Fair Bayesian Data Selection via Generalized Discrepancy Measures

Yixuan Zhang[1*], Jiabin Luo[2*], Zhenggang Wang[1], Feng Zhou[3,4†], Quyu Kong[5]

[1]School of Statistics and Data Science, Southeast University, China
[2]School of Software and Microelectronics, Peking University, China
[3]Center for Applied Statistics and School of Statistics, Renmin University of China, China
[4]Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, China
[5]Alibaba Cloud, China
zh1xuan@hotmail.com, jbluo25@stu.pku.edu.cn, zgwangsrmkph@gmail.com, feng.zhou@ruc.edu.cn,
kongquyu.kqy@alibaba-inc.com

Abstract

Fairness concerns are increasingly critical as machine learning models are deployed in high-stakes applications. While existing fairness-aware methods typically intervene at the model level, they often suffer from high computational costs, limited scalability, and poor generalization. To address these challenges, we propose a Bayesian data selection framework that ensures fairness by aligning group-specific posterior distributions of model parameters and sample weights with a shared central distribution. Our framework supports flexible alignment via various distributional discrepancy measures, including Wasserstein distance, maximum mean discrepancy, and f-divergence, allowing geometry-aware control without imposing explicit fairness constraints. This data-centric approach mitigates group-specific biases in training data and improves fairness in downstream tasks, with theoretical guarantees. Experiments on benchmark datasets show that our method consistently outperforms existing data selection and model-based fairness methods in both fairness and accuracy.

1 Introduction

Artificial intelligence is rapidly expanding into key areas such as clinical diagnosis (Tiu et al. 2022), text generation (Gallegos et al. 2024), and financial credit approval (Khandani, Kim, and Lo 2010). While these advanced models are powerful, they often exhibit uneven performance across different groups, such as those defined by gender, race, or socioeconomic status, which leads to unfair decisions and raising concerns about fairness risks in real-world applications (Bird et al. 2016). As a result, ensuring fairness and preventing AI from exacerbating social inequalities have become critical challenges for both researchers and industry. Fairness-aware machine learning has thus emerged as a key area of study to address these issues.

Currently, most fairness-aware machine learning strategies focus on modifying the model itself (Zafar et al. 2015; Agarwal et al. 2018; Baharlouei, Patel, and Razaviyayn 2024; Chen et al. 2024). However, these approaches often re-
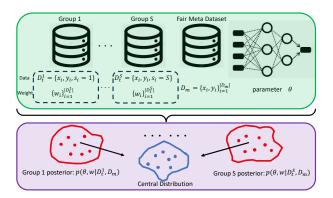
Figure 1: An illustration of Fair-BADS. Fair-BADS jointly infers model parameters and sample weights while reducing bias via posterior alignment to a central distribution.

quire building new models from scratch for each task, resulting in high computational costs and low efficiency, which makes them difficult to scale to large datasets and deep neural networks. Moreover, even trained fair models may still encounter unfairness issues during transfer due to generalization errors (Dutt et al. 2024).

Rather than addressing fairness solely during model training or transfer, a key challenge is to tackle the problem directly from the data. This is increasingly important as modern models rely heavily on massive raw data, which often contain imbalance and systemic biases. Meanwhile, manually identifying high-quality data from such large-scale sources is impractical. As a result, selecting high-quality and fairness-aware training data has become critical for building both effective and fair models (Xu et al. 2024).

Data selection offers a practical solution for improving model fairness by identifying or reweighting training examples that encourages fair outcomes. However, most existing methods focus on maximizing data utility or informativeness, often overlooking fairness and inadvertently reinforcing biases against underrepresented groups. Moreover, many rely on bi-level optimization or meta-learning, which are computationally intensive and difficult to scale (Fan et al. 2017). To address these limitations, Bayesian data

selection (Xu et al. 2024) formulates the task as posterior inference over model parameters and sample weights, using stochastic gradient Langevin dynamics (SGLD) for efficient optimization. This avoids nested optimization, enables standard gradient-based updates, and therefore scales well to large models and datasets. As manually curating high-quality data becomes infeasible at scale, principled data selection has become essential for reducing noise, imbalance, and bias, making it a critical component of scalable, fairness-aware learning.

Building on this motivation, we propose a fairness-aware data selection framework. While existing approaches often attempt to enforce fairness by adjusting model parameters in Euclidean space, this overlooks the fact that model parameters naturally reside in a more complex, generally non-Euclidean space. To better reflect fairness in this setting, it is crucial to consider the intrinsic geometry of the parameter space and the distribution of group-specific posteriors within it. We propose a Bayesian framework that formulates fairness as the alignment of posterior distributions across demographic groups toward a shared central distribution under a general class of divergence based objectives. Specifically, we jointly infer model parameters and sample weights using a fairness-aware meta-dataset, encouraging group-specific posteriors to align toward this central distribution via divergences such as Wasserstein distance, maximum mean discrepancy (MMD), or f-divergence. To efficiently approximate high-dimensional posteriors, we adopt Stein variational gradient descent (SVGD), which deterministically updates particles while preserving diversity. This enables stable and scalable inference without injecting noise and prevents dominant-group bias from overwhelming the learned posterior, making the framework well-suited for fairness-aware learning in large-scale settings.

Our contributions can be summarized as follows: (1) We propose a data-centric Bayesian framework for fairness-aware learning that jointly infers model parameters and instance weights, providing a scalable alternative to traditional model-centric approaches. (2) We propose a unified divergence-based formulation that aligns group-specific posteriors toward a shared central distribution, enabling flexible and geometry-aware fairness across demographic groups. (3) We provide theoretical guarantees by deriving discrepancy based bounds that approximate average group risk and bound intergroup performance gaps. (4) We use SVGD for efficient posterior inference, enabling stable updates without nested optimization and ensuring scalability for fairness-aware learning.

2 Related Works

Data selection. Most established selection strategies rely on bi-level optimization or meta-learning frameworks (Grangier, Ablin, and Hannun 2023; Ren et al. 2018; Shu et al. 2019; Zhang and Pfister 2021), which introduce an additional outer optimization loop to improve training data by maximizing model performance on a held-out meta set. These approaches, including sample-weighting (Grangier, Ablin, and Hannun 2023; Ren et al. 2018) and mini-batch reweighting (Fan et al. 2017), often require expensive meta-gradients or reinforcement learning, making them difficult to scale to large datasets and deep models. Other strategies rely on heuristics such as loss or confidence scores, for instance, curriculum learning (Bengio et al. 2009) favors easy samples, online methods (Loshchilov and Hutter 2015; Katharopoulos and Fleuret 2018; Jiang et al. 2019) prioritize high-loss or high-gradient examples, and confidence-based approaches (Cordeiro et al. 2023; Berthon et al. 2021) select uncertain instances. Most methods focus on performance and neglect fairness, with only a few adjusting sampling to meet fairness metrics (Roh et al. 2021). We address this gap with a Bayesian data selection framework that aligns group-specific posteriors to incorporate fairness directly.

