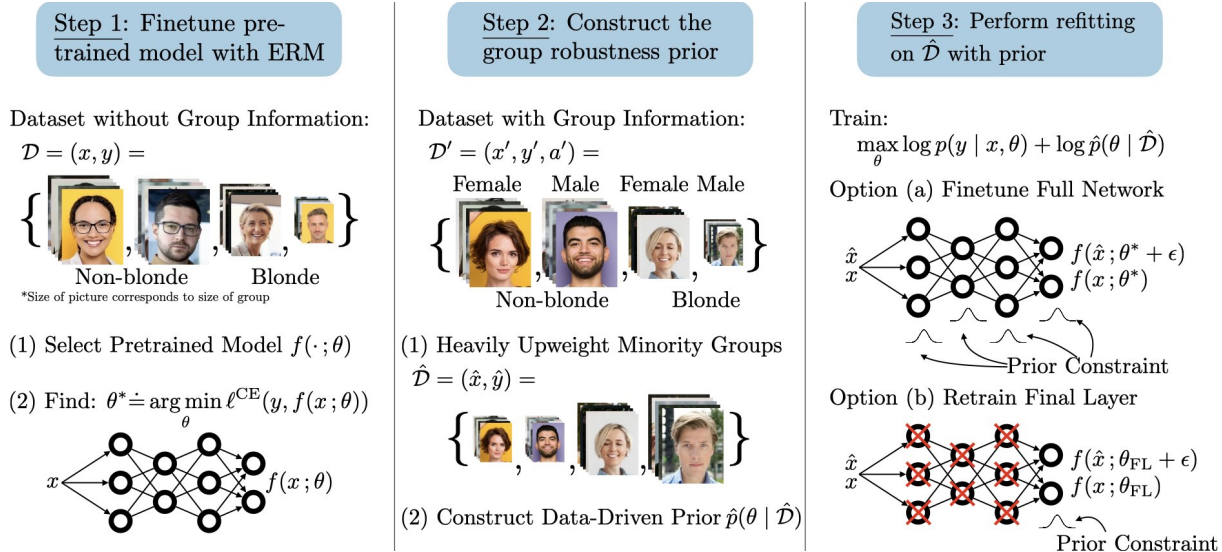


Figure 1: **Training with a Group-Aware Prior (GAP)**. Step 1: Train a neural network with empirical risk minimization (ERM). Step 2: Construct a group-aware prior by defining a tractable joint distribution that places high probability density on parameters that achieve high worst group accuracy. Step 3: Find the most likely parameters under the posterior induced by the group-aware prior and the data.²

uation, we consider the realistic setting where only a small set of data with group information is available and construct a simple example of a group-aware prior. We then show that for established subpopulation shift benchmarking tasks (i) finetuning a previously trained model with a group-aware prior *leads to state-of-the-art results on all benchmarks* and (ii) only retraining the final layer of a previously trained model with a group-aware prior leads to state-of-the-art results on two of the benchmarks, and remains competitive on the other.¹ Excitingly, this probabilistic formulation of group robustness opens up new routes for bringing to bear the vast arsenal of Bayesian inference methods to obtain even higher levels of group robustness.

To summarize, our key contributions are as follows:

1. We present a general framework for constructing tractable data-driven priors to achieve group robustness under subpopulation shifts.
2. We design a simple group-aware prior (GAP) that places high probability density on parameter values that lead to high group robustness.
3. We show empirically that finetuning a previously trained model with this prior leads to state-of-the-art results on standard benchmarking tasks.
4. We consider a more constrained setting in which we freeze a previously trained network and only retrain the final layer. We find that even in this highly constrained setting, retraining only a few hundred parameters with group-aware priors leads to state-of-the-art results.

¹Our code is available at <https://github.com/timrudner/group-aware-priors>.

2 BACKGROUND

2.1 Learning as Probabilistic Inference

Consider supervised learning problems with N i.i.d. data realizations $\mathcal{D} = \{x_{\mathcal{D}}^{(n)}, y_{\mathcal{D}}^{(n)}\}_{n=1}^N = (x_{\mathcal{D}}, y_{\mathcal{D}})$ of inputs $X \in \mathcal{X}$ and labels $Y \in \mathcal{Y}$ with input space \mathcal{X} and label space \mathcal{Y} . For supervised learning tasks, we define a parametric observation model $p_{Y|X, \Theta}(y | x, \theta; f)$ with a neural network mapping, $f(\cdot; \theta)$, and a *prior* distribution over the parameters, $p_{\Theta}(\theta)$ with the goal of inferring a *posterior distribution* from the data.

Since, by Bayes' Theorem, the posterior under this model is proportional to the joint probability density given by the product of the likelihood of the parameters under the data $p_{Y|X, \Theta}(y_{\mathcal{D}} | x_{\mathcal{D}}, \theta)$ and the prior $p_{\Theta}(\theta)$,

$$p_{\Theta|Y, X}(\theta | y_{\mathcal{D}}, x_{\mathcal{D}}) \propto p_{Y|X, \Theta}(y_{\mathcal{D}} | x_{\mathcal{D}}, \theta) p_{\Theta}(\theta),$$

the most likely parameters under the posterior are given by the mode of $p_{Y|X, \Theta}(y_{\mathcal{D}} | x_{\mathcal{D}}, \theta) p_{\Theta}(\theta)$. Maximum a posteriori (MAP) estimation seeks to find this mode, θ^{MAP} [Bishop, 2006, Murphy, 2013]. Under a likelihood that factorizes across the data points given parameters θ , the MAP optimization objective can be expressed as

$$\mathcal{F}^{\text{MAP}}(\theta) = \sum_{n=1}^N \log p_{Y|X, \Theta}(y_{\mathcal{D}}^{(n)} | x_{\mathcal{D}}^{(n)}, \theta) + \log p_{\Theta}(\theta).$$

Under Gaussian and categorical likelihood functions, the log-likelihood term in the MAP optimization objective corresponds to a scaled negative mean squared error (MSE) loss function and a negative cross-entropy

²All images in this paper are taken from <https://unsplash.com> or <https://www.istockphoto.com> under unlimited, perpetual, nonexclusive, worldwide license.

loss function, respectively. Similarly, under Gaussian and Laplace priors, the log-density of the prior is proportional to L_2 regularization and L_1 regularization, respectively. We will use this probabilistic perspective to obtain a tractable optimization objective that allows incorporating data-driven priors designed to improve the group robustness of neural networks into training.

2.2 Subpopulation Shifts

We consider classification problems on data $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where we assume that the data consists of several groups (subpopulations) $g \in \mathcal{G}$, which are often defined by a combination of a label $y \in \mathcal{Y}$ and spurious attribute $a \in \mathcal{A}$ (sometimes called *environment*). The attribute $a \in \mathcal{A}$ may or may not be available during training. For instance, in CelebA hair color prediction [Liu et al., 2015], the labels are ‘blond’ and ‘brown’, and the groups are ‘non-blond women’ (g_1), ‘blond women’ (g_2), ‘non-blond men’ (g_3), and ‘blond men’ (g_4) with proportions 44%, 14%, 41%, and 1% of the data, respectively; the group g_4 is the minority group, and gender serves as the attribute (spurious feature).

