ν-BAYES: ROBUST BELIEF UPDATING UNDER SUBPOPULATION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Subpopulation shift changes the prevalence of latent groups between training and test distributions. Without group annotations or an explicit model of this shift, ordinary Bayesian updating treats the training mixture as representative, potentially favoring patterns supported by overrepresented groups. We introduce ν-Bayes, a group-annotation-free Bayesian method that learns the influence of each observation on the posterior, which in practice upweights informative minority examples while downweighting overrepresented ones. Starting from the principle of minimum discrimination information, we represent each observation through a soft constraint on its posterior expected loss. The associated Lagrange multipliers become learned likelihood powers, with ordinary Bayes recovered under a uniform allocation. Rather than specifying a separate tolerance for every observation, ν-Bayes learns a shared loss threshold, while ν controls the degree to which the likelihood powers may concentrate. We show that the resulting dual is an empirical distributionally robust optimization (DRO) problem over an order-∞ Rényi ambiguity set. Consequently, without using group annotations during optimization, its robust objective controls the empirical risk of every latent group containing at least a ν fraction of the training sample and provides population worst-group guarantees under subpopulation mixture shift. We combine primal–dual optimization with IVON to make the posterior update practical for deep neural networks. Experiments on controlled and benchmark shifts demonstrate improved worst-group performance over ordinary IVON and show how the learned likelihood powers redistribute influence across observations.

1 INTRODUCTION

Distribution shift is a central obstacle to the reliable deployment of machine learning models (Quiñonero-Candela et al., 2009; Koh et al., 2021). Bayesian methods are not immune: under certain forms of covariate shift, Bayesian model averaging can perform worse than classical point estimation, while even high-fidelity posteriors over Bayesian neural network parameters can yield poor predictive generalization under domain shift (Izmailov et al., 2021a;b). We focus on subpopulation mixture shift, where the distributions within latent groups remain stable but their relative proportions change between training and test.

In this setting, strong average performance can conceal substantially worse performance on underrepresented groups (Hashimoto et al., 2018). Such failures are especially consequential when groups correspond to protected attributes or meaningful subpopulations (Oakden-Rayner et al., 2020; Kreiman & Krishnapriyan, 2025). Many leading approaches rely on group annotations during training or model selection, including SUBG (Idrissi et al., 2022), GroupDRO (Sagawa et al., 2019), and group-aware priors (Rudner et al., 2024). Other methods reduce this reliance by identifying difficult or influential examples (Liu et al., 2021; Qiao et al., 2025; Cheng & Paige, 2026), or by retraining or replacing the final classifier (Qiu et al., 2023; To et al., 2025).

Unlike these approaches, we begin with Kullback's principle of minimum discrimination information (Kullback, 1959), also known as the minimum-change principle (May & Harper, 1976; Williams, 1980). From this principle, we derive a constrained formulation of Bayesian updating in which each observation imposes a separate constraint on its posterior expected loss. We refer to this formulation as ϵ-Bayes. Each Lagrange multiplier in ϵ-Bayes is associated with a sample-wise constraint and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

measures the marginal cost of tightening it. For negative log-likelihood losses, the multiplier becomes an observation-specific likelihood power, with ordinary Bayes recovered under uniform allocation. This makes observation-level influence explicit while recovering ordinary Bayesian inference as a special case.

Nonetheless, the constrained formulation requires a tolerance ϵ_i for every observation, which is rarely known in advance. We therefore introduce ν-Bayes, which replaces these individual tolerances with a shared learned loss threshold. The parameter ν controls how concentrated the resulting sample weights may become. We show that the dual is an empirical distributionally robust optimization problem over an order-∞ Rényi ambiguity set. Consequently, without using group annotations during optimization, the objective controls the empirical risk of every latent group containing at least a ν fraction of the training sample and yields population worst-group guarantees under subpopulation mixture shift.

To make this posterior update practical for deep neural networks, we combine projected primal–dual optimization with IVON (Shen et al., 2024). The resulting procedure learns observation-specific likelihood powers while maintaining a variational posterior and supporting posterior predictive inference. This distinguishes ν-Bayes from Bayesian approaches to subpopulation shift that encode robustness through the prior, either by modifying it or restricting its support (Wang et al., 2023; Rudner et al., 2024). Instead, ν-Bayes leaves the prior unchanged and adapts how the likelihood contribution of each observation enters the posterior update. In this respect, our method is more closely related to Bayesian data reweighting (Wang et al., 2017), with one important distinction: rather than placing a probability distribution over the observation weights, ν-Bayes obtains them from normalized Lagrange multipliers associated with the sample-wise constraints and learns them through primal–dual optimization.

Our contributions are threefold:

- **Group-annotation-free Bayesian updating.** We derive a sample-wise constrained formulation of Bayesian inference in which learned likelihood powers determine how strongly individual observations shape the posterior. We introduce ν-Bayes to learn these powers through a shared loss threshold without requiring group annotations.
- **Distribution-shift guarantees.** We establish the equivalence between ν-Bayes and empirical DRO over an order-∞ Rényi ambiguity set. This yields control over every latent group containing at least a ν fraction of the training sample, together with population worst-group guarantees under subpopulation mixture shift.
- **Scalable posterior inference.** We combine primal–dual optimization with IVON to obtain a practical variational learning algorithm for deep neural networks. We evaluate its robustness, posterior predictions, and learned allocation of likelihood influence on controlled and benchmark subpopulation shifts.

2 Information as Constraints

To derive the ϵ-Bayes formulation introduced above, we begin with the principle of minimum discrimination information (MDI) (Kullback, 1959):

Minimum Discrimination Information

Among all distributions compatible with the available information, choose the one that is least distinguishable from the reference distribution.

MDI provides a general rule for revising a reference distribution in light of new information. When distinguishability is measured by Kullback–Leibler (KL) divergence, Jaynes's maximum-entropy principle is recovered as the special case in which the reference distribution is uniform (Jaynes, 1957).

Taking the prior p as the reference distribution, we instantiate MDI by selecting a posterior q that minimizes its KL divergence from p, subject to constraints encoding the information supplied by the data. More broadly, this construction reflects the view that constraints encode information by restricting the set of admissible beliefs (Pratt, 1987). We apply this perspective at the level of

individual observations: each training example contributes a separate constraint that bounds its expected loss under q.