Fairness-aware learning. Existing bias mitigation methods generally fall into three categories: preprocessing, in-processing, and post-processing. Preprocessing methods aim to reduce discriminatory information in the input data through fair representation learning (Louizos et al. 2015; Zemel et al. 2013; Lum and Johndrow 2016; Creager et al. 2019), fair data generation (Jang, Zheng, and Wang 2021), and data mapping (Calmon et al. 2017). In-processing methods reduce bias during training by incorporating fairness constraints into the learning process (Roh et al. 2020; Baharlouei, Patel, and Razaviyayn 2024; Donini et al. 2018; Gordaliza et al. 2019; Chiappa et al. 2020; Zhang, Lemoine, and Mitchell 2018). These can be model-specific (Bilal Zafar et al. 2015; Calders, Kamiran, and Pechenizkiy 2009) or model-agnostic (Agarwal et al. 2018; Lowy et al. 2022). Post-processing methods adjust model outputs to meet fairness criteria (Hardt, Price, and Srebro 2016). However, these approaches often face scalability and generalization challenges, motivating a shift toward data selection. Tahir, Cheng, and Liu (2023) is conceptually related in addressing both data distribution and posterior weight biases while overcoming the SGLD limits, ours instead enforces fairness via Bayesian data selection optimized with SVGD.

3 Preliminaries

In this section, we review the framework of Bayesian data selection from a fairness perspective and motivate the need for an efficient approach to posterior inference in this setting.

3.1 Bayesian Formulation for Fair Data Selection

Consider a training dataset $\mathcal{D}_t = \{(\mathbf{x}_i, y_i, s_i)\}_{i=1}^N$, where $\mathbf{x}_i$ denotes the non-sensitive features, $y_i \in \{0, 1\}$ is the binary label, and $s_i \in \{0, 1, \ldots, S\}$ represents the sensitive attribute, such as gender or race. The training set $\mathcal{D}_t$ may contain biased samples due to label corruption influenced by sensitive attributes. For instance, a qualified individual (i.e., $y_{\text{true}} = 1$) might be assigned a negative label ($y_{\text{obs}} = 0$) due to group-based prejudice, as in $p(y_{\text{obs}} = 0 \mid y_{\text{true}} = 1, s)$. Such biases can significantly degrade a model's fairness performance, especially when these patterns are learned and amplified during training.

To mitigate this issue, we assume access to a small meta-dataset $\mathcal{D}_m$ drawn from a fair target distribution, where labels are unaffected by sensitive attributes: $p(\mathbf{x}, y, s) =$

$p(y \mid \mathbf{x})p(\mathbf{x})p(s)$. Traditional data selection methods typically rely on bi-level optimization or meta-learning, where model parameters are trained on a reweighted training set, and the weights are optimized in an outer loop guided by $\mathcal{D}_m$. However, such approaches often incur high computational overhead and instability due to nested optimization.

The Bayesian formulation (Xu et al. 2024) offers a principled alternative by introducing a probabilistic model over both model parameters $\boldsymbol{\theta} \in \mathbb{R}^P$ and instance-level sample weights $\mathbf{w} \in \mathbb{R}^N$ applied to the training data $\mathcal{D}_t$. Then, the posterior distribution over $(\boldsymbol{\theta}, \mathbf{w})$ is given by:

$$
\begin{aligned}
p(\boldsymbol{\theta}, \mathbf{w} \mid \mathcal{D}_t, \mathcal{D}_m) &= \frac{p(\boldsymbol{\theta}, \mathcal{D}_m \mid \mathbf{w}, \mathcal{D}_t)p(\mathbf{w})}{p(\mathcal{D}_m \mid \mathcal{D}_t)} \\
&\propto p(\boldsymbol{\theta} \mid \mathbf{w}, \mathcal{D}_t)\, p(\mathcal{D}_m \mid \boldsymbol{\theta})\, p(\mathbf{w}),
\end{aligned}
\tag{1}
$$

where $p(\boldsymbol{\theta} \mid \mathbf{w}, \mathcal{D}_t)$ denotes the conditional distribution of model parameters given the sample weights and training data, $p(\mathcal{D}_m \mid \boldsymbol{\theta})$ encourages $\boldsymbol{\theta}$ toward fairness-aware generalization, and $p(\mathbf{w})$ is a prior over sample weights (e.g., sparsity-inducing or uniform). This formulation enables learning weights that prioritize training examples most compatible with the fairness-oriented meta-dataset $\mathcal{D}_m$.

3.2 Efficient Posterior Approximation

Inferring the joint posterior in Eq. (1) is generally intractable, particularly when $\boldsymbol{\theta}$ and $\mathbf{w}$ are high-dimensional. To enable scalable inference, we approximate $p(\boldsymbol{\theta}, \mathbf{w} \mid \mathcal{D}_t, \mathcal{D}_m)$ using tractable methods. Xu et al. (2024) proposed using SGLD, which augments stochastic gradient descent with Gaussian noise in each update step to simulate Langevin dynamics and sample from the posterior. While effective in many scenarios, SGLD suffers from slow convergence, sensitivity to step size and noise scale, often resulting in unstable training dynamics.

To overcome these limitations, we adopt SVGD, a deterministic, particle-based variational inference method that approximates the posterior by iteratively updating a set of particles via functional gradients in a reproducing kernel Hilbert space (RKHS) (Liu and Wang 2016; Wei et al. 2025). Each particle represents a "sample" from the posterior, and is transported toward high-density regions while maintaining diversity. Unlike SGLD, SVGD avoids the randomness and instability of stochastic samplers while better capturing the complex posterior. Formally, given M particles $\{\mathbf{z}^{(i)}\}_{i=1}^M$, the update rule is:

$$
\begin{aligned}
\mathbf{z}^{(i)} \leftarrow \mathbf{z}^{(i)} + \frac{\epsilon}{N} \sum_{j=1}^{N} &\Big[k(\mathbf{z}^{(j)}, \mathbf{z}^{(i)}) \nabla_{\mathbf{z}^{(j)}} \log p(\mathbf{z}^{(j)}) \\
&+ \nabla_{\mathbf{z}^{(j)}} k(\mathbf{z}^{(j)}, \mathbf{z}^{(i)}) \Big],
\end{aligned}
\tag{2}
$$

where $k(\cdot, \cdot)$ is a positive-definite kernel that defines particle interactions, and $\nabla_{\mathbf{z}^{(j)}} \log p(\mathbf{z}^{(j)})$ denotes the gradient of the log-posterior with respect to particle $\mathbf{z}^{(j)}$. This update encourages convergence to the posterior while mitigating particle collapse.

4 Methodology

In this section, we propose the *Fair Bayesian Data Selection* (Fair-BADS) framework (see Fig. 1), which jointly infers model parameters and sample weights with fairness considerations. While Bayesian data selection provides a principled and scalable alternative to bi-level optimization or meta-learning, existing approaches often overlook disparities across demographic groups. As a result, models trained under such frameworks may overfit to majority groups due to issues like class imbalance or group-dependent label bias in $\mathcal{D}_t$, leading to unfair performance across subpopulations.

To tackle this issue, Fair-BADS explicitly models group-specific posteriors and softly aligns them toward a central distribution. This central distribution serves as the group alignment target, defined via a divergence-based objective across group-specific posteriors. Fairness is then introduced at the distributional level by regularizing the divergence between each group-specific posterior and the central distribution. This allows the model to preserve group-specific signals while emphasizing globally fair samples.