We assume that the training data comes from a mixture of group-wise distributions $p_{\text{train}} = \sum_{g \in \mathcal{G}} \alpha_g p_g$, where $\alpha \in \Delta_{|\mathcal{G}|}$ (group weights summing to 1) and p_g are distributions over elements of the group $g \in \mathcal{G}$. We speak of *subpopulation shift* when the test data comes from a differently weighted distribution $p_{\text{test}} = \sum_{g \in \mathcal{G}} \beta_g p_g$, where the weighting $\beta \in \Delta_{|\mathcal{G}|}$ is not known at train-time. This leads to the natural goal of maximizing *worst group test accuracy* (WGA), that is, the lowest accuracy across groups \mathcal{G} represented in the test data. We follow the shift taxonomy proposed in Yang et al. [2023], and our benchmarking datasets exhibit combinations of all of the following three shifts.

Spurious correlations are present when a label y is correlated with an attribute a in the training distribution but not in the test distribution. For instance, it was shown that thoracic X-ray images contain spurious correlations between labels and attributes such as patient age, scanning position, or text font, which may not be present in the test data [Heaven, 2021, DeGrave et al., 2021].

An *attribute shift* is present when the distribution of an attribute a for a specific label y and context combination differs between the training and test distributions. For instance, MultiNLI has a high prevalence of negation words (‘no’, ‘never’)—one of the attributes.

A *Class shift* is present when the distribution of classes differs between the training and test distributions. For example, in the CelebA dataset, the ‘blond’ class in the training set constitutes 15% of the training examples but a larger fraction in the test set.

3 RELATED WORK

Subpopulation shifts are omnipresent in real-world applications, which has led to a large body of literature concerned with reducing the adverse effects of spurious correlations, attribute shifts, and class shifts on model performance through more robust models.

Ensuring robustness to subpopulation shifts is a long-standing challenge in machine learning. A broad range of methods has been developed to mitigate different types of subpopulation shifts, including *data reweighting* [Idrissi et al., 2022], *data augmentation* [Zhang et al., 2018], *domain-invariant feature learning* [Arjovsky et al., 2020, Li et al., 2018], and *(sub-)group robustness* methods [Sagawa et al., 2020, Liu et al., 2021, Nam et al., 2022, Zhang et al., 2022, Kirichenko et al., 2023]. Several methods designed to achieve high worst group accuracy build on the distributionally robust optimization (DRO) framework [Rahimian and Mehrotra, 2022], where worst-case accuracy—instead of average case accuracy—is explicitly maximized during training [Ben-Tal et al., 2013, Hu et al., 2018, Oren et al., 2019, Zhang et al., 2020]. Notably, GroupDRO [Sagawa et al., 2020] has become a standard baseline for group robustness.

Group Labeling Models. In most real-world settings, obtaining group information is expensive and only feasible for a small number of data points. To emulate real-world settings where only a small number of training data with group information is available, existing methods have used the validation sets of benchmarking datasets to achieve high worst group accuracy. Nam et al. [2022] and Sohoni et al. [2022] train group labeling models on group-labeled validation data, generate group labels for the training dataset, and then train a second model on the full dataset using the generated group labels. In contrast, our proposed method does not require training a group labeling model and instead only requires reweighting the validation set during group-robustness finetuning to obtain state-of-the-art performance, outperforming all group labeling techniques.

Last-Layer Retraining. A particularly simple and efficient approach to improving group robustness in settings with limited group label availability is *deep feature reweighting* [DFR; Kirichenko et al., 2023], which first finetunes a neural network using expected risk minimization (ERM) on a training dataset without group information and then retrains the last layer on a group-balanced reweighting dataset using ERM regularized with an L_1 -norm between the current and the previously fine-tuned last-layer parameters (see also Izmailov et al. [2022], Qiu et al. [2023], and Le et al. [2023] for follow-up works). This work is notable for

its marked simplicity and low complexity compared to related methods, as it only requires last-layer retraining on a validation set to reach state-of-the-art performance on some benchmarks and competitive performance on other benchmarks. We present results for our method in the setting where we only retrain the final layer with a validation set and show that it outperforms DFR on a majority of benchmarking tasks. While Kirichenko et al. [2023] showed that DFR leads to worse performance when group-robustness finetuning on the full network, we find that group-robustness finetuning the full network with our method leads to significant improvements over last-layer retraining, resulting in state-of-the-art results.

Data-driven Priors. Prior distributions encode prior information about random variables, but designing informative prior distributions over neural network parameters is challenging in practice due to the limited interpretability of neural networks. To address this challenge, previous work has proposed data-driven priors by pertaining neural networks and using the trained parameters to specify priors for related downstream tasks. For example, Shwartz-Ziv et al. [2022] propose to reshape the loss surface using data-driven priors. These informative priors are learned from the source task—similar to the pre-training paradigm—and lead to improved transfer generalization. Taking an alternative route, Rudner et al. [2023] propose *function-space empirical Bayes*, deriving an optimization objective that allows approximating the most likely parameters under posteriors (i.e., the maximum a posteriori estimate) computed from sophisticated data-driven priors and use it to learn models with improved uncertainty quantification. We use the optimization objective proposed in Rudner et al. [2023] to perform maximum a posteriori estimation with group-aware priors.

Generalization and Parameter Perturbations. Prior work on improving generalization of neural networks, notably *Sharpness-Aware Minimization* [SAM; Foret et al., 2021] and *Stochastic Weight-Averaging* [SWA; Izmailov et al., 2018], has attempted to improve model performance by finding flatter minima of the loss landscape. The underlying idea is that minimizing a loss under different types of perturbations to the neural network parameters will steer the learned neural network parameters into regions of parameter space where small perturbations to the parameters do not lead to significant increases in the training loss and ultimately lead to learned parameters that correspond to flat loss minima, which in turn have been related to improved generalization.

4 GROUP-AWARE PRIORS

In this section, we first define a general family of group-aware priors. Then, we use this framework to develop a simple and scalable instantiation of a group-aware prior. Finally, we describe how to apply the maximum a posteriori estimation procedure proposed in Rudner et al. [2023] to use this prior to train a neural network.