Formally, let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denote the training data, and let $\ell_i(\theta)$ denote a nonnegative loss of model parameters θ on example i. The resulting optimization problem is

$$\begin{aligned} \min_{q \in \mathcal{Q}} \quad & D_{\mathrm{KL}}(q\|p) \\ \text{s.t.} \quad & \mathbb{E}_q[\ell_i(\theta)] \le \epsilon_i, \qquad i = 1, \dots, N. \end{aligned} \tag{1}$$

Rather than aggregating the data into a single empirical-risk objective, equation 1 represents the information supplied by the dataset through a collection of sample-wise constraints. This yields a sample-centric inference procedure similar to the feasible-learning paradigm (Ramirez et al., 2025).

Unfortunately, the hard constraints in equation 1 may be infeasible or overly restrictive, particularly when observations are noisy, mutually conflicting, or poorly described by the model. One possible response is to discard observations that cannot be accommodated. Motivated instead by de Finetti's argument that suspected outliers should be accommodated within a Bayesian analysis rather than rejected (de Finetti, 1961), we soften each constraint by introducing a nonnegative slack variable. Penalizing these variables linearly allows difficult observations to incur a cost without being discarded, yielding

$$\begin{aligned} \min_{q \in \mathcal{Q},\, \xi \ge 0} \quad & D_{\mathrm{KL}}(q\|p) + \mathcal{C} \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & \mathbb{E}_q[\ell_i(\theta)] \le \epsilon_i + \xi_i, \qquad i = 1, \dots, N. \end{aligned} \tag{2}$$

Here, $\mathcal{C} > 0$ determines the penalty for constraint violations. We refer to equation 2 as ϵ-Bayes. We also show that, under the regularity conditions stated in Appendix A, the soft-constrained problem is feasible and satisfies strong duality. Moreover, unlike the hard-constrained formulation, it contains ordinary Bayesian inference as a special case. We empirically verify this result in a controlled experiment reported in Appendix D.1.

Proposition 1 (Bayes as a special case)**.** *Suppose that $\epsilon_i = 0$ for every i. Then ϵ-Bayes reduces to*

$$\min_{q \in \mathcal{Q}} D_{\mathrm{KL}}(q\|p) + \mathcal{C} \sum_{i=1}^N \mathbb{E}_q[\ell_i(\theta)]. \tag{3}$$

If $\ell_i(\theta) = -\log p(y_i \mid x_i, \theta)$ and $\mathcal{C} = 1$, then, when $\mathcal{Q}$ is unrestricted, the unique minimizer of equation 3 is the ordinary Bayesian posterior (Csiszár, 1975; Donsker & Varadhan, 1975; Zellner, 1988).

Furthermore, under conditions ensuring strong duality and primal–dual attainment, the KKT conditions imply that a primal–dual solution of ϵ-Bayes has a Gibbs-form posterior. For negative log-likelihood losses, the multiplier $\lambda_i^\star$ becomes the power of the likelihood factor associated with observation i.

Theorem 1 (ϵ-Bayes as Bayesian data reweighting)**.** *Suppose the regularity conditions for strong duality are met and $\mathcal{Q}$ contains all probability measures absolutely continuous with respect to p. If $(q^\star, \xi^\star, \lambda^\star)$ is a primal–dual solution of ϵ-Bayes, then*

$$dq^\star(\theta) = \frac{1}{Z(\lambda^\star)} \exp\left(-\sum_{i=1}^N \lambda_i^\star \ell_i(\theta) \right) dp(\theta), \qquad 0 \le \lambda_i^\star \le \mathcal{C}. \tag{4}$$

In particular, if $\ell_i(\theta) = -\log p(y_i \mid x_i, \theta)$, then

$$dq^\star(\theta) = \frac{1}{Z(\lambda^\star)} \prod_{i=1}^N p(y_i \mid x_i, \theta)^{\lambda_i^\star}\, dp(\theta). \tag{5}$$

Thus, equation 2 can be viewed as a form of Bayesian data reweighting (Wang et al., 2017), in which the multipliers $\lambda_i^\star$ determine how strongly the corresponding likelihood factors enter the posterior.

The use of MDI to motivate constrained learning under distribution shift has previously been explored by Sutter et al. (2021). Their construction projects the training distribution onto a family of test

distributions defined by known target moments. In contrast, our construction projects a prior over model parameters subject to sample-wise constraints on posterior expected losses. More broadly, KL projection problems with expectation constraints are well established in the information-theoretic and statistical literature (Friedlander & Gupta, 2006; Kumar & Sundaresan, 2015). Our contribution therefore lies not in the general constrained optimization form itself, but in the perspective from which we formulate the problem and the resulting interpretation of its dual variables.

Our ϵ-Bayes formulation is also closely related to generalized Bayesian updating (Bissiri et al., 2016), posterior regularization (Ganchev et al., 2007; 2010), and regularized Bayesian inference (Zhu et al., 2014). Indeed, both the hard- and soft-constrained formulations, equation 1 and equation 2, fall within the broad RegBayes framework. Our emphasis, however, is on our particular sample-wise construction, in which the posterior expected loss of each observation is constrained separately, as a result of applying MDI to inference. This setup also gives the corresponding dual variables a direct interpretation as shadow prices: each multiplier measures the marginal cost of tightening the associated constraint (Boyd & Vandenberghe, 2004). For negative log-likelihood losses, this interpretation becomes especially concrete: the multiplier λ_i is the power applied to observation i's likelihood factor. To our knowledge, this latter interpretation has not appeared in the works cited in this paragraph.

Our use of linearly penalized slack variables parallels that of slack variables in soft-margin support vector machines (Cortes & Vapnik, 1995); we make this connection precise in Appendix B. This relationship is natural because RegBayes subsumes maximum entropy discrimination (Jaakkola et al., 1999), which provides a probabilistic large-margin framework closely related to support vector machines (Zhu et al., 2014).