Formally, we partition the training set $\mathcal{D}_t$ into demographic groups and define, for each group s, a posterior over model parameters $\boldsymbol{\theta}$ and sample weights $\mathbf{w}$ is:

$$
p_s(\boldsymbol{\theta}, \mathbf{w}) := p(\boldsymbol{\theta}, \mathbf{w} \mid \mathcal{D}_t^s, \mathcal{D}_m),
$$

where $\mathcal{D}_t^s \subseteq \mathcal{D}_t$ is the subset of $\mathcal{D}_t$ from group s with size $N_s = |\mathcal{D}_t^s|$. To allow alignment across groups with differing sizes, we embed each posterior into a common space of dimension $P + \bar{N}$, where P is the model parameter dimension and $\bar{N} = \max_s N_s$. For each group, the weight vector $\mathbf{w}$ is zero-padded to this common dimensionality, allowing all particles to reside in a consistent joint space and enabling consistent divergence computation across groups.

We assume that each demographic group induces a distinct posterior reflecting its statistical characteristics and potential biases. To mitigate inter-group disparities, we introduce a fairness-aware alignment mechanism during the inference by softly aligning the group specific posteriors $\{p_s(\boldsymbol{\theta}, \mathbf{w})\}_{s=1}^S$ toward the central distribution $p^\star(\boldsymbol{\theta}, \mathbf{w})$, defined as the minimizer of a divergence-based objective:

$$
p^\star(\boldsymbol{\theta}, \mathbf{w}) = \operatorname*{argmin}_p \sum_{s=1}^{S} \lambda_s D(p, p_s(\boldsymbol{\theta}, \mathbf{w})),
\tag{3}
$$

where $D(\cdot, \cdot)$ is a user-specified distributional discrepancy, such as Wasserstein distance, MMD, or f-divergence. The coefficients $\lambda_s \in (0, 1)$ satisfy $\sum_s \lambda_s = 1$ and control the contribution of each group ($\lambda_s = 1/S$ in our setting to ensure equal contribution). This central distribution guides the model to balance fairness across demographic groups under the chosen D. Specific instantiations and optimization strategies are discussed in subsequent sections.

4.1 Inference via SVGD

To approximate the joint posterior over $(\boldsymbol{\theta}, \mathbf{w})$, we adopt the SVGD algorithm, inspired by Wei et al. (2025). For each group s, we maintain a set of M particles $\{z_s^{(m)} =$

$(\boldsymbol{\theta}_s^{(m)}, \mathbf{w}_s^{(m)})\}_{m=1}^M$, where each particle represents a sample from the group-specific posterior $p_s(\boldsymbol{\theta}, \mathbf{w})$ and $z_s^{(m)} \in \mathbb{R}^{P+\bar{N}}$. According to Eq. (1), the log-posterior can be decomposed into three terms:

$$\log p_s(\boldsymbol{\theta}, \mathbf{w}) = \underbrace{-\sum_{i=1}^{\bar{N}} \sigma(w_i) \cdot \mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i)}_{\text{weighted training loss}}$$

$$-\underbrace{\sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_m} \mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i)}_{\text{meta loss}} + \underbrace{\left(\sum_{i=1}^{\bar{N}} \sigma(w_i) - \beta\bar{N}\right)^2}_{\text{weight prior (soft constraint)}}, \quad (4)$$

where $\mathcal{L}$ denotes the cross-entropy loss, and σ is the sigmoid function used to constrain each weight to $(0, 1)$. Following Xu et al. (2024), we define $p(\mathbf{w})$ implicitly via a soft prior, implemented as a regularization term that encourages the average weight to remain close to a predefined sparsity level β. Each particle is then updated using the SVGD (Eq. (2)):

$$\mathbf{z}_s^{(m)} \leftarrow \mathbf{z}_s^{(m)} + \frac{\epsilon}{M} \sum_{l=1}^M [k(\mathbf{z}_s^{(l)}, \mathbf{z}_s^{(m)}) \cdot \nabla_{\mathbf{z}_s^{(l)}} \log p_s(\mathbf{z}_s^{(l)})$$
$$+ \nabla_{\mathbf{z}_s^{(m)}} k(\mathbf{z}_s^{(l)}, \mathbf{z}_s^{(m)})], \quad (5)$$

where $\mathbf{z} = (\boldsymbol{\theta}, \mathbf{w}) \in \mathbb{R}^{P+\bar{N}}$ denote a sample in the padded parameter-weight space, $k(\cdot, \cdot)$ is a kernel function defined over the joint space to ensure smoothness and diversity across particles and the gradient term is:

$$\nabla_{\mathbf{z}} \log p_s(\mathbf{z})$$
$$= \begin{bmatrix} \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta} \mid \mathbf{w}, \mathcal{D}_t^s) + \nabla_{\boldsymbol{\theta}} \log p(\mathcal{D}_m \mid \boldsymbol{\theta}) \\ \nabla_{\mathbf{w}} \log p(\boldsymbol{\theta} \mid \mathbf{w}, \mathcal{D}_t^s) + \nabla_{\mathbf{w}} \log p(\mathbf{w}) \end{bmatrix}.$$

After completing the SVGD updates for each group, we obtain particle sets $\{z_s^{(m)}\}_{m=1}^M$, which are used to construct an empirical approximation of the group-specific posterior:

Group-specific Posterior: $\quad \tilde{p}_s(\mathbf{z}) = \frac{1}{M} \sum_{m=1}^M \delta(\mathbf{z} - \mathbf{z}_s^{(m)}),$

where $\delta(\cdot)$ denotes the Dirac delta function, and $\tilde{p}$ indicates that it is an empirical distribution supported on discrete particles. To encourage fairness across groups, we aim to align $\tilde{p}_s(\mathbf{z})$ toward a central distribution:

Central Distribution: $\quad \tilde{p}^\star(\mathbf{z}) = \frac{1}{M} \sum_{m=1}^M \delta(\mathbf{z} - \bar{\mathbf{z}}^{(m)}),$

where the central distribution is represented by a set of particles $\{\bar{\mathbf{z}}^{(m)}\}_{m=1}^M$, which serve as discrete support points summarizing the shared structure across all group-specific posteriors. Each particle $\bar{\mathbf{z}}^{(m)}$ lies in the same space $\mathbb{R}^{P+\bar{N}}$ as the group particles, allowing consistent comparison and alignment across distributions. These central particles are later obtained via central distribution computation (see Section 4.2), where we minimize a chosen distributional discrepancy to update $\{\bar{\mathbf{z}}^{(m)}\}_{m=1}^M$. We will detail this computation and update procedure in the following.

To incorporate fairness into the posterior inference, we modify the SVGD update by replacing $\nabla_{\mathbf{z}_s^{(l)}} \log p_s(\mathbf{z}_s^{(l)})$ in Eq. (5) with $\nabla_{\mathbf{z}_s^{(l)}} \log p_{\text{fair}}(\mathbf{z}_s^{(l)})$ that is defined as:

$$\log p_{\text{fair}}(\mathbf{z}) := \log p_s(\mathbf{z}) + \log p^\star(\mathbf{z}),$$

where a regularization term $\log p^\star(\mathbf{z})$ is introduced to softly encourage each group-specific posterior $\tilde{p}_s(\mathbf{z})$ to align with a shared central distribution $\tilde{p}^\star(\mathbf{z})$. This design guides particle updates to not only fit the group-specific posteriors but also remain close to the central distribution, thereby promoting fairness at the population level.