4.1 A Family of Group-Aware Priors

We begin by specifying the auxiliary inference problem. Let \hat{x} be a set of *context points* $\hat{x} = \{\hat{x}_1, \dots, \hat{x}_M\}$ with corresponding *context labels* $\hat{y} = \{\hat{y}_1, \dots, \hat{y}_M\}$, and let \hat{Z} be a Bernoulli random variable denoting whether a given set of neural network parameters induces predictions with some desired property (e.g., high uncertainty on certain evaluation points, high accuracy on evaluation points with a certain group attribute, etc.). We define an auxiliary likelihood function $\hat{p}_{\hat{Z}|\Theta}(\hat{z}|\theta; f, p_{\hat{X}, \hat{Y}})$ —which denotes the likelihood of observing a yet-to-be-specified outcome \hat{z} under $\hat{p}_{\hat{Z}|\Theta}$ given θ and $p_{\hat{X}, \hat{Y}}$ —and a prior over the model parameters, $p_{\Theta}(\theta)$. For notational simplicity, we will drop the subscripts going forward except when needed for clarity. By Bayes’ Theorem, the posterior under this model and observation \hat{z} is given by

$$\hat{p}(\theta|\hat{z}; f, p_{\hat{X}, \hat{Y}}) = \frac{\hat{p}(\hat{z}|\theta; f, p_{\hat{X}, \hat{Y}})p(\theta)}{\hat{p}(\hat{z}|\theta; f, p_{\hat{X}, \hat{Y}})}. \quad (1)$$

To define a family of data-driven priors that place high probability density on neural network parameter values that induce predictive functions that achieve high group robustness, we define a specific Bernoulli auxiliary observation model $\hat{p}_{\hat{Z}|\Theta}$ in which $\hat{Z} = 1$ denotes the outcome of ‘achieving group robustness’ and $\hat{p}(\hat{z} = 1|\theta; f, p_{\hat{X}, \hat{Y}})$ denotes the likelihood of $\hat{z} = 1$ given θ and $p_{\hat{X}, \hat{Y}}$. We can now define a general family of group-aware priors by specifying the Bernoulli observation model

$$\begin{aligned} \hat{p}(\hat{z} = 1|\theta; f, p_{\hat{X}, \hat{Y}}) &= \exp(-\lambda \mathbb{E}_{p_{\hat{X}, \hat{Y}}}[c(\hat{X}, \hat{Y}, \theta)]) \\ \hat{p}(\hat{z} = 0|\theta; f, p_{\hat{X}, \hat{Y}}) &= 1 - \hat{p}(\hat{z} = 1|\theta; f, p_{\hat{X}, \hat{Y}}), \end{aligned} \quad (2)$$

where $c: \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^P \rightarrow \mathbb{R}_{\geq}$ is a ‘cost’ function and $\lambda > 0$ is a scaling parameter. By specifying an auxiliary dataset with $\hat{\mathcal{D}} = \hat{z} = \{1, \dots, 1\}$ and a distribution $p_{\hat{X}, \hat{Y}}$ we obtain a posterior $\hat{p}(\theta|\hat{z}; f, p_{\hat{X}, \hat{Y}})$, the distribution over neural network parameters that we *would* infer if we observed outcomes $\hat{z} = \{1, \dots, 1\}$ under the likelihood defined above. As with every Bayesian method, the quality of this posterior is determined by the quality of the observation model $\hat{p}_{\hat{Z}|\Theta}$, the data, and the prior. Therefore, if the observation model is poor, so will be the posterior. As a result, the main challenge in designing useful group-aware priors is to construct

an observation model—that is, a cost function c —that is as well-specified as possible. The better specified the observation model, the more useful the data-driven prior will be. Below, we present a specific instantiation of a group-aware prior by proposing a simple cost function, paired with a suitable set of context points and context labels.

4.2 A Simple Group-Aware Prior

To specify a practically useful group-aware prior, we need two ingredients: (i) We need to specify the distribution $p_{\hat{X}, \hat{Y}}$ for which observing $\hat{z} = \{1, \dots, 1\}$ would be most informative, and (ii) we need to specify a cost function c which—for suitably chosen $p_{\hat{X}, \hat{Y}}$ —is a good proxy for group robustness on unknown test points. Both (i) and (ii) are generally challenging and could be tackled with sophisticated methods—for example, by learning tailored generative models or handcrafting complex cost functions. However, to demonstrate the usefulness of group-aware priors, we instead limit ourselves to fixed, prespecified distribution $p_{\hat{X}, \hat{Y}}$ and a very simple cost function.

First, to specify a useful context distribution $p_{\hat{X}, \hat{Y}}$, we assume that we have access to at least a small dataset, \mathcal{D}_{val} , with group information and simply upsample the dataset to create a distribution

$$p_{\hat{X}, \hat{Y}} = \sum_{g \in \mathcal{G}} \alpha_g p_g \quad (3)$$

with $\alpha_g = \tilde{\alpha}_g / \sum_{g \in \mathcal{G}} \tilde{\alpha}_g$ and $\tilde{\alpha}_g = (|\mathcal{D}_{\text{val}}| / |\mathcal{D}_{\text{val}}^g|)^\gamma$, where $|\mathcal{D}_{\text{val}}|$ is the number of data points in \mathcal{D}_{val} , $|\mathcal{D}_{\text{val}}^g|$ is the number of data points from group g in \mathcal{D}_{val} , and $\gamma \geq 1$ is a scaling parameter. The larger the hyperparameter γ , the stronger rare groups will be upweighted. The smaller $|\mathcal{D}_{\text{val}}^g|$ as a fraction of $|\mathcal{D}_{\text{val}}|$, the larger α_g will be.

Second, to specify a useful cost function, we define

$$c(\hat{x}, \hat{y}, \theta) \doteq \ell(\hat{y}, f(\hat{x}; \theta + \rho \epsilon(\theta))) \quad (4)$$

where $\rho > 0$ is a scaling parameter and ℓ is a cross-entropy loss function for classification tasks and a mean-squared error loss function for regression tasks. $\epsilon(\theta)$ is a worst-case perturbation to the model parameters proposed in Foret et al. [2021] and given by

$$\epsilon(\theta, \hat{x}, \hat{y}) \doteq \perp \frac{\nabla_{\theta} \ell(\hat{y}, f(\hat{x}; \theta))}{\|\nabla_{\theta} \ell(\hat{y}, f(\hat{x}; \theta))\|_2}, \quad (5)$$

where \perp is the `stop_gradient` operator. The intuition for this cost function is simple: we know that achieving low loss tends to correspond to good generalization, but given that the context dataset has—by assumption—few data points from rare groups, generalizing to test points from these groups is difficult. To design a cost

function that leads to a data-driven prior with high probability density on parameters that enable good generalization to points from rare groups, we add a worst-case perturbation defined to lead to a maximum increase in the loss, as proposed in Foret et al. [2021].

Unfortunately, the full posterior $\hat{p}(\theta | \hat{z}; f, p_{\hat{X}, \hat{Y}})$ is intractable, since the parameters θ appear non-linearly in c . However, taking the log of the analytically tractable joint density $\hat{p}(\hat{z} | \theta; f, p_{\hat{X}, \hat{Y}}) p(\theta)$, we obtain

$$\begin{aligned} & \log \hat{p}(\hat{z} | \theta; f, p_{\hat{X}, \hat{Y}}) + \log p(\theta) \\ & \propto -\mathbb{E}_{p_{\hat{X}, \hat{Y}}} [\ell(\hat{Y}, f(\hat{X}; \theta + \rho \epsilon(\theta)))] + \log p(\theta) \quad (6) \\ & \doteq \mathcal{J}(\theta, p_{\hat{X}, \hat{Y}}), \end{aligned}$$

with proportionality up to an additive constant independent of θ . This implies that

$$\arg \max_{\theta} \hat{p}(\theta | \hat{z}; f, p_{\hat{X}, \hat{Y}}) = \arg \max_{\theta} \mathcal{J}(\theta, p_{\hat{X}, \hat{Y}}),$$

that is, maximizing the analytically tractable expression $\mathcal{J}(\theta, p_{\hat{X}, \hat{Y}})$ with respect to θ is mathematically equivalent to maximizing the posterior $\hat{p}(\theta | \hat{z}; f, p_{\hat{X}, \hat{Y}})$. Next, we will show how to use this insight to train neural networks with this prior.