This section has therefore instantiated, through MDI, the perspective that constraints encode information about the data, exposed its consequences for likelihood allocation, and situated the formulation within the wider literature. Proofs are provided in Appendix A.

3 ν-BAYES: DERIVATION, PROPERTIES, AND ALGORITHM

Although ϵ-Bayes makes the contribution of individual observations explicit, it requires specifying a separate tolerance ϵ_i for each training example. These tolerances are rarely known in advance, and treating all N of them as hyperparameters is impractical for large datasets. Motivated by the ν-SVM parameterization (Schölkopf et al., 2000), which replaces a prescribed accuracy level with a learned threshold and an interpretable control on the number of support vectors, we replace the individual tolerances with a shared loss threshold ρ learned jointly with the posterior. This yields the following formulation, which we call ν-Bayes:

$$\begin{aligned} \min_{q\in\mathcal{Q},\,\rho\in\mathbb{R},\,\xi\geq 0} \quad & \nu D_{\mathrm{KL}}(q\|p) + \mathcal{C}\left(\nu N\rho + \sum_{i=1}^{N}\xi_i\right) \\ \text{s.t.} \quad & \mathbb{E}_q[\ell_i(\theta)] \leq \rho + \xi_i, \qquad i = 1,\ldots,N. \end{aligned} \tag{6}$$

Here, ρ is a data-adaptive loss threshold shared across all observations and serves as a candidate worst-case bound on sample loss, with the slacks permitting violations. The parameter $\nu \in (0, 1]$ controls the trade-off between lowering this threshold and penalizing aggregate violations above it, while $\mathcal{C} > 0$ controls the overall likelihood strength. The analogy with ν-SVM extends beyond the form of the objective: under the KKT conditions, ν lower-bounds the fraction of observations with positive dual weight and upper-bounds the fraction with positive slack. Consequently, $\nu = 1$ recovers uniform empirical weighting and, for negative log-likelihood loss with $\mathcal{C} = 1$, ordinary Bayesian inference. Smaller values of ν permit likelihood influence to concentrate on a smaller set of high-loss observations without changing its average strength. Further details are in Appendix C.1.

3.1 ν-BAYES AND SUBPOPULATION SHIFT

Interestingly, ν-Bayes admits an adversarial interpretation, and more specifically, an empirical DRO interpretation, motivated in part by the adversarial view of Bayesian inference developed by Husain & Knoblauch (2022). A key distinction is that their adversary perturbs the likelihood, or equivalently the loss in generalized Bayesian inference, whereas the adversary considered here reallocates probability

mass across the observed training examples. The shared threshold in ν-Bayes constrains these weights to an order-∞ Rényi ambiguity set, making ν-Bayes equivalent to an empirical DRO problem.

Proposition 2 (ν-Bayes as empirical DRO)**.** *Let $R_i(q) := \mathbb{E}_q[\ell_i(\theta)]$, $\widehat{P}_N := N^{-1}\sum_i \delta_{z_i}$, and $P_\omega := \sum_i \omega_i \delta_{z_i}$ for $\omega \in \Delta_N$, the probability simplex. For $\nu \in (0,1]$, define the order-∞ Rényi ambiguity set*

$$\mathcal{U}_\nu := \left\{\omega \in \Delta_N : D_\infty(P_\omega \| \widehat{P}_N) \le \log(1/\nu)\right\}, \tag{7}$$

where $D_\infty(P\|Q) := \log \operatorname{ess\,sup}_Q(dP/dQ)$. Then the minimizers of equation 6 coincide with those of

$$\min_{q\in\mathcal{Q}} \sup_{\omega\in\mathcal{U}_\nu} \left\{\tau D_{\mathrm{KL}}(q\|p) + \sum_{i=1}^N \omega_i R_i(q)\right\}. \tag{8}$$

Here and henceforth, $\tau := 1/(\mathcal{C}N) > 0$ denotes the posterior temperature in the normalized objective. Thus, ν-Bayes guards against the worst empirical reweighting whose order-∞ Rényi divergence from $\widehat{P}_N$ is at most $\log(1/\nu)$.

Beyond its adversarial interpretation, the empirical DRO characterization yields the following worst-group guarantee under subpopulation mixture shift.

Theorem 2 (Worst-group robustness of ν-Bayes)**.** *Let P_g denote the data distribution within group g, and let $G_g := \{i : z_i \text{ belongs to group } g\}$ and $n_g := |G_g|$. Define*

$$\widehat{R}_g(q) := \frac{1}{n_g}\sum_{i\in G_g} R_i(q), \qquad R_g(q) := \mathbb{E}_{\substack{z\sim P_g \\ \theta\sim q}}[\ell(\theta; z)].$$

Suppose $\ell(\theta; z) \in [0,1]$, the prior p is independent of the training sample, and, conditional on their group counts, the examples in group g are drawn independently from P_g. Let $n_{\min} := \min_g n_g$ and assume $\nu \le n_{\min}/N$. Define

$$\widehat{\mathcal{R}}_\nu(q) := \sup_{\omega\in\mathcal{U}_\nu} \sum_{i=1}^N \omega_i R_i(q).$$

Then, deterministically, for every posterior q,

$$\max_{1\le g\le K} \widehat{R}_g(q) \le \widehat{\mathcal{R}}_\nu(q). \tag{9}$$

Moreover, with probability at least $1-\delta$, simultaneously for every $q \ll p$,

$$\max_{1\le g\le K} R_g(q) \le \widehat{\mathcal{R}}_\nu(q) + \sqrt{\frac{D_{\mathrm{KL}}(q\|p) + \log\big(2K\sqrt{N}/\delta\big)}{2n_{\min}}}. \tag{10}$$

Consequently, this bound holds for the posterior q_ν returned by ν-Bayes. For any test distribution that changes only the group mixture, $P_{\text{test}} = \sum_{g=1}^K \alpha_g P_g$ with $\alpha \in \Delta_K$,

$$R_{\text{test}}(q_\nu) \le \max_g R_g(q_\nu) \le \widehat{\mathcal{R}}_\nu(q_\nu) + \sqrt{\frac{D_{\mathrm{KL}}(q_\nu\|p) + \log\big(2K\sqrt{N}/\delta\big)}{2n_{\min}}}. \tag{11}$$

For the zero–one loss, these risks are the classification errors of the Gibbs predictor induced by q_ν.