4.2 Computation of Central Distribution

To compute the central distribution, we solve a divergence minimization problem that aligns group-specific posteriors toward a central distribution, as defined in Eq. (3). We consider three representative divergence measures: Wasserstein distance, MMD, and f-divergence.

Wasserstein Distance. The Wasserstein distance $W_2(\cdot, \cdot)$ between two particle-based distributions is relatively easy to compute. To measure the discrepancy between the central distribution $\tilde{p}$ and a group posterior $\tilde{p}_s$, we define a cost matrix $\mathbf{C}_s \in \mathbb{R}_+^{M \times M}$, where each element is $\mathbf{C}_s[i, j] = \|\mathbf{z}_s^{(i)} - \bar{\mathbf{z}}^{(j)}\|_2^2$, representing the squared L2 cost between group and central particles. A transport plan $\mathbf{T}_s \in \mathbb{R}_+^{M \times M}$ specifies the amount of probability mass transported from $\mathbf{z}_s^{(i)}$ to $\bar{\mathbf{z}}^{(j)}$, subject to uniform marginal constraints:

$$W_2^2(\tilde{p}, \tilde{p}_s) = \min_{\mathbf{T}_s} \langle \mathbf{C}_s, \mathbf{T}_s \rangle_F,$$
$$\text{s.t.} \quad \mathbf{T}_s \mathbf{1} = \frac{1}{M}\mathbf{1}, \quad \mathbf{T}_s^\top \mathbf{1} = \frac{1}{M}\mathbf{1}, \quad (6)$$

where $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product and $\mathbf{1} \in \mathbb{R}^M$ is the all-ones vector. After solving Eq. (6) and obtaining the optimal plan $\mathbf{T}_s^\star$, the central distribution is computed by minimizing the weighted sum of Wasserstein distances:

$$\tilde{p}^\star = \underset{\{\bar{\mathbf{z}}^{(m)}\}_{m=1}^M}{\arg\min} \sum_{s=1}^S \lambda_s \langle \mathbf{C}_s, \mathbf{T}_s^\star \rangle_F.$$

Since the objective is quadratic in $\{\bar{\mathbf{z}}^{(m)}\}$, the closed-form solution for the central particles exists:

$$\bar{\mathbf{Z}}^\star = \sum_{s=1}^S \lambda_s \mathbf{T}_s^\star \mathbf{Z}_s \text{diag}(M^{-1}),$$

where $\mathbf{Z}_s \in \mathbb{R}^{M \times (P+\bar{N})}$ stacks group-specific particles, and $\bar{\mathbf{Z}}^\star \in \mathbb{R}^{M \times (P+\bar{N})}$ denotes the barycenter particles.

MMD. The MMD distance between two particle-based distributions can be computed via the kernel trick, using a kernel function $\tilde{k}(\cdot, \cdot)$ to measure the discrepancy between two sample sets. Specifically, the MMD between the central distribution $\tilde{p}$ and a group posterior $\tilde{p}_s$ is:

$$\text{MMD}^2(\tilde{p}, \tilde{p}_s) = \frac{1}{M^2} \sum_{i,j} \tilde{k}(\bar{\mathbf{z}}^{(i)}, \bar{\mathbf{z}}^{(j)}) + \frac{1}{M^2} \sum_{i,j} \tilde{k}(\mathbf{z}_s^{(i)}, \mathbf{z}_s^{(j)})$$
$$- \frac{2}{M^2} \sum_{i,j} \tilde{k}(\bar{\mathbf{z}}^{(i)}, \mathbf{z}_s^{(j)}).$$

The central distribution is obtained by minimizing the weighted sum of the MMDs across all groups:

$$\tilde{p}^{\star} = \operatorname*{argmin}_{\{\bar{\mathbf{z}}^{(m)}\}_{m=1}^{M}} \sum_{s=1}^{S} \lambda_s \mathrm{MMD}^2(\tilde{p}, \tilde{p}_s).$$

Due to the presence of the kernel, the objective no longer admits a closed-form solution. Therefore, the optimal central particles $\bar{\mathbf{Z}}^{\star}$ must be obtained via gradient descent.

f-divergence. The f-divergence is a general class of divergence measures, defined as:

$$D_f(p \,\|\, q) = \int q(\mathbf{z})\, f\left(\frac{p(\mathbf{z})}{q(\mathbf{z})}\right)\, d\mathbf{z},$$

where f is a convex function. By choosing different f, we recover various divergence measures. For example, when $f(t) = t \log t$, we obtain the KL divergence; when $f(t) = -\log t$, we get the reverse KL divergence; and when $f(t) = t \log \frac{2t}{t+1} + \log \frac{2}{t+1}$, we obtain the JS divergence.

The f-divergence between two particle-based distributions is generally intractable due to the need to integrate density ratios. We approximate it using kernel density estimation (KDE). Given a kernel function $\tilde{k}(\cdot, \cdot)$ and bandwidth $h > 0$, the KDEs for $\tilde{p}$ and $\tilde{p}_s$ are:

$$\hat{p}(\mathbf{z}) = \frac{1}{M} \sum_{i=1}^{M} \tilde{k}_h(\mathbf{z}, \bar{\mathbf{z}}^{(i)}), \quad \hat{p}_s(\mathbf{z}) = \frac{1}{M} \sum_{j=1}^{M} \tilde{k}_h(\mathbf{z}, \mathbf{z}_s^{(j)}).$$

Then the f-divergence between $\tilde{p}$ and $\tilde{p}_s$ is approximated as:

$$D_f(\tilde{p}\|\tilde{p}_s) \approx \frac{1}{M} \sum_{j=1}^{M} f\left(\frac{\hat{p}(\mathbf{z}_s^{(j)})}{\hat{p}_s(\mathbf{z}_s^{(j)}) + \epsilon}\right),$$

where $\epsilon > 0$ is a small constant added for numerical stability. The optimal central particles $\bar{\mathbf{Z}}^{\star}$ are then obtained by minimizing the weighted sum of f-divergences across all groups using gradient descent.

5   Theoretical Analysis

We present the theoretical guarantee for Fair-BADS from two perspectives: (i) a *discrepancy transfer* bound, showing that evaluating the model on the empirical central distribution $\tilde{p}^{\star}$ approximates the average group risk; and (ii) a *group fairness disparity* bound, demonstrating that performance gaps across groups are controlled when their posteriors align with the shared central distribution. Define

$$R_s(p) \triangleq \mathbb{E}_{\mathbf{z}\sim p}\big[\mathcal{L}_s(\mathbf{z})\big], \qquad R(p) \triangleq \mathbb{E}_{\mathbf{z}\sim p}\big[\mathcal{L}(\mathbf{z})\big],$$

where $\mathcal{L}$ is the loss function used in Eq. (4), and $R(\cdot)$ denotes the expected risk. The subscript s indicates it is computed w.r.t. group s only. And we use the following discrepancy-specific regularity assumptions.