4.3 Maximum A Posteriori Estimation

To train a neural network with data-driven group-aware priors, we follow the approach described in Rudner et al. [2023] and derive a tractable objective function for neural network training with group-aware priors that allows us to find the most likely neural network parameters under the posterior distribution,

$$p(\theta | y_{\mathcal{D}}, x_{\mathcal{D}}, \hat{z}; f, p_{\hat{X}, \hat{Y}}) = \frac{p(y_{\mathcal{D}} | x_{\mathcal{D}}, \theta) \hat{p}(\theta | \hat{z}; f, p_{\hat{X}, \hat{Y}})}{p(y_{\mathcal{D}} | x_{\mathcal{D}}, \hat{z}; f, p_{\hat{X}, \hat{Y}})}. \quad (7)$$

To do so, we use two insights: (i) that maximum a posteriori estimation only requires an analytically tractable function proportional to the log-joint distribution $\log(p(y_{\mathcal{D}} | x_{\mathcal{D}}, \theta) \hat{p}(\theta | \hat{z}; f, p_{\hat{X}, \hat{Y}}))$ and (ii) that the group-aware prior $\hat{p}(\theta | \hat{z}; f, p_{\hat{X}, \hat{Y}})$ —which is itself an analytically intractable posterior distribution—is proportional to the analytically tractable function $\mathcal{J}(\theta, p_{\hat{X}, \hat{Y}})$. Noting that by taking the logarithm of the posterior in Equation (7), we get

$$\begin{aligned} & \log p(\theta | y_{\mathcal{D}}, x_{\mathcal{D}}; f, p_{\hat{X}, \hat{Y}}) \\ & \propto \log p(y_{\mathcal{D}} | x_{\mathcal{D}}, \theta) + \log \hat{p}(\theta | \hat{z}; f, p_{\hat{X}, \hat{Y}}) \quad (8) \\ & \propto \log p(y_{\mathcal{D}} | x_{\mathcal{D}}, \theta) + \mathcal{J}(\theta, p_{\hat{X}, \hat{Y}}), \end{aligned}$$

which we can write in the form of a standard optimization objective

$$\mathcal{F}(\theta) \doteq \sum_{n=1}^N \log p(y_{\mathcal{D}}^{(n)} | x_{\mathcal{D}}^{(n)}, \theta) + \mathcal{J}(\theta, p_{\hat{X}, \hat{Y}}),$$

where $\log p(y_{\mathcal{D}}^{(n)} | x_{\mathcal{D}}^{(n)}, \theta)$ is a data-fit term and $\mathcal{J}(\theta, p_{\hat{X}, \hat{Y}})$ is a regularization term that favors parameter values that have a high level of group robustness. For a Gaussian prior $p(\theta) = \mathcal{N}(\theta; \mu, \tau_{\theta}^{-1})$, we have $\log p(\theta) \propto -\frac{\tau_{\theta}}{2} \|\theta - \mu\|_2^2$, where μ could be any set of prior mean parameters, the final minimization objective takes the simple form

$$\mathcal{L}(\theta) = \sum_{n=1}^N \ell(y_{\mathcal{D}}^{(n)}, f(x_{\mathcal{D}}^{(n)}; \theta)) + \frac{\tau_{\theta}}{2} \|\theta - \mu\|_2^2 + \lambda \mathbb{E}_{p_{\hat{X}, \hat{Y}}}[\ell(\hat{Y}, f(\hat{X}; \theta + \rho\epsilon(\theta)))] \quad (9)$$

which we can compute via simple Monte Carlo estimation as

$$\hat{\mathcal{L}}(\theta) \doteq \underbrace{\sum_{n=1}^N \ell(y_{\mathcal{D}}^{(n)}, f(x_{\mathcal{D}}^{(n)}; \theta)) + \frac{\tau_{\theta}}{2} \|\theta - \mu\|_2^2}_{\text{standard } L_2\text{-regularized loss}} + \underbrace{\frac{\lambda}{S} \sum_{s=1}^S \ell(\hat{y}^{(s)}, f(\hat{x}^{(s)}; \theta + \rho\epsilon(\theta)))}_{\text{robustness regularization}} \quad (10)$$

where $(\hat{x}^{(s)}, \hat{y}^{(s)}) \sim p_{\hat{X}, \hat{Y}}$. This objective is amenable to optimization with stochastic gradient descent.

4.4 Practical Considerations

Deconstructing Loss Components. The optimization objective of Equation (10) contains two distinct terms: (i) The *standard L_2 -regularized loss* is computed on the training set and does not require any group labels. The *robustness regularization* is evaluated on the context distribution $p_{\hat{X}, \hat{Y}}$ sampled from the validation set. Defining $p_{\hat{X}, \hat{Y}}$ requires group labels to upweight the data as shown in Equation (3). Lastly, it is worth noting that only the robustness regularization incorporates a perturbation of the parameters. This reflects that we would like the prior to favor flatter minima that specifically improve generalization to minority groups.

Hyperparameters. The optimization objective in Equation (10) has three additional hyperparameters, compared to training with L_2 -regularized ERM. The parameter λ governs the strength of the empirical prior, the parameter ρ governs the strength of the parameter perturbation, and the parameter γ governs how strongly the distribution is reweighted.

Computational Complexity. The objective requires two additional forward passes on the M points sampled from the context distribution: one to compute epsilon and one to compute the loss under the parameter perturbation. In practice, this slowdown in training speed is not a problem since we only ever train for a handful of epochs, as detailed in Appendix A.

5 EMPIRICAL EVALUATION

Our experimental evaluation has two moving pieces: (i) datasets and (ii) last-layer retraining versus full finetuning on the group-labeled validation set. We outline the datasets we use, prior benchmark we compare to (including current state-of-the-art), and our experimental setup and show that our method achieves new SOTA or is competitive with the best methods. We also discuss the components that go into our designed prior and provide ablation studies to show their impact.

5.1 Datasets

We evaluate our method on both image classification and text datasets that are commonly used to benchmark the performance of group robustness methods.

Waterbirds. Waterbirds is a binary image classification problem generated synthetically by combining images of birds from the CUB dataset [Wah et al., 2011] and backgrounds from the Places dataset [Zhou et al., 2018]; the class corresponds to either land- or waterbird. 73% of images correspond to the majority group (waterbirds on water), 22% are landbirds on land and a sharply pronounced minority group of 1% landbirds on water, as well as 4% waterbirds on land. The distribution of backgrounds (land/water) on the validation and test sets is balanced.