The parameter ν determines the resolution of the robust objective: it specifies the smallest fraction of the training sample that can be isolated and fully protected. The sufficient condition $\nu \le n_{\min}/N$ therefore requires ν to be no larger than the proportion of the smallest group. At this resolution, the robust objective controls the risk of every group, yielding a worst-group risk guarantee and, under zero–one loss, a corresponding guarantee on worst-group accuracy.

Another consequence of Proposition 2 is that the inner maximization over empirical reweightings is exactly the empirical conditional value-at-risk (CVaR) of the posterior-expected sample losses. Let $[u]_+ := \max\{u, 0\}$ and, using ν as the upper-tail probability, define

$$\mathrm{CVaR}_\nu(\mathbf{R}(q)) := \inf_{\rho\in\mathbb{R}} \left\{\rho + \frac{1}{\nu N}\sum_{i=1}^N [R_i(q) - \rho]_+\right\}, \tag{12}$$

where $\mathbf{R}(q) := (R_1(q), \ldots, R_N(q))$ and $R_i(q) := \mathbb{E}_q[\ell_i(\theta)]$ (Rockafellar & Uryasev, 2000).

Corollary 1 (ν-Bayes as regularized empirical CVaR optimization)**.** *For every posterior q,*

$$\sup_{\omega \in \mathcal{U}_\nu} \sum_{i=1}^{N} \omega_i R_i(q) = \mathrm{CVaR}_\nu(\mathbf{R}(q)).$$

Consequently, the minimizers of the empirical DRO problem in equation 8 coincide with those of

$$\min_{q \in \mathcal{Q}} \left\{ \tau D_{\mathrm{KL}}(q\|p) + \mathrm{CVaR}_\nu(\mathbf{R}(q)) \right\}. \tag{13}$$

This representation provides a second interpretation of the learned threshold ρ: it is an empirical $(1-\nu)$-quantile of the posterior-expected sample losses. In other words, ρ marks the onset of the upper loss tail formed by the worst-performing ν-fraction of the sample, and thus serves as an empirical value-at-risk threshold. Note also that CVaR is applied here to $R_i(q) = \mathbb{E}_q[\ell_i(\theta)]$ rather than to losses evaluated at a single parameter estimate. This distinction motivates a point-estimate CVaR baseline: comparing ν-Bayes against it allows us to assess the contribution of maintaining a posterior. Appendix C.4 derives the corresponding L_2-regularized CVaR objective as a concentrating-posterior limit. The proofs of Proposition 2 and Theorem 2 appear in Appendices C.2 and C.3, respectively.

3.2 ν-BAYES ALGORITHM

The empirical DRO formulation provides not only an interpretation of ν-Bayes, but also a practical route to its optimization. For this purpose, it is convenient to rescale the DRO weights as $\lambda_i = \nu N \omega_i$ and define the capped simplex

$$\Lambda_\nu := \left\{ \lambda \in [0,1]^N : \sum_{i=1}^{N} \lambda_i = \nu N \right\}. \tag{14}$$

Under this reparameterization, equation 8 becomes the following saddle-point problem

$$\min_{q \in \mathcal{Q}} \max_{\lambda \in \Lambda_\nu} \mathcal{L}_\nu(q, \lambda), \qquad \mathcal{L}_\nu(q,\lambda) := \tau D_{\mathrm{KL}}(q\|p) + \frac{1}{\nu N} \sum_{i=1}^{N} \lambda_i R_i(q). \tag{15}$$

The effective weight of observation i is therefore λ_i/ν, and $N^{-1}\sum_i \lambda_i/\nu = 1$, so the reweighting preserves the average scale of the empirical loss. Inspired by the primal–dual approach of Ramirez et al. (2025), we alternate between minibatch primal updates and epoch-level projected dual ascent. Algorithm 1 summarizes this procedure. Here, q_θ denotes the IVON posterior with mean θ, and $\widetilde{\theta} \sim q_\theta$ denotes sampled model parameters.

Algorithm 1 Alternating GDA for ν-Bayes

Require: Prior p, initial model parameters θ_0, posterior temperature τ, $\nu \in (0,1]$, primal and dual step sizes $\eta_\theta, \eta_\lambda$, dual-momentum coefficient $\beta \in [0,1)$, posterior samples M, and epochs T

1: Initialize $\theta \leftarrow \theta_0$, $\lambda \leftarrow \nu \mathbf{1}_N$, and $\bar{\lambda} \leftarrow \lambda$
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: Initialize the epoch loss vector $\widehat{\mathbf{R}}_t \leftarrow \mathbf{0}_N$
4: **for** each minibatch B in a shuffled pass through the training data **do**
5: Draw $\widetilde{\theta}^{(1)}, \ldots, \widetilde{\theta}^{(M)} \sim q_\theta$
6: Estimate $\widehat{r}_i \leftarrow M^{-1} \sum_{s=1}^{M} \ell_i(\widetilde{\theta}^{(s)})$ for each $i \in B$
7: Update θ with step size η_θ on $\tau D_{\mathrm{KL}}(q_\theta\|p) + |B|^{-1} \sum_{i \in B} (\bar{\lambda}_i/\nu)\widehat{r}_i$
8: Set $\widehat{R}_{t,i} \leftarrow \widehat{r}_i$ for each $i \in B$
9: **end for**
10: $\lambda \leftarrow \Pi_{\Lambda_\nu}(\lambda + \eta_\lambda \widehat{\mathbf{R}}_t)$
11: $\bar{\lambda} \leftarrow \beta\bar{\lambda} + (1-\beta)\lambda$
12: **end for**
13: **return** Posterior q_θ and sample weights $\bar{\lambda}/\nu$

However, unlike the basic update of Ramirez et al. (2025), our implementation introduces momentum into the dual weights (Algorithm 1, line 11). We find that this improves training stability by preventing the effective training mass from shifting too quickly toward a small number of examples, thereby preserving a more diverse training signal.