(A1) Loss Regularity. For each group s, the loss $\mathcal{L}_s$ is bounded and satisfies:

$$\big|\mathbb{E}_p\mathcal{L}_s(\mathbf{z}) - \mathbb{E}_q\mathcal{L}_s(\mathbf{z})\big| \leq C_s D(p, q),$$

with C_s depending on the choice of D:

- $C_s = L_s$ if $D = W_2$ and $\mathcal{L}_s$ is L_s–Lipschitz;
- $C_s = \|\mathcal{L}_s\|_{\mathcal{H}_{\tilde{k}}}$ if $D = \mathrm{MMD}_{\tilde{k}}$ and $\mathcal{L}_s \in \mathcal{H}_{\tilde{k}}$;
- $C_s = B_s\sqrt{2c_f}$ if $D = D_f$ and $\mathcal{L}_s \in [0, B_s]$, where $\mathcal{H}_{\tilde{k}}$ is a reproducing kernel Hilbert space, B_s and c_f are some constants whose definitions are provided in Section A.

(A2) Cross–Group Compatibility. There exists $K < \infty$ such that for any $\mathbf{z}$ and s, s', $|\mathcal{L}_s(\mathbf{z}) - \mathcal{L}_{s'}(\mathbf{z})| \leq K$.

Theorem 1 (Discrepancy Transfer Bound) *Let $\tilde{p}^{\star}$ be the empirical central distribution minimizing Eq.* (3) *and define $\bar{R} \triangleq \sum_{s=1}^{S} \lambda_s R_s(\tilde{p}_s)$. Under (A1),*

$$\big|R(\tilde{p}^{\star}) - \bar{R}\big| \leq \sum_{s=1}^{S} \lambda_s C_s D(\tilde{p}_s, \tilde{p}^{\star}). \tag{7}$$

Concretely:

(Wasserstein) $\quad \big|R(\tilde{p}^{\star}) - \bar{R}\big| \leq \sum_s \lambda_s L_s W_2(\tilde{p}_s, \tilde{p}^{\star}).$

(MMD) $\quad \big|R(\tilde{p}^{\star}) - \bar{R}\big| \leq \sum_s \lambda_s \|\mathcal{L}_s\|_{\mathcal{H}_{\tilde{k}}} \mathrm{MMD}_{\tilde{k}}(\tilde{p}_s, \tilde{p}^{\star}).$

(f-divergence) $\big|R(\tilde{p}^{\star}) - \bar{R}\big| \leq \sum_s \lambda_s B_s\sqrt{2c_f} D_f(\tilde{p}_s\|\tilde{p}^{\star}).$

Theorem 2 (Group Fairness Disparity Bound) *Suppose Fair-BADS is run for $t = 0, 1, \ldots$, producing group posteriors $\{\tilde{p}_s^{(t)}\}_{s=1}^{S}$ and central $\tilde{p}_{\star}^{(t)}$. Define the* effective *cross–group gap restricted to the support of the current central:*

$$K_{\mathrm{eff}}(t) \triangleq \sup_{\mathbf{z}\in\mathrm{supp}(\tilde{p}_{\star}^{(t)})} \max_{s,s'} \big|\mathcal{L}_s(\mathbf{z}) - \mathcal{L}_{s'}(\mathbf{z})\big|.$$

Then for any s, s',

$$\begin{aligned}
&\left|\mathbb{E}_{\tilde{p}_s^{(t)}}\mathcal{L}_s - \mathbb{E}_{\tilde{p}_{s'}^{(t)}}\mathcal{L}_{s'}\right| \\
&\leq C_s D(\tilde{p}_s^{(t)}, \tilde{p}_{\star}^{(t)}) + C_{s'} D(\tilde{p}_{s'}^{(t)}, \tilde{p}_{\star}^{(t)}) + K_{\mathrm{eff}}(t) \quad (8) \\
&\leq 2C_{\max} \max_s D(\tilde{p}_s, \tilde{p}^{\star}) + K_{\mathrm{eff}}(t),
\end{aligned}$$

where $C_{\max} = \max_s C_s$.

The term $K_{\mathrm{eff}}(t)$ reflects intrinsic group-level difficulty and cannot be fully eliminated by alignment. However, it typically *decreases over iterations* as (i) $\tilde{p}_t^{\star}$ concentrates on low-loss parameter regions, and (ii) learned weights $\mathbf{w}$ downweight disparity-inducing samples. Section A provides sufficient conditions for $K_{\mathrm{eff}}(t) \to 0$ along with full assumptions and proofs.

6   Experiments

In the following sections, we first outline the experimental setup, then compare our method with related work across diverse image classification tasks under varying levels of label bias, followed by ablation studies on our selection strategy.