CelebA. We consider a binary classification problem (‘blond vs. non-blond hair color’) with gender serving as the spurious feature. 94% of images with the blond labels show females. Models were trained on pixel intensities at the top of each image into a binary ‘blonde vs. not-blond’ label. No individual face characteristics, landmarks, keypoints, facial mapping, metadata, or any other information was used for training.

MultiNLI. MultiNLI is a text classification problem where the task is to classify the relationship between a given pair of sentences as a contradiction, entailment, or neither. In this dataset, the presence of negation words (e.g., ‘never’) in the second sentence is spuriously correlated with the ‘contradiction’ class.

5.2 Baselines

We consider seven baseline methods that make different assumptions about the availability of group attributes at training time. Empirical risk minimization (ERM) represents conventional training without any procedures for improving worst group accuracy. Just Train Twice [JTT; Liu et al., 2021] is a method that detects the minority group examples on train data, only using group labels on the validation set to tune hyperparameters. Correct-n-Contrast [CnC; Zhang et al.,

Table 1: Worst group and average accuracy on the test set of our method against a variety of other baselines in recent literature. We follow Sagawa et al. [2020] and reweigh the test accuracy for each group based on their proportion in the training data. The Group Info column details whether a method uses group labels in the training and validation dataset and whether it uses an auxiliary group labeling model. Accuracies of our method are estimated over ten trials. For a description of the baselines methods, see Section 5.2. We report the standard error of the mean, which sometimes requires adjusting the error bars of other baselines. The best-performing method is highlighted in gray and bolded, and the second-best-performing method is only bolded.

| Method | Group Info | | | Waterbirds | | CelebA | | MultiNLI | |
|-----------------------|------------|------|------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| | Tr. | Val. | Aux. | Worst | Average | Worst | Average | Worst | Average |
| ERM | N | N | N | 74.9±1.0 | 98.1±0.0 | 46.9±1.3 | 95.3±0.0 | 65.9±0.1 | 82.8±0.0 |
| JTT | N | Y | Y | 86.7 | 93.3 | 81.1 | 88.0 | 72.6 | 78.6 |
| CnC | N | Y | Y | 88.5±0.2 | 90.9±0.1 | 88.8±0.5 | 89.9±0.3 | — | — |
| SSA | N | Y | Y | 89.0±0.3 | 92.2±0.5 | 89.8±0.8 | 92.8±0.1 | 76.6±0.4 | 79.9±0.5 |
| DFR | N | Y | N | 92.9±0.1 | 94.2±0.2 | 88.3±0.5 | 91.3±0.1 | 74.7±0.3 | 82.1±0.1 |
| SUBG | Y | Y | N | 89.1±0.5 | — | 85.6±1.0 | — | 68.9±0.4 | — |
| G-DRO | Y | Y | N | 91.4 | 93.5 | 88.9 | 92.9 | 77.7 | 81.4 |
| GAP Last Layer | N | Y | N | 93.2±0.2 | 94.6±0.2 | 90.2±0.3 | 91.7±0.2 | 74.3±0.2 | 81.9±0.0 |
| GAP All Layers | N | Y | N | 93.8±0.1 | 95.6±0.1 | 90.2±0.3 | 91.5±0.1 | 77.8±0.6 | 82.5±0.1 |

2022] detects the minority group examples similarly to JTT and uses a contrastive objective to learn representations robust to spurious correlations. Group DRO [Sagawa et al., 2020] uses group information to train on a worst group loss objective and is a commonly used baseline. Deep Feature Reweighting [DFR; Kirichenko et al., 2023], uses a network finetuned with ERM and performs last-layer feature reweighting. SUBG [Idrissi et al., 2022] is carefully tuned ERM on a random subset of the data where the groups are equally represented. Finally, Spread Spurious Attribute [SSA; Nam et al., 2022] attempts to fully exploit the group-labeled validation data with a semi-supervised approach that propagates the group labels to the training data.

All baselines use pretrained models (a ResNet-50, pretrained on ImageNet1K for Waterbirds and CelebA, and a pretrained BERT model for MultiNLI), and all of them finetune the entire network except for DFR, where finetuning is only performed on the last layer (after full-parameter ERM training).

5.3 Experimental setup

For all experiments, we import a pretrained model and finetune all layers of the network using ERM on the target training dataset. For vision datasets, we finetune a ResNet-50 [He et al., 2016] pretrained on ImageNet [Russakovsky et al., 2015] for 30 epochs. For language datasets, we use pretrained BERT [Devlin et al., 2019] and finetune for five epochs. We assume that the validation dataset \mathcal{D}_{val} is group-annotated, and we use 85% of it to sample the context distribution $p_{\hat{x}, \hat{y}}$; the

remaining 15% are reserved for hyperparameter tuning. The total size of the validation datasets is 19,867 (out of 202,599) for CelebA, 1,199 (out of 11,788) for Waterbirds, and 82,462 (out of 412,349) for MultiNLI. In all cases, we heavily upweight the minority groups in the robust regularization term with $\gamma = 4, 1.5,$ and 2 for Waterbirds, CelebA, and MultiNLI, respectively. We use adversarial perturbations of the parameters in the robust regularization term with strength $\rho = 0.15$ for Waterbirds and CelebA and $\rho = 0.1$ for MultiNLI. We use the same configurations for last-layer and full-network training. For further details, see Appendix A.

5.4 Results

Table 1 presents our results. GAP achieves state-of-the-art performance for all three benchmarking tasks, demonstrating the value of optimizing with even a very simple data-driven prior. In particular, on Waterbirds and MultiNLI, GAP improves both worst-group and average accuracy even when compared to methods that require training data with group labels or use an auxiliary model to create attribute labels. For CelebA, GAP improves the worst group accuracy over all baselines, with only marginally lower average accuracy.

Perhaps most remarkably, we achieve state-of-the-art and close-to-state-of-the-art performance even when only retraining the last layer using group-labeled validation data. In particular, GAP applied to the last layer outperforms DFR, another method that only requires last-layer refitting, on the Waterbirds and CelebA) tasks, and is competitive to DFR on MultiNLI.

Table 2: Ablation on Adversarial Perturbation. We ablate the scaling parameter ρ in the GAP robustness regularizer (last-layer retraining only). GAP uses $\rho = 0.15$ for Waterbirds and CelebA and $\rho = 0.1$ for MultiNLI. The $\rho = 0$ setting indicates no parameter perturbation. Means and standard errors are estimated from ten trials.

| ρ | Waterbirds | | CelebA | | MultiNLI | |
|------------|------------|----------|----------|----------|----------|----------|
| | Worst | Average | Worst | Average | Worst | Average |
| GAP | 93.2±0.2 | 94.6±0.2 | 90.2±0.3 | 91.7±0.2 | 74.3±0.2 | 81.9±0.0 |
| 0 | 92.7±0.2 | 94.8±0.1 | 83.2±0.8 | 94.1±0.1 | 74.0±0.2 | 82.2±0.0 |

5.5 Ablation Studies

While the group-aware prior constructed in Section 4.2 is relatively simple, it adds several additional degrees of freedom. Most notably, it involves a scaled parameter perturbation, as can be seen in the robust regularizer of Equation (10), meant to favor flatter minima and better generalization to minority groups. Furthermore, the expected cost function is computed under a context distribution, which we construct by upweighting minority groups in the data as per Equation (3). Below, we show the impact of these design choices in two ablation studies (for further details, see Appendix A).