Table 1: Synthetic test results, reported as mean (SD) over five paired seeds; means and SDs are rounded to at most three and two significant figures, respectively. All methods use the same training, validation, and test sets within each seed. Boldface and underlining indicate the best and second-best predictive results. See Appendix D.2.

Method	Avg. acc. (%) ↑	Worst-group acc. (%) ↑	NLL ↓	$\|\theta_{\text{noise}}\|_2$	$\|\theta_{\text{core}}\|/\|\theta_{\text{spu}}\|$
Point-estimate methods					
ERM	57.4 (0.9)	22.9 (1.7)	1.01 (0.11)	0.599 (0.13)	1.28 (0.11)
GroupDRO	59.5 (3.7)	36.8 (5.9)	1.31 (0.41)	1.03 (0.59)	1.96 (0.40)
JTT	77.9 (4.6)	57.6 (8.5)	0.716 (0.21)	1.22 (0.13)	4.89 (1.0)
CVaR-Adam	<u>92.5 (3.3)</u>	<u>84.2 (7.0)</u>	0.372 (0.054)	0.162 (0.046)	9.68 (2.1)
Posterior methods					
Ordinary IVON	84.3 (6.3)	67.5 (12)	<u>0.339 (0.056)</u>	0.427 (0.021)	5.46 (1.2)
ν-Bayes ($\bar{\lambda}_i/\nu$)	**99.4 (0.5)**	**98.2 (1.5)**	**0.181 (0.008)**	0.258 (0.029)	16.3 (2.6)

4 Numerical Experiments

4.1 Synthetic Experiments

Following Idrissi et al. (2022), we construct a binary logistic-regression problem with label $y \in \{-1, +1\}$ and spurious attribute $a \in \{-1, +1\}$. Each example belongs to one of the four groups $g = (y, a)$ and has input

$$x = \begin{bmatrix} \gamma_{\text{spu}} x_{\text{spu}} \\ \gamma_{\text{core}} x_{\text{core}} \\ \gamma_{\text{noise}} x_{\text{noise}} \end{bmatrix}, \quad x_{\text{spu}} \sim \mathcal{N}(a, \sigma^2), \quad x_{\text{core}} \sim \mathcal{N}(y, \sigma^2), \quad x_{\text{noise}} \sim \mathcal{N}(0, \sigma^2 I_d).$$

During training, the spurious attribute agrees with the label with probability $\Pr(a = y) = 0.9$; at test time, we balance the four (y, a) groups. This construction gives the model a strong, easily learned spurious feature, a weaker invariant feature, and enough high-dimensional nuisance information to memorize individual training examples.

To assess ν-Bayes under subpopulation shift, we report worst-group accuracy, average accuracy, and predictive negative log-likelihood (NLL). We compare it with five baselines that probe the roles of group information, robust reweighting, and posterior inference. ERM (Vapnik, 1998) provides the standard average-risk reference. GroupDRO (Sagawa et al., 2019) uses group annotations to minimize worst-group loss. JTT (Liu et al., 2021) reweights examples misclassified by an initial model without using group annotations. CVaR-Adam optimizes the group-oblivious tail-risk objective from Section 3.1 with a point estimate, allowing us to assess whether robust reweighting alone explains the gains of ν-Bayes. Finally, ordinary IVON (Shen et al., 2024) provides a Bayesian baseline with uniform likelihood powers for assessing the value of adaptive powers.

On the synthetic dataset, ν-Bayes achieves the highest worst-group accuracy (98.2%) among the algorithms tested, approximately 14 percentage points above CVaR-Adam (84.2%) on paired data under the same training and selection protocol. It also achieves higher average accuracy (99.4% versus 92.5%) and lower NLL (0.181 versus 0.372). Because CVaR-Adam uses robust sample reweighting with a point estimate, these results suggest that reweighting alone does not explain the performance of ν-Bayes. Its larger core-to-spurious weight ratio (16.3 versus 9.68) is consistent with greater reliance on the invariant feature. ν-Bayes also exceeds the group-annotated GroupDRO baseline in worst-group accuracy, despite using no group annotations during optimization. Ordinary IVON shares the posterior optimizer settings of ν-Bayes; its lower worst-group accuracy (67.5%) supports a benefit from adaptive likelihood powers.

4.2 Image Classification

We evaluate ν-Bayes on three image-classification datasets: Colored MNIST (cMNIST), Waterbirds, and CelebA. We adapt cMNIST from Arjovsky et al. (2020) and use the SubpopBench setup (Yang et al., 2023) for Waterbirds and CelebA. On cMNIST, the task is to classify each digit into one of two classes. The label is spuriously correlated with color during training, making examples whose

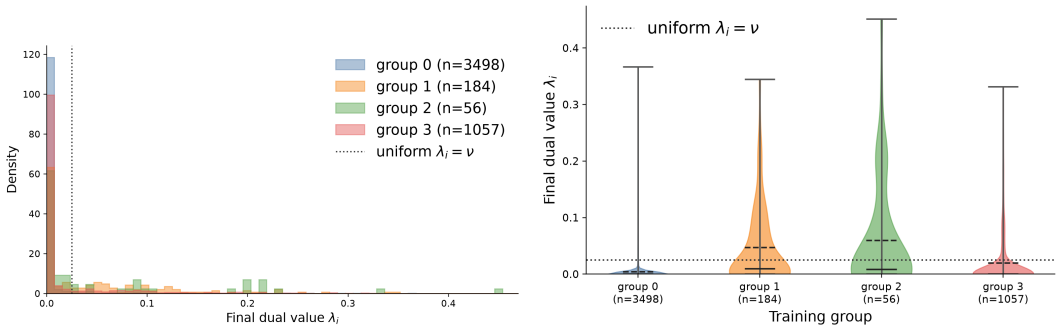
Table 2: Test worst-group accuracy (%), reported as mean (SD) over three paired seeds; means and SDs are rounded to one decimal place. Boldface and underlining indicate the best and second-best predictive results.