Method	Bias amount: 0.2				Bias amount: 0.4			
	ACC(↑)	DP(↓)	DDP(↓)	EO(↓)	ACC(↑)	DP(↓)	DDP(↓)	EO(↓)
UTKFace								
ERM	$0.848_{\pm 0.005}$	$\mathbf{0.062}_{\pm 0.011}$	$0.027_{\pm 0.011}$	$\mathbf{0.088}_{\pm 0.027}$	$0.770_{\pm 0.009}$	$0.252_{\pm 0.012}$	$0.051_{\pm 0.027}$	$0.291_{\pm 0.032}$
FairBatch	$0.803_{\pm 0.034}$	$0.082_{\pm 0.024}$	$0.048_{\pm 0.033}$	$0.123_{\pm 0.069}$	$0.721_{\pm 0.016}$	$0.254_{\pm 0.025}$	$0.058_{\pm 0.030}$	$0.294_{\pm 0.041}$
FERM	$0.843_{\pm 0.006}$	$0.069_{\pm 0.002}$	$0.030_{\pm 0.014}$	$0.095_{\pm 0.029}$	$0.780_{\pm 0.003}$	$0.225_{\pm 0.020}$	$0.059_{\pm 0.018}$	$0.296_{\pm 0.057}$
BLO	$0.805_{\pm 0.005}$	$0.099_{\pm 0.007}$	$0.028_{\pm 0.014}$	$0.154_{\pm 0.019}$	$0.726_{\pm 0.008}$	$0.249_{\pm 0.006}$	$0.057_{\pm 0.023}$	$\mathbf{0.286}_{\pm 0.027}$
BADS	$0.816_{\pm 0.013}$	$0.068_{\pm 0.013}$	$0.050_{\pm 0.013}$	$0.141_{\pm 0.023}$	$0.751_{\pm 0.015}$	$0.225_{\pm 0.008}$	$0.053_{\pm 0.014}$	$0.303_{\pm 0.066}$
Fair-BADS-W	$\mathbf{0.851}_{\pm 0.002}$	$\mathbf{0.062}_{\pm 0.006}$	$0.026_{\pm 0.003}$	$0.116_{\pm 0.011}$	$\mathbf{0.787}_{\pm 0.007}$	$0.219_{\pm 0.021}$	$0.060_{\pm 0.028}$	$0.309_{\pm 0.040}$
Fair-BADS-M	$0.849_{\pm 0.009}$	$0.073_{\pm 0.012}$	$0.031_{\pm 0.006}$	$0.132_{\pm 0.009}$	$0.781_{\pm 0.007}$	$0.214_{\pm 0.026}$	$0.050_{\pm 0.013}$	$0.295_{\pm 0.035}$
Fair-BADS-F	$0.849_{\pm 0.006}$	$0.069_{\pm 0.014}$	$\mathbf{0.023}_{\pm 0.002}$	$0.286_{\pm 0.014}$	$0.786_{\pm 0.007}$	$\mathbf{0.211}_{\pm 0.021}$	$\mathbf{0.043}_{\pm 0.012}$	$\mathbf{0.286}_{\pm 0.029}$
LFW-A								
ERM	$0.884_{\pm 0.005}$	$0.142_{\pm 0.020}$	$0.011_{\pm 0.006}$	$0.041_{\pm 0.020}$	$0.821_{\pm 0.019}$	$0.273_{\pm 0.025}$	$0.079_{\pm 0.040}$	$0.192_{\pm 0.048}$
FairBatch	$0.889_{\pm 0.008}$	$0.131_{\pm 0.006}$	$0.016_{\pm 0.015}$	$0.051_{\pm 0.015}$	$0.780_{\pm 0.014}$	$0.251_{\pm 0.009}$	$0.080_{\pm 0.016}$	$0.182_{\pm 0.018}$
FERM	$0.860_{\pm 0.056}$	$0.142_{\pm 0.070}$	$0.013_{\pm 0.003}$	$0.035_{\pm 0.013}$	$0.829_{\pm 0.013}$	$0.237_{\pm 0.042}$	$0.028_{\pm 0.021}$	$0.130_{\pm 0.044}$
BLO	$0.888_{\pm 0.002}$	$0.147_{\pm 0.010}$	$0.023_{\pm 0.012}$	$0.068_{\pm 0.016}$	$0.798_{\pm 0.024}$	$0.253_{\pm 0.024}$	$0.084_{\pm 0.015}$	$0.185_{\pm 0.028}$
BADS	$0.884_{\pm 0.004}$	$0.140_{\pm 0.025}$	$0.014_{\pm 0.012}$	$0.056_{\pm 0.023}$	$0.834_{\pm 0.018}$	$0.217_{\pm 0.019}$	$0.031_{\pm 0.017}$	$0.122_{\pm 0.025}$
Fair-BADS-W	$\mathbf{0.902}_{\pm 0.011}$	$\mathbf{0.129}_{\pm 0.004}$	$\mathbf{0.006}_{\pm 0.004}$	$0.042_{\pm 0.005}$	$\mathbf{0.859}_{\pm 0.006}$	$\mathbf{0.162}_{\pm 0.018}$	$\mathbf{0.012}_{\pm 0.007}$	$\mathbf{0.052}_{\pm 0.019}$
Fair-BADS-M	$0.901_{\pm 0.006}$	$0.133_{\pm 0.015}$	$0.010_{\pm 0.004}$	$0.034_{\pm 0.007}$	$0.850_{\pm 0.006}$	$0.186_{\pm 0.032}$	$0.024_{\pm 0.012}$	$0.090_{\pm 0.041}$
Fair-BADS-F	$0.900_{\pm 0.003}$	$0.132_{\pm 0.014}$	$0.014_{\pm 0.003}$	$\mathbf{0.033}_{\pm 0.009}$	$\mathbf{0.859}_{\pm 0.018}$	$0.189_{\pm 0.013}$	$0.014_{\pm 0.007}$	$0.079_{\pm 0.021}$
FairFace								
ERM	$0.716_{\pm 0.014}$	$0.170_{\pm 0.038}$	$0.045_{\pm 0.016}$	$0.198_{\pm 0.022}$	$0.656_{\pm 0.007}$	$0.392_{\pm 0.020}$	$0.030_{\pm 0.003}$	$0.402_{\pm 0.022}$
FairBatch	$0.685_{\pm 0.011}$	$0.139_{\pm 0.016}$	$0.044_{\pm 0.015}$	$0.168_{\pm 0.003}$	$0.629_{\pm 0.003}$	$0.328_{\pm 0.019}$	$0.044_{\pm 0.018}$	$0.357_{\pm 0.036}$
FERM	$0.699_{\pm 0.006}$	$0.416_{\pm 0.002}$	$\mathbf{0.026}_{\pm 0.013}$	$0.156_{\pm 0.032}$	$0.628_{\pm 0.006}$	$0.416_{\pm 0.002}$	$0.026_{\pm 0.016}$	$0.422_{\pm 0.017}$
BLO	$0.680_{\pm 0.005}$	$0.128_{\pm 0.002}$	$0.048_{\pm 0.020}$	$0.162_{\pm 0.021}$	$0.618_{\pm 0.002}$	$0.320_{\pm 0.014}$	$0.054_{\pm 0.008}$	$\mathbf{0.335}_{\pm 0.007}$
BADS	$0.661_{\pm 0.011}$	$0.159_{\pm 0.023}$	$0.055_{\pm 0.010}$	$0.203_{\pm 0.031}$	$0.632_{\pm 0.012}$	$0.369_{\pm 0.096}$	$0.047_{\pm 0.008}$	$0.400_{\pm 0.097}$
Fair-BADS-W	$0.718_{\pm 0.009}$	$0.140_{\pm 0.033}$	$0.038_{\pm 0.011}$	$0.165_{\pm 0.021}$	$0.662_{\pm 0.009}$	$0.341_{\pm 0.038}$	$0.026_{\pm 0.001}$	$0.350_{\pm 0.038}$
Fair-BADS-M	$\mathbf{0.719}_{\pm 0.008}$	$0.141_{\pm 0.031}$	$0.040_{\pm 0.011}$	$0.168_{\pm 0.022}$	$0.660_{\pm 0.009}$	$0.342_{\pm 0.036}$	$0.027_{\pm 0.008}$	$0.353_{\pm 0.028}$
Fair-BADS-F	$\mathbf{0.719}_{\pm 0.008}$	$\mathbf{0.126}_{\pm 0.035}$	$0.045_{\pm 0.012}$	$\mathbf{0.156}_{\pm 0.008}$	$\mathbf{0.663}_{\pm 0.008}$	$0.327_{\pm 0.042}$	$\mathbf{0.025}_{\pm 0.005}$	$\mathbf{0.335}_{\pm 0.045}$

Table 1: Evaluation results under different bias amount. For Fair-BADS, we report the results using three different variants.

6.1 Experimental Setup

For each dataset, we simulate label bias using group-dependent corruption strategies (Wick, Tristan et al. 2019). Unless otherwise specified, we use 20 particles per group and set the weight prior strength to $\beta = 0.005$. We use the JS divergence as our choice of f-divergence. The kernel k in SVGD and $\tilde{k}$ in MMD and f-divergence are both Gaussian kernels with adaptive bandwidth $h = 0.1$, numerical stability constant $\epsilon = $ 1e-3. The heuristic kernel may degrade in high-dimensional spaces, where norm-regularized (Grathwohl et al. 2020) or PDE-based kernels (Liu et al. 2019) provide more robust alternatives.