Ablation on Parameter Perturbation. Table 2 compares the performance of GAP (last-layer retraining only) with and without the perturbation of the parameters in the robustness regularization term. As expected, optimizing against a worst-case perturbation in the parameters leads to a small decrease in average accuracy, which is largely compensated for by a significant gain in worst group accuracy.

Ablation on Context Distribution. We have designed the context distribution $p_{\hat{X}, \hat{Y}}$ to upweight the minority groups (see Equation (3) and the paragraph following it). In Table 3, we compare different exponential weighting schemes from a completely group-balanced setting ($\gamma = 1$) to very strong upweighting ($\gamma = 4$) for Waterbirds (for last-layer retraining). Stronger upweighting of minority groups beyond the balanced setting is beneficial for improved worst-case accuracy.

6 Discussion

We presented a simple probabilistic framework for learning models that are robust to subpopulation shifts using group-aware prior distributions. Our empirical evaluation has shown that a probabilistically principled—and yet simple—prior distribution over neural network parameters reflecting group robustness desiderata, is able to achieve state-of-the-art performance on standard

Table 3: Ablation on Context Distribution. We ablate the upweighting strength in the context distribution $p_{\hat{X}, \hat{Y}}$, for exponential upweighting schemes with $\gamma \in \{1, 2, 4\}$ on the Waterbirds dataset using GAP (last-layer retraining). Means and standard errors are estimated from three trials for $\gamma \in \{1, 2\}$ and ten trials for $\gamma = 4$.

| Upweight Strength | Waterbirds | |
|---------------------------|-----------------|-----------------|
| | Worst | Average |
| Balanced ($\gamma = 1$) | 90.8±0.6 | 95.8±0.4 |
| Moderate ($\gamma = 2$) | 93.1±0.5 | 95.1±0.3 |
| Strong ($\gamma = 4$) | 93.2±0.2 | 94.6±0.2 |

benchmarking tasks without the need for any pseudo-labeling routine—even in the highly constrained setting of only retraining the last layer. This is achieved with minimal computational overhead and implementation complexity since, as we showed in Equation (9), MAP estimation with the simple group-aware prior can be reduced to adding an additional regularization term to the ERM optimization objective.

Flexibility: We have presented a very simple first example of a group-aware prior, which has already proven to be effective at improving robustness to subpopulation shifts. However, we are not limited to such simple priors. We have derived a general family of group robustness priors parameterized by a cost function and a context distribution, each of which can be specified using sophisticated models that satisfy group robustness desiderata. For example, a more sophisticated context distribution $p_{\hat{X}, \hat{Y}}$ could be defined by learning a generative model or by using a larger dataset without group information in conjunction with a learned group labeling model.

Complementarity: As noted above, methods proposed in related work—such as learned group labeling models, including those that do not require any group labels [e.g., Pezeshki et al., 2023], and data reweighting schemes—complement the framework presented in this paper.

Full Bayesian Inference: While we used MAP estimation to find the most likely parameters under the posterior, the probabilistic formulation of learning group robustness via uncertainty-aware priors lends itself to Bayesian inference and, as such, opens up routes for bringing to bear the vast arsenal of Bayesian inference methods. Inferring full posterior distribution using group-aware priors can improve generalization via Bayesian model averaging Wilson and Izmailov [2020] and lead to more reliable uncertainty estimation Rudner et al. [2023].

7 Acknowledgements

JK acknowledges support through NSF NRT training grant award 1922658. AGW acknowledges support through NSF HDR-2118310, CDS&E-MSS 2134216, CAREER IIS-2145492, I-DISRE 193471. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.
- Solon Barocas and Andrew D. Selbst. Big data’s disparate impact. *California Law Review*, 104(3):671–732, 2016. ISSN 00081221.
- Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenare, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.
- Christopher M. Bishop. Pattern recognition and machine learning (information science and statistics). 2006.
- Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, 2018.
- Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 2021.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.
- Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.
- Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938. PMLR, 10–15 Jul 2018.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Will Douglas Heaven. Hundreds of ai tools have been built to catch covid. none of them helped., 2021.
- Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.
- Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 336–351. PMLR, 11–13 Apr 2022.
- Pavel Izmailov, Dmitrii Podoprikin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, 34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018, pages 876–885. Association For Uncertainty in Artificial Intelligence (AUAI), 2018.
- Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35:38516–38532, 2022.
- Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023.
- Phuong Quynh Le, Jörg Schlötterer, and Christin Seifert. Is last layer re-training truly sufficient for robustness to spurious correlations?, 2023.
- Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018. doi: 10.1109/CVPR.2018.00566.
- Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6781–6792. PMLR, 18–24 Jul 2021.

-
- Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Kevin P. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013. ISBN 9780262018029 0262018020.
- Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *International Conference on Learning Representations*, 2022.
- Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*, 2019.
- Mohammad Pezeshki, Diane Bouchacourt, Mark Ibrahim, Nicolas Ballas, Pascal Vincent, and David Lopez-Paz. Discovering environments with xrm, 2023.
- Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and Fast Group Robustness by Automatic Feature Reweighting. *International Conference on Machine Learning (ICML)*, 2023.
- Joaquin Quiñero Candela. *Dataset shift in machine learning*. MIT Press, 2009.
- Hamed Rahimian and Sanjay Mehrotra. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:1–85, jul 2022. doi: 10.5802/ojmo.15.
- Tim G. J. Rudner, Sanyam Kapoor, Shikai Qiu, and Andrew Gordon Wilson. Function-space regularization in neural networks: A probabilistic perspective. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org, 2023.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.
- Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.
- Ravid Shwartz-Ziv, Micah Goldblum, Hossein Souri, Sanyam Kapoor, Chen Zhu, Yann LeCun, and Andrew Gordon Wilson. Pre-train your loss: Easy bayesian transfer learning with informative priors. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
- Nimit Sharad Sohoni, Maziar Sanjabi, Nicolas Ballas, Aditya Grover, Shaoliang Nie, Hamed Firooz, and Christopher Re. BARACK: Partially supervised group robustness with guarantees. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022.
- Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998. ISBN 978-0-471-03003-4.
- Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. *The Caltech-UCSD Birds-200-2011 Dataset*. Jul 2011.
- Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
- Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. In *International Conference on Machine Learning*, 2023.
- Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.
- Jingzhao Zhang, Aditya Menon, Andreas Veit, Srinadh Bhojanapalli, Sanjiv Kumar, and Suvrit Sra. Coping with label shift via distributionally robust optimisation. *arXiv preprint arXiv:2010.12230*, 2020.
- Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Re. Correct-contrast: a contrastive approach for improving robustness to spurious correlations. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26484–26516. PMLR, 17–23 Jul 2022.
- Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009.