	Without group annotations					With group annotations		
Dataset	ERM	JTT	CVaR-Adam	IVON	ν-Bayes	Group-DRO	ERM-DFR	ν-Bayes-DFR
cMNIST	15.0 (2.7)	55.6 (6.1)	37.6 (16.5)	29.9 (11.6)	62.2 (5.7)	**69.3 (1.0)**	64.0 (3.3)	<u>67.6 (0.7)</u>
Waterbirds	71.7 (1.8)	68.6 (1.1)	77.5 (2.2)	70.1 (1.5)	<u>81.3 (4.3)</u>	74.4 (2.1)	80.4 (4.8)	**82.3 (1.6)**
CelebA	55.0 (3.6)	60.0 (8.9)	68.7 (1.2)	37.4 (14.3)	80.5 (5.1)	84.6 (4.2)	**89.4 (0.5)**	<u>86.0 (0.5)</u>

colors conflict with their labels the most challenging groups. We train a two-hidden-layer MLP following the architecture of Arjovsky et al. (2020). On Waterbirds, the task is to classify each image as containing either a waterbird or a landbird. Bird type is spuriously correlated with the background, making rare bird–background combinations the most challenging groups. On CelebA, the task is to predict whether a person has blond hair. Hair color is spuriously correlated with gender, making rare hair-color–gender combinations, particularly blond males, the most challenging groups. We train a ResNet-50 end to end on both Waterbirds and CelebA. Group annotations are withheld from ERM, JTT, CVaR-Adam, IVON, and ν-Bayes during fitting; GroupDRO and the DFR variants use them for fitting or retraining.

As shown in Table 2, robust reweighting alone is already beneficial: CVaR-Adam improves worst-group accuracy over ERM on all three datasets. ν-Bayes, in turn, outperforms CVaR-Adam in every case, suggesting that its posterior-based formulation provides benefits beyond point-estimate reweighting. The improvement of ν-Bayes over ordinary IVON further supports the value of learning observation-specific likelihood powers rather than weighting all observations uniformly. Together, these comparisons indicate that posterior-based adaptive likelihood allocation is key to the performance of ν-Bayes. Moreover, ν-Bayes remains competitive with methods that use group annotations during optimization or retraining. Applying deep feature reweighting (DFR) (Kirichenko et al., 2023) to the ν-Bayes backbone further improves worst-group accuracy on all three datasets, indicating that the two approaches are complementary. Appendix D.3 reports additional metrics beyond worst-group accuracy, following the recommendations of Yang et al. (2023).

Figure 1a shows that most training examples receive dual allocations close to zero, whereas the two bird–background mismatch groups account for most of the upper tail. The violin plots in Figure 1b further show that these groups have broader distributions and higher mean dual allocations than the matched groups, with the rarest group exhibiting the largest values. Since the corresponding likelihood power is $\lambda_i/(\tau\nu N)$, these results indicate that ν-Bayes assigns greater posterior influence to selected informative minority examples from underrepresented groups.



(a) Density histograms of the per-example dual variables λ_i, stratified by training group. Relative to Groups 0 and 3, Groups 1 and 2 account for most of the upper tail.

(b) Violin plots of the per-example dual variables λ_i, stratified by training group. Relative to Groups 0 and 3, Groups 1 and 2 exhibit heavier upper tails and higher means.

Figure 1: Distributions of the per-example dual variables λ_i on Waterbirds. The dotted line marks the uniform allocation $\lambda_i = \nu$.

5 Related Work

Subpopulation shift. Subpopulation shift changes the proportions of latent mixture components between training and deployment while leaving the component distributions unchanged. This setting has been studied in the statistical literature through finite-mixture models, robust estimation under contamination, and prior-probability shift (McLachlan & Peel, 2000; Huber, 1964; Saerens et al., 2002; Quiñonero-Candela et al., 2009). DRO provides a framework for promoting uniform performance across heterogeneous populations (Duchi & Namkoong, 2018); in particular, Duchi et al. (2023) establish finite-sample guarantees for every latent subpopulation whose mass exceeds a specified threshold. Despite substantial progress on subpopulation shift in machine learning, performance remains sensitive to the underlying shift mechanism (Yang et al., 2023). This is particularly true under attribute generalization, where some attribute values are absent from p_{train} but present in p_{test}; since no reweighting of the training data can recover a missing component, this case lies beyond the mixture-shift setting considered here. Our Bayesian approach requires no group annotations and derives robust sample weights as learned likelihood powers. This grounds adaptive reweighting in MDI and connects subpopulation robustness to principled belief updating. The perspective is increasingly relevant to LLM post-training, where aggregating heterogeneous feedback can underrepresent minority preferences, which motivates pluralistic and group-robust alignment methods (Chakraborty et al., 2024; Ramesh et al., 2024; Halpern et al., 2025).

Robust Bayesian inference. Bayesian inference can be sensitive to model misspecification, producing brittle posteriors and poor predictive performance (Owhadi et al., 2015; Grünwald & van Ommen, 2017; Izmailov et al., 2021a). Robust Bayesian methods mitigate this sensitivity by modifying the prior or likelihood, or by departing from ordinary conditioning (Box & Tiao, 1962; O'Hagan, 1979; Husain & Knoblauch, 2022). One prominent approach raises every likelihood contribution to a common learning rate η, yielding a fractional or tempered posterior as a special case of generalized Bayes (Barron & Cover, 1991; Walker & Hjort, 2001; Zhang, 2006; Bissiri et al., 2016; Grünwald & van Ommen, 2017). Although $\eta < 1$ is commonly used to reduce the influence of a misspecified likelihood, the appropriate rate depends on the learning problem and may even exceed one (de Heide et al., 2020). Coarsened posteriors take a different approach by conditioning on a neighborhood of the empirical distribution (Miller & Dunson, 2019), whereas Bayesian data reweighting places a probabilistic model over observation-specific weights (Wang et al., 2017). Our contribution is not the use of likelihood powers per se. Rather than applying a common learning rate to every observation, ν-Bayes derives observation-specific powers from normalized Lagrange multipliers of sample-wise MDI constraints. Moreover, unlike Bayesian data reweighting, it does not place a probability model over these powers; instead, ν controls their collective allocation. The resulting powers may vary across observations and fall below or above one, allowing the posterior update to attenuate some observations while emphasizing others. This adaptive allocation underpins our group-oblivious DRO interpretation and corresponding robustness guarantees.