Datasets. We evaluate our method on three image datasets: **UTKFace** (Zhang, Song, and Qi 2017), Labeled Faces in the Wild with Attributes (**LFW-A**) (Wolf, Hassner, and Taigman 2011; Kumar et al. 2009), and **Fair-Face** (Karkkainen and Joo 2021). In UTKFace, race is used as the sensitive attribute and gender as the prediction target. For LFW-A, we predict gender and treat "HeavyMakeup" as the sensitive attribute due to its observed correlation with gender bias. In FairFace, we perform binary gender classification using race as the sensitive variable, grouping individuals as "White" or "Black" to evaluate fairness.

Baselines and Metrics. We evaluate our proposed method against several representative baselines, including standard empirical risk minimization (**ERM**), a sampling-based approach (**FairBatch** (Roh et al. 2021)), an in-processing fairness method using f-divergence (**FERM** (Baharlouei, Patel, and Razaviyayn 2024)), a reweighting-based data selection method (**BLO**) and a Bayesian data selection method (**BADS** (Xu et al. 2024)). For our Fair-BADS, we implement three variants based on different discrepancy measures: Wasserstein distance (Fair-BADS-W), MMD (Fair-BADS-M) and f-divergence (Fair-BADS-F). All methods are evaluated under a consistent experimental setup, using both accuracy and fairness metrics (Demographic Parity (**DP**), Difference in Demographic Parity (**DDP**) and Equal Opportunity (**EO**)). Each experiment is conducted with three runs, and we report the mean $\pm$ standard deviation.

6.2 Comparison Results

Table 1 summarizes the results across UTKFace, LFW-A and FairFace under varying levels of label bias. On both UTKFace and LFW-A, our Fair-BADS variants consistently achieve the best and competitive accuracy while reducing fairness disparities compared to other baselines. In particular, Fair-BADS-W yields the best overall trade-off, outperforming the original BADS in both accuracy and fairness metrics. Fair-BADS-W shows more stable improvements, though MMD and f-divergence variants also perform well. ERM and BLO tend to suffer from increasing fairness gaps under higher bias, while FairBatch and FERM reduce disparities at the cost of performance. Though BLO and BADS are originally designed to handle low quality data via data selection, they do not explicitly address fairness and thus in-

Method	LFW-A (bias amount: 0.2)				LFW-A (bias amount: 0.4)			
	ACC(↑)	**DP**(↓)	**DDP**(↓)	**EO**(↓)	**ACC**(↑)	**DP**(↓)	**DDP**(↓)	**EO**(↓)
Fair-BADS-W	$\mathbf{0.891}_{\pm 0.018}$	$0.144_{\pm 0.023}$	$0.014_{\pm 0.008}$	$0.049_{\pm 0.029}$	$0.847_{\pm 0.016}$	$0.192_{\pm 0.044}$	$0.034_{\pm 0.008}$	$0.095_{\pm 0.053}$
Fair-BADS-M	$0.890_{\pm 0.017}$	$0.142_{\pm 0.026}$	$\mathbf{0.010}_{\pm 0.005}$	$0.046_{\pm 0.028}$	$0.846_{\pm 0.005}$	$\mathbf{0.187}_{\pm 0.054}$	$\mathbf{0.029}_{\pm 0.015}$	$\mathbf{0.090}_{\pm 0.057}$
Fair-BADS-F	$0.889_{\pm 0.020}$	$\mathbf{0.133}_{\pm 0.025}$	$0.017_{\pm 0.004}$	$\mathbf{0.039}_{\pm 0.031}$	$\mathbf{0.850}_{\pm 0.020}$	$0.198_{\pm 0.040}$	$\mathbf{0.029}_{\pm 0.015}$	$0.093_{\pm 0.049}$

Table 2: Evaluation results on LFW-A, with CLIP-RN50 used as a zero-shot predictor for meta loss approximation.

Method	ACC (↑)	DP (↓)	DDP (↓)	EO (↓)
	Backbone: ResNet-18			
Fair-BADS-W	$0.820_{\pm 0.015}$	$\mathbf{0.049}_{\pm 0.017}$	$0.031_{\pm 0.016}$	$0.002_{\pm 0.001}$
Fair-BADS-M	$0.819_{\pm 0.007}$	$0.051_{\pm 0.007}$	$\mathbf{0.030}_{\pm 0.007}$	$\mathbf{0.001}_{\pm 0.001}$
Fair-BADS-F	$\mathbf{0.821}_{\pm 0.016}$	$0.051_{\pm 0.018}$	$\mathbf{0.030}_{\pm 0.017}$	$0.002_{\pm 0.001}$
	Backbone: DenseNet-121			
Fair-BADS-W	$0.841_{\pm 0.009}$	$0.080_{\pm 0.002}$	$\mathbf{0.004}_{\pm 0.002}$	$\mathbf{0.001}_{\pm 0.001}$
Fair-BADS-M	$\mathbf{0.845}_{\pm 0.011}$	$0.081_{\pm 0.003}$	$0.008_{\pm 0.004}$	$0.003_{\pm 0.002}$
Fair-BADS-F	$0.837_{\pm 0.008}$	$\mathbf{0.076}_{\pm 0.006}$	$0.007_{\pm 0.011}$	$0.002_{\pm 0.001}$
	Backbone: ViT-B/16			
Fair-BADS-W	$0.867_{\pm 0.014}$	$0.160_{\pm 0.005}$	$0.021_{\pm 0.006}$	$0.079_{\pm 0.005}$
Fair-BADS-M	$\mathbf{0.874}_{\pm 0.011}$	$\mathbf{0.156}_{\pm 0.007}$	$\mathbf{0.018}_{\pm 0.006}$	$\mathbf{0.073}_{\pm 0.007}$
Fair-BADS-F	$0.866_{\pm 0.016}$	$0.160_{\pm 0.005}$	$0.021_{\pm 0.008}$	$0.079_{\pm 0.007}$

Table 3: Comparison of Fair-BADS variants across backbones under bias level 0.4.

Figure 2: Comparison of sample weight distributions across demographic groups. Left: KDE of sample weights **w** at the final training epoch for groups $s = 0$ and $s = 1$. Right: Wasserstein distance between group-specific weight distributions over training epochs.

advertently reinforce bias by prioritizing samples from the majority group. To validate these improvements, we conduct paired t-tests and find that on UTKFace, Fair-BADS-W significantly outperforms the next-best method (BADS) in both accuracy and fairness across all bias levels ($p < 0.001$). On LFW-A, it also shows significant gains, especially under high bias ($p < 0.01$). On FairFace, Fair-BADS-F and Fair-BADS-M consistently outperform the next-best method, with significant improvements in accuracy and DP ($p < 0.05$).