Supplementary Material

Appendix A Experimental Details

Neural Network Architecture and Optimization. We follow the precedents set by Sagawa et al. [2020], Yang et al. [2023], and others. That is, all image datasets use a pre-trained ResNet-50 [He et al., 2016], and language datasets use a pre-trained BERT [Devlin et al., 2019]. Images are resized and cropped in the center to 224x224 pixels. For image datasets, we use SGD with momentum 0.9. For language datasets, we use AdamW with default parameters [Loshchilov and Hutter, 2017]. Table 4 shows the hyperparameters we use for the ERM-training step and Table 5 (last-layer) resp. Table 6 (all-layer) show the hyperparameters we use for the fine-tuning stage. We note that in the fine-tuning step with group-aware prior, we use a portion of the validation set for both terms of the loss in Equation (10), i.e. both terms are evaluated on points from the context set, which is chosen to be the validation dataset in our case. In principle, we could use the training set for the first term in Equation (10), as it does not require group labels, but we have chosen to evaluate both terms on the context distribution for simplicity.

Parameters for Ablation Studies. For Table 2, we use the hyperparameters detailed in Table 5 and set the relevant parameter to its ‘trivial’ realization. That is, we set $\rho = 0$. In this ablation study, we average our results over ten trials. For Table 3, we also use a ‘trivial’ parameter choice $\gamma = 1$, but also investigate results under different strengths of upweighting. In this study, we estimate our mean and standard errors from three trials, except for when the parameter choice coincides with our parameter choice from Table 5 (e.g. $\gamma = 4$), in which case we use ten.

Table 4: Table of Hyperparameter Choice for Initial ERM Finetuning. For all of our experiments that use a pretrained network fine-tuned on the target task with ERM, we detail the hyperparameters used. Initial LR means the initial learning rate input into the learning rate scheduler. Note that α is the minimum multiplier value for adjusting the learning rate for the cosine decay scheduler.

| Hyperparameter | Waterbirds | CelebA | MultiNLI |
|----------------|--------------|--------------|--------------|
| Epochs | 30 | 5 | 3 |
| Initial LR | 0.005 | 0.005 | 0.00002 |
| LR Scheduler | Cosine Decay | Cosine Decay | Linear Decay |
| α | 0.01 | 0.001 | — |
| Batch Size | 128 | 128 | 32 |
| Weight Decay | 0 | 0 | 0.01 |

Table 5: **Table of Hyperparameter Choice for Last-Layer Retraining.** For our experiments involving last-layer retraining on our group-aware prior, we detail the hyperparameters used. Epochs specifically indicate the number of epochs we re-train our last layer on, and does not include the previous ERM fine-tuning steps. Initial LR means the initial learning rate input into the learning rate scheduler. Note that α is the minimum multiplier value for adjusting the learning rate for the cosine decay scheduler. The parameters λ , γ , and ρ are hyperparameters specific to our method.

| Hyperparameter | Waterbirds | CelebA | MultiNLI |
|----------------|--------------|--------------|--------------|
| Epochs | 40 | 20 | 5 |
| Initial LR | 0.001 | 0.0001 | 0.00004 |
| LR Scheduler | Cosine Decay | Cosine Decay | Linear Decay |
| α | 1 | 1 | — |
| Batch Size | 128 | 128 | 32 |
| Weight Decay | 0 | 0 | 0 |
| λ | 1 | 30 | 10 |
| γ | 4 | 1.5 | 2 |
| ρ | 0.15 | 0.15 | 0.1 |

Table 6: **Table of Hyperparameter Choice for All-Layer Finetuning.** For our experiments involving all-layer finetuning on our group-aware prior, we detail the hyperparameters used. Epochs specifically indicate the number of epochs we finetune the entire network on, and does not include the previous ERM fine-tuning steps. Initial LR means the initial learning rate input into the learning rate scheduler. The parameters λ , γ , and ρ are hyperparameters specific to our method.

| Hyperparameter | Waterbirds | CelebA | MultiNLI |
|----------------|--------------|--------------|--------------|
| Epochs | 40 | 10 | 1 |
| Initial LR | 0.001 | 0.0001 | 0.000005 |
| LR Scheduler | Linear Decay | Linear Decay | Linear Decay |
| Batch Size | 128 | 128 | 32 |
| Weight Decay | 0 | 0 | 0 |
| λ | 15 | 200 | 10 |
| γ | 4 | 4 | 4 |
| ρ | 0.15 | 0.1 | 0.1 |

Appendix B Method Details

We define a general family of group-aware priors by specifying a Bernoulli observation model

$$\begin{aligned}\hat{p}(\hat{z} = 1 \mid \theta; f, p_{\hat{X}, \hat{Y}}) &= \exp(-\lambda \mathbb{E}_{p_{\hat{X}, \hat{Y}}} [c(\hat{X}, \hat{Y}, \theta)]) \\ \hat{p}(\hat{z} = 0 \mid \theta; f, p_{\hat{X}, \hat{Y}}) &= 1 - \hat{p}(\hat{z} = 1 \mid \theta; f, p_{\hat{X}, \hat{Y}}),\end{aligned}\tag{B.1}$$

where $c : \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^P \rightarrow \mathbb{R}_{\geq}$ is a ‘cost’ function and $\lambda > 0$ is a scaling parameter.

To define a specific member of this family of priors, we specify a cost function

$$c(\hat{x}, \hat{y}, \theta) \doteq \ell(\hat{y}, f(\hat{x}; \theta + \rho \epsilon(\theta, \hat{x}, \hat{y})))\tag{B.2}$$

where $\rho > 0$ is a scaling parameter, ℓ is a cross-entropy loss function for classification tasks and a mean-squared error loss function for regression tasks, and $\epsilon(\theta, \hat{x}, \hat{y})$ is a worst-case perturbation to the model parameters proposed in Foret et al. [2021] and given by

$$\epsilon(\theta, \hat{x}, \hat{y}) \doteq \perp \frac{\nabla_{\theta} \ell(\hat{y}, f(\hat{x}; \theta))}{\|\nabla_{\theta} \ell(\hat{y}, f(\hat{x}; \theta))\|_2},\tag{B.3}$$

where \perp is the `stop_gradient` operator.