6 Discussion and Conclusion

We introduced ν-Bayes, a Bayesian approach to subpopulation shift. We showed that ν-Bayes admits an empirical DRO interpretation, which offers an adversarial perspective on the method and leads to a practical algorithm for inference with deep neural networks. A natural extension is to test-time adaptation, where unlabeled target inputs could guide how influence is allocated towards the training observations most relevant to the deployment distribution. More broadly, we believe that treating information as a constraint offers a promising foundation for robust Bayesian updating under distribution shift, and one that merits further exploration.

AI Use Statement

We used generative AI tools to draft and polish sections of the manuscript, retrieve and discover relevant literature, develop and check mathematical arguments, and support research ideation and execution. The latter included refining hypotheses and experimental designs, implementing and debugging code, suggesting hyperparameters, and assisting with analysis and interpretation. This use followed an iterative, author-led process: AI outputs served as provisional drafts or prompts for ideation and were often rejected or substantially rewritten through repeated refinement. The authors

independently verified all citations, mathematical claims, code, data, and experimental results and take full responsibility for the paper's content.

Ethics Statement

This work studies robustness to latent subpopulation shift, but worst-group guarantees alone do not establish fairness or safety. Our experiments use synthetic data and the established cMNIST, Waterbirds, and CelebA benchmarks. CelebA contains images of people and demographic attributes; we use the existing benchmark release and do not attempt to identify individuals. The base ν-Bayes method does not use group annotations for fitting, although group labels are used for evaluation and by group-aware baselines and DFR. Deployment, especially in high-stakes settings, requires context-specific validation.

Reproducibility Statement

The appendix provides complete proofs and the full synthetic experimental protocol. The main text documents the datasets, baselines, metrics, and model selection criterion, while the accompanying anonymous code provides the ν-Bayes implementation, image-experiment configurations, and experiment scripts.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. URL https://arxiv.org/abs/1907.02893.

Andrew R. Barron and Thomas M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054, 1991. doi: 10.1109/18.86996. URL https://doi.org/10.1109/18.86996.

Pier Giovanni Bissiri, Chris C. Holmes, and Stephen G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B*, 78(5):1103–1130, 2016. doi: 10.1111/rssb.12158. URL https://doi.org/10.1111/rssb.12158.

George E. P. Box and George C. Tiao. A further look at robustness via bayes's theorem. *Biometrika*, 49(3/4):419–432, 1962. doi: 10.1093/biomet/49.3-4.419. URL https://www.jstor.org/stable/2333976.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. MaxMin-RLHF: Alignment with diverse human preferences. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6116–6135. PMLR, 2024. URL https://proceedings.mlr.press/v235/chakraborty24b.html.

Alice ST Cheng and Brooks Paige. Leave no one out: Mitigating subpopulation shift via leave-one-out upsampling. In *Symposium on Probabilistic Machine Learning*, 2026. URL https://openreview.net/forum?id=cJDphjWgB8.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. doi: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

Imre Csiszár. i-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975. doi: 10.1214/aop/1176996454.

Bruno de Finetti. The bayesian approach to the rejection of outliers. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 199–210. University of California Press, 1961.

Rianne de Heide, Alisa Kirichenko, Peter Grunwald, and Nishant Mehta. Safe-bayesian generalized linear regression. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2623–2633. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/heide20a.html.

Monroe D. Donsker and S. R. Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975. doi: 10.1002/cpa.3160280102.

John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018. URL https://arxiv.org/abs/1810.08750.

John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses for latent covariate mixtures. *Operations Research*, 71(2):649–664, 2023. doi: 10.1287/opre.2022.2363. URL https://doi.org/10.1287/opre.2022.2363.

Michael P Friedlander and Maya R Gupta. On minimizing distortion and relative entropy. *IEEE Transactions on Information Theory*, 52(1):238–245, 2006.

Kuzman Ganchev, Ben Taskar, and João V. Graça. Expectation maximization and posterior constraints. In *Advances in Neural Information Processing Systems 20*, pp. 569–576. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper/2007/hash/73e5080f0f3804cb9cf470a8ce895dac-Abstract.html.

Kuzman Ganchev, João V. Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049, 2010. URL https://jmlr.org/papers/v11/ganchev10a.html.

Peter Grünwald and Thijs van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017. doi: 10.1214/17-BA1085. URL https://doi.org/10.1214/17-BA1085.

Daniel Halpern, Evi Micha, Ariel D. Procaccia, and Itai Shapira. Pairwise calibrated rewards for pluralistic alignment. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1940. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/53dbd7e34fab703a639964e2d3ee9e84-Abstract-Conference.html.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1929–1938. PMLR, 2018. URL https://proceedings.mlr.press/v80/hashimoto18a.html.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964.

Hisham Husain and Jeremias Knoblauch. Adversarial interpretation of Bayesian inference. In *Proceedings of the 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pp. 553–572. PMLR, 2022. URL https://proceedings.mlr.press/v167/husain22a.html.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 336–351. PMLR, 2022. URL https://proceedings.mlr.press/v177/idrissi22a.html.

Pavel Izmailov, Patrick Nicholson, Sanae Lotfi, and Andrew Gordon Wilson. Dangers of Bayesian model averaging under covariate shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 3309–3322, 2021a. URL https://proceedings.neurips.cc/paper/2021/hash/1ab60b5e8bd4eac8a7537abb5936aadc-Abstract.html.

Pavel Izmailov, Sharad Vikram, Matthew D. Hoffman, and Andrew Gordon Wilson. What are bayesian neural network posteriors really like? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 4629–4640. PMLR, 2021b. URL https://proceedings.mlr.press/v139/izmailov21a.html.

Tommi Jaakkola, Marina Meila, and Tony Jebara. Maximum entropy discrimination. *Advances in neural information processing systems*, 12, 1999.

Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957. doi: 10.1103/PhysRev.106.620. URL https://doi.org/10.1103/PhysRev.106.620.

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. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts, 2021. URL https://arxiv.org/abs/2012.07421.

Tobias Kreiman and Aditi S. Krishnapriyan. Understanding and mitigating distribution shifts for machine learning force fields, 2025. URL https://arxiv.org/abs/2503.08674.