It is worth briefly noting the implications of using the `stop_gradient` operator in a prior probability density function. Since the perturbation $\epsilon(\theta, \hat{x}, \hat{y})$ is a function of θ , it would normally be differentiable with respect to θ , which would affect the gradient of the robustness regularization term in the final objective given in Equation (10). Since we apply the `stop_gradient` operator, $\epsilon(\theta, \hat{x}, \hat{y})$ is treated as a constant for the purposes of gradient computation. Since the gradients of the robustness regularization term with respect to θ are therefore different when the `stop_gradient` operator is applied, this implies that the application of the `stop_gradient` operator implicitly changes the prior probability density function. To see how the application of the `stop_gradient` operator changes the prior probability density function, we first note that the gradient of the cost function *without* the `stop_gradient` operator in $\epsilon(\theta, \hat{x}, \hat{y})$, evaluated at parameters θ_t , is given by

$$\nabla_{\theta} \ell(\hat{y}, f(\hat{x}; \theta + \rho \tilde{\epsilon}(\theta)))|_{\theta=\theta_t} = \frac{d(\theta + \tilde{\epsilon}(\theta))}{d\theta} \Big|_{\theta=\theta_t} \nabla_{\theta} \ell(\hat{y}, f(\hat{x}; \theta))|_{\theta=\theta_t + \tilde{\epsilon}(\theta_t)}\tag{B.4}$$

$$= \nabla_{\theta} \ell(\hat{y}, f(\hat{x}; \theta))|_{\theta=\theta_t + \tilde{\epsilon}(\theta_t)} + \frac{d\tilde{\epsilon}(\theta)}{d\theta} \Big|_{\theta=\theta_t} \nabla_{\theta} \ell(\hat{y}, f(\hat{x}; \theta))|_{\theta=\theta_t + \tilde{\epsilon}(\theta_t)},\tag{B.5}$$

where

$$\tilde{\epsilon}(\theta) \doteq \frac{\nabla_{\theta} \ell(\hat{y}, f(\hat{x}; \theta))}{\|\nabla_{\theta} \ell(\hat{y}, f(\hat{x}; \theta))\|_2}.\tag{B.6}$$

In contrast, *with* the `stop_gradient` operator in $\epsilon(\cdot)$, the gradient of the cost function, evaluated at θ_t is given by

$$\nabla_{\theta} \ell(\hat{y}, f(\hat{x}; \theta + \rho \epsilon(\theta, \hat{x}, \hat{y})))|_{\theta=\theta_t} = \nabla_{\theta} \ell(\hat{y}, f(\hat{x}; \theta))|_{\theta=\theta_t + \epsilon(\theta_t)},\tag{B.7}$$

that is, it does not contain the term $\frac{d\tilde{\epsilon}(\theta)}{d\theta} \Big|_{\theta=\theta_t} \nabla_{\theta} \ell(\hat{y}, f(\hat{x}; \theta))|_{\theta=\theta_t + \tilde{\epsilon}(\theta_t)}$.

In order to obtain this gradient when using the perturbation function $\tilde{\epsilon}(\theta)$ (which does not use the `stop_gradient` operator), we need to include an additive term in the cost function whose gradient is equal to $\frac{d\tilde{\epsilon}(\theta)}{d\theta} \nabla_{\theta} \ell(\hat{y}, f(\hat{x}; \theta))$. To obtain this term, all we need to do is to take the integral of $\frac{d\tilde{\epsilon}(\theta)}{d\theta} \nabla_{\theta} \ell(\hat{y}, f(\hat{x}; \theta))$ with respect to θ :

$$A(\theta) \doteq - \int \frac{d\tilde{\epsilon}(\theta)}{d\theta} \nabla_{\theta} \ell(\hat{y}, f(\hat{x}; \theta)) d\theta.\tag{B.8}$$



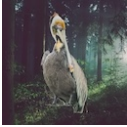
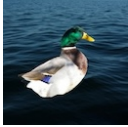




This integral exists for all evaluation points for which $\tilde{\epsilon}(\cdot)$ and $\ell(\hat{y}, f(\hat{x}; \cdot))$ are differentiable with respect to θ . While this integral may be analytically intractable, we do not need to compute it unless we want to compute the value of the prior probability density for a given parameter configuration, and since we only need the prior density for optimization, we can simply use the `stop_gradient` operator to compute the desired gradients.

Appendix C Dataset Details

Below, we provide information about each of the datasets used in our empirical evaluation. All images used in this manuscript are illustrative only (i.e., they differ from the actual images in the datasets used in this paper) and are taken from <https://unsplash.com> (for birds) and <https://www.istockphoto.com> (for people) under unlimited, perpetual, nonexclusive, worldwide license.

C.1 Vision Data

Table 7: Summary of Vision Dataset Properties.

| Waterbirds | | | | |
|-------------------|---|---|--|---|
| | g_1 | g_2 | g_3 | g_4 |
| Example Image |  |  |  |  |
| Group Description | Landbird on Land | Landbird on Water | Waterbird on Land | Waterbird on Water |
| Class Label | 0 | 0 | 1 | 1 |
| Attribute Label | 0 | 1 | 0 | 1 |
| Group Label | 0 | 1 | 2 | 3 |
| # Training Data | 3,498 | 184 | 56 | 1,057 |
| # Validation Data | 467 | 466 | 133 | 133 |
| # Test Data | 2,255 | 2,255 | 642 | 642 |
| CelebA | | | | |
| | g_1 | g_2 | g_3 | g_4 |
| Example Image |  |  |  |  |
| Group Description | Non-blonde Woman | Non-blonde Man | Blonde Woman | Blonde Man |
| Class Label | 0 | 0 | 1 | 1 |
| Attribute Label | 0 | 1 | 0 | 1 |
| Group Label | 0 | 1 | 2 | 3 |
| # Training Data | 71,629 | 66,874 | 22,880 | 1,387 |
| # Validation Data | 8,535 | 8,276 | 2,874 | 182 |
| # Test Data | 9,767 | 7,535 | 2,480 | 180 |

C.2 Language Data

Table 8: Summary of Language Dataset Properties.

| Multi-Genre Natural Language Inference (MultiNLI) corpus | | | |
|--|---|---|--|
| | <i>g</i> ₁ | <i>g</i> ₂ | <i>g</i> ₃ |
| Example Text | (P): <i>if residents are unhappy, they can put wheels on their homes and go someplace else, she said.</i> | (P): <i>within this conflict of values is a clash about art.</i> | (P): <i>there was something like amusement in the old man’s voice.</i> |
| | (H): <i>residents are stuck here but they can’t go anywhere else.</i> | (H): <i>there is no clash about art.</i> | (H): <i>the old man showed amusement.</i> |
| Group Description | Contradiction without Negations | Contradiction with Negations | Entailment without Negations |
| Class Label | 0 | 0 | 1 |
| Attribute Label | 0 | 1 | 0 |
| Group Label | 0 | 1 | 2 |
| # Training Data | 57,498 | 11,158 | 67,376 |
| # Validation Data | 22,814 | 4,634 | 26,949 |
| # Test Data | 34,597 | 6,655 | 40,496 |
| | <i>g</i> ₄ | <i>g</i> ₅ | <i>g</i> ₆ |
| Example Text | (P): <i>in 1988, the total cost for the postal service was about \$36.</i> | (P): <i>yeah but even even cooking over an open fire is a little more fun isn’t it.</i> | (P): <i>that’s not too bad.</i> |
| | (H): <i>the postal service cost us citizens almost nothing in the late 80’s.</i> | (H): <i>i like the flavour of the food.</i> | (H): <i>it’s better than nothing.</i> |
| Group Description | Entailment with Negations | Neutral without Negations | Neutral with Negations |
| Class Label | 1 | 2 | 2 |
| Attribute Label | 1 | 0 | 1 |
| Group Label | 3 | 4 | 5 |
| # Training Data | 1,521 | 66,630 | 1,992 |
| # Validation Data | 613 | 26,655 | 797 |
| # Test Data | 886 | 39,930 | 1,148 |