Solomon Kullback. *Information Theory and Statistics*. John Wiley & Sons, New York, 1959.

M Ashok Kumar and Rajesh Sundaresan. Minimization problems based on relative α-entropy I: Forward projection. *IEEE Transactions on Information Theory*, 61(9):5063–5080, 2015.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792. PMLR, 2021. URL https://proceedings.mlr.press/v139/liu21f.html.

Sherry May and William Harper. Toward an optimization procedure for applying minimum change principles in probability kinematic. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science: Volume I Foundations and Philosophy of Epistemic Applications of Probability Theory*, pp. 137–166. Springer, 1976.

Geoffrey J. McLachlan and David Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2000. doi: 10.1002/0471721182. URL https://doi.org/10.1002/0471721182.

Jeffrey W Miller and David B Dunson. Robust bayesian inference via coarsening. *Journal of the American Statistical Association*, 2019.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 151–159. ACM, 2020. doi: 10.1145/3368555.3384468. URL https://doi.org/10.1145/3368555.3384468.

Anthony O'Hagan. On outlier rejection phenomena in bayes inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(3):358–367, 1979. doi: 10.1111/j.2517-6161.1979.tb01090.x. URL https://academic.oup.com/jrsssb/article/41/3/358/7027507.

Houman Owhadi, Clint Scovel, and Tim Sullivan. On the brittleness of bayesian inference. *SIAM Review*, 57(4):566–582, 2015. doi: 10.1137/130938633.

Ian Pratt. Constraints, meaning and information. *Linguistics and Philosophy*, 10(3):299–324, 1987. URL https://www.jstor.org/stable/25001277.

Rui Qiao, Zhaoxuan Wu, Jingtan Wang, Pang Wei Koh, and Bryan Kian Hsiang Low. Group-robust sample reweighting for subpopulation shifts via influence functions. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5b6c1a77153c9d11861e26f92780cf2f-Abstract-Conference.html.

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28448–28467. PMLR, 2023. URL https://proceedings.mlr.press/v202/qiu23c.html.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence (eds.). *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA, 2009. doi: 10.7551/mitpress/9780262170055.001.0001. URL https://doi.org/10.7551/mitpress/9780262170055.001.0001.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free RLHF. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1171. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4147dfaa46cd7e20a2aecb91097ae8cc-Abstract-Conference.html.

Juan Ramirez, Ignacio Hounie, Juan Elenter, Jose Gallego-Posada, Meraj Hashemizadeh, Alejandro Ribeiro, and Simon Lacoste-Julien. Feasible learning. *arXiv preprint arXiv:2501.14912*, 2025. URL https://arxiv.org/abs/2501.14912.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

Tim G. J. Rudner, Ya Shi Zhang, Andrew Gordon Wilson, and Julia Kempe. Mind the GAP: Improving robustness to subpopulation shifts with group-aware priors. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 127–135. PMLR, 2024. URL https://proceedings.mlr.press/v238/rudner24a.html.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002. doi: 10.1162/089976602753284446. URL https://doi.org/10.1162/089976602753284446.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. URL https://arxiv.org/abs/1911.08731.

Bernhard Schölkopf, Alex J Smola, Robert C Williamson, and Peter L Bartlett. New support vector algorithms. *Neural computation*, 12(5):1207–1245, 2000.

Matthias Seeger. PAC-bayesian generalisation error bounds for gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002. URL https://jmlr.org/papers/v3/seeger02a.html.

Yuesong Shen, Nico Daheim, Bai Cong, Peter Nickl, Gian Maria Marconi, Clement Bazan, Rio Yokota, Iryna Gurevych, Daniel Cremers, Mohammad Emtiyaz Khan, and Thomas Mollenhoff. Variational learning is effective for large deep networks. *arXiv preprint arXiv:2402.17641*, 2024. URL https://arxiv.org/abs/2402.17641.

Tobias Sutter, Andreas Krause, and Daniel Kuhn. Robust generalization despite distribution shift via minimum discriminating information. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29754–29767, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f86890095c957e9b949d11d15f0d0cd5-Abstract.html.

Nguyen Nhat Minh To, Paul F. R. Wilson, Viet Nguyen, Mohamed Harmanani, Michael Cooper, Fahimeh Fooladgar, Purang Abolmaesumi, Parvin Mousavi, and Rahul Krishnan. Diverse prototypical ensembles improve robustness to subpopulation shift. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59761–59783. PMLR, 2025. URL https://proceedings.mlr.press/v267/to25a.html.

Vladimir Naumovich Vapnik. *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. Wiley and Sons, 1998.

Stephen Walker and Nils Lid Hjort. On Bayesian consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):811–821, 2001. doi: 10.1111/1467-9868.00314. URL https://doi.org/10.1111/1467-9868.00314.

Yixin Wang, Alp Kucukelbir, and David M Blei. Robust probabilistic modeling with bayesian data reweighting. In *International conference on machine learning*, pp. 3646–3655. PMLR, 2017.

Ziyu Wang, Binjie Yuan, Jiaxun Lu, Bowen Ding, Yunfeng Shao, Qibin Wu, and Jun Zhu. A constrained Bayesian approach to out-of-distribution prediction. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2248–2258. PMLR, 2023. URL https://proceedings.mlr.press/v216/wang23f.html.

Peter M. Williams. Bayesian conditionalisation and the principle of minimum information. *The British Journal for the Philosophy of Science*, 31(2):131–144, 1980.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39584–39622. PMLR, 2023. URL https://proceedings.mlr.press/v202/yang23s.html.

Arnold Zellner. Optimal information processing and bayes's theorem. *The American Statistician*, 1988. doi: 10.1080/00031305.1988.10475585.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006. doi: 10.1109/TIT.2005.864439. URL https://doi.org/10.1109/TIT.2005.864439.

Jun Zhu, Ning Chen, and Eric P. Xing. Bayesian inference with posterior regularization and applications to infinite latent SVMs. *Journal of Machine Learning Research*, 15:1799–1847, 2014. URL https://jmlr.org/papers/v15/zhu14a.html.