6.3 Learning without Meta Dataset

In scenarios where no explicit meta dataset $\mathcal{D}_m$ is available, we approximate the meta objective using a zero-shot predictor $f^*(\mathbf{x})$ trained on external data (e.g., CLIP-RN50). Instead of directly evaluating $p(\mathcal{D}_m \mid \boldsymbol{\theta})$, we estimate it as:

$$\log p(\mathcal{D}_m \mid \boldsymbol{\theta}) \approx -\mathrm{KL}\big[p_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}) || p(\mathbf{y} \mid f^*(\mathbf{x}))\big], \quad (9)$$

where $p_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})$ is the model's output distribution and $p(\mathbf{y} \mid f^*(\mathbf{x}))$ is the pseudo label distribution induced by the zero-shot predictor. This KL divergence acts as a surrogate meta loss, allows us to avoid explicit collection of a meta set. To compute it, we reserve a small fraction (1%) of the training data as $\mathcal{D}_m^{\mathrm{pseudo}}$, which is excluded from training loss throughout training. As shown in Table 2, even under the situation when meta set is not available, our method still outperforms all baseline approaches in both accuracy and fairness metrics, despite showing slightly lower performance compared to the performance use explicit meta set. This highlights the framework's practical advantage in settings where collecting a clean meta set is infeasible.
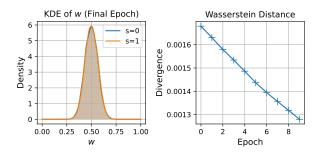
6.4 Ablation Studies

To assess architectural impact, we compare three backbones: ResNet-18 (He et al. 2016), DenseNet-121 (Huang et al. 2017), and ViT-B/16 (Dosovitskiy et al. 2021). As shown in Table 3, ViT-B/16 yields the best accuracy, while all variants maintain low fairness metrics. This confirms that our method generalizes well across architectures and that MMD and Wasserstein distances provide more stable fairness control than f-divergence.

Beyond architectural variations, we also examine how fairness emerges throughout training. In Fig. 2, the KDE plot (left) shows near-identical sample weight distributions across groups by the final epoch, indicating unbiased data selection. The Wasserstein distance (right) decreases during training, confirming that group posteriors align progressively. This supports the effectiveness of barycenter-based alignment in improving fairness.

7 Conclusions

We propose *Fair-BADS*, a framework that addresses fairness by combining group-specific inference with distributional alignment. Unlike prior methods that overlook group disparities, we model group-specific posteriors and align them via a shared central distribution, acting as a soft regularizer in SVGD. This approach ensures inter-group consistency without adversarial training or hard constraints. Our particle-based inference is scalable and naturally promotes distributional fairness. Experiments demonstrate improved fairness with strong task performance. Future directions include extending to continuous attributes, dynamic group discovery, and complex tasks like language generation.

8 Acknowledgments

References

Agarwal, A.; Beygelzimer, A.; Dudik, M.; Langford, J.; and Wallach, H. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 60–69. PMLR.

Baharlouei, S.; Patel, S.; and Razaviyayn, M. 2024. f-FERM: A Scalable Framework for Robust Fair Empirical Risk Minimization. In *The Twelfth International Conference on Learning Representations*.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum Learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, 41–48. New York, NY, USA: Association for Computing Machinery.

Berthon, A.; Han, B.; Niu, G.; Liu, T.; and Sugiyama, M. 2021. Confidence Scores Make Instance-dependent Label-noise Learning Possible. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 825–836. PMLR.

Bilal Zafar, M.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2015. Fairness Constraints: Mechanisms for Fair Classification. *arXiv e-prints*, arXiv:1507.05259.

Bird, S.; Barocas, S.; Crawford, K.; and Wallach, H. 2016. Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning, New York University*, 4.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building Classifiers with Independency Constraints. In *2009 IEEE International Conference on Data Mining Workshops*, 13–18.

Calmon, F.; Wei, D.; Vinzamuri, B.; Natesan Ramamurthy, K.; and Varshney, K. R. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems 30*, 3992–4001.

Chen, C.; Zhang, Y.; Li, Y.; Wang, J.; Qi, L.; Xu, X.; Zheng, X.; and Yin, J. 2024. Post-training attribute unlearning in recommender systems. *ACM Transactions on Information Systems*, 43(1): 1–28.

Chiappa, S.; Jiang, R.; Stepleton, T.; Pacchiano, A.; Jiang, H.; and Aslanides, J. 2020. A General Approach to Fairness with Optimal Transport. In *AAAI*, 3633–3640. AAAI Press.

Cordeiro, F. R.; Sachdeva, R.; Belagiannis, V.; Reid, I.; and Carneiro, G. 2023. LongReMix: Robust learning with high confidence samples in a noisy label environment. *Pattern Recognition*, 133: 109013.

Creager, E.; Madras, D.; Jacobsen, J.; Weis, M. A.; Swersky, K.; Pitassi, T.; and Zemel, R. S. 2019. Flexibly Fair Representation Learning by Disentanglement. *CoRR*, abs/1906.02589.

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J.; and Pontil, M. 2018. Empirical risk minimization under fairness constraints. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 2796–2806. Red Hook, NY, USA.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Dutt, R.; Bohdal, O.; Tsaftaris, S. A.; and Hospedales, T. 2024. FairTune: Optimizing Parameter Efficient Fine Tuning for Fairness in Medical Image Analysis. In *The Twelfth International Conference on Learning Representations*.

Fan, Y.; Tian, F.; Qin, T.; Bian, J.; and Liu, T.-Y. 2017. Learning What Data to Learn. arXiv:1702.08635.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3): 1097–1179.

Gordaliza, P.; Barrio, E. D.; Fabrice, G.; and Loubes, J.-M. 2019. Obtaining Fairness using Optimal Transport Theory. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2357–2365. PMLR.

Grangier, D.; Ablin, P.; and Hannun, A. 2023. Adaptive Training Distributions with Scalable Online Bilevel Optimization. arXiv:2311.11973.

Grathwohl, W.; Wang, K.-C.; Jacobsen, J.-H.; Duvenaud, D.; and Zemel, R. 2020. Learning the Stein Discrepancy for Training and Evaluating Energy-Based Models without Sampling. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 3732–3747. PMLR.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. *CoRR*, abs/1610.02413.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778.

Huang, G.; Liu, Z.; van der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *CVPR*, 2261–2269.

Jang, T.; Zheng, F.; and Wang, X. 2021. Constructing a Fair Classifier with Generated Fair Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9): 7908–7916.

Jiang, A. H.; Wong, D. L.; Zhou, G.; Andersen, D. G.; Dean, J.; Ganger, G. R.; Joshi, G.; Kaminsky, M.; Kozuch, M.; Lipton, Z. C.; and Pillai, P. 2019. Accelerating Deep Learning by Focusing on the Biggest Losers. *CoRR*, abs/1910.00762.

Karkkainen, K.; and Joo, J. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1548–1558.

Katharopoulos, A.; and Fleuret, F. 2018. Not All Samples Are Created Equal: Deep Learning with Importance Sampling. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2525–2534. PMLR.

Khandani, A. E.; Kim, A. J.; and Lo, A. W. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11): 2767 – 2787.

Kumar, N.; Berg, A. C.; Belhumeur, P. N.; and Nayar, S. K. 2009. Attribute and simile classifiers for face verification. In *2009 IEEE 12th International Conference on Computer Vision*, 365–372.

Liu, C.; Zhuo, J.; Cheng, P.; Zhang, R.; and Zhu, J. 2019. Understanding and Accelerating Particle-Based Variational Inference. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 4082–4092. PMLR.

Liu, Q.; and Wang, D. 2016. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. In *Advances in Neural Information Processing Systems*, volume 29.

Loshchilov, I.; and Hutter, F. 2015. Online Batch Selection for Faster Training of Neural Networks. *CoRR*, abs/1511.06343.

Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; and Zemel, R. 2015. The Variational Fair Autoencoder. *arXiv e-prints*, arXiv:1511.00830.

Lowy, A.; Baharlouei, S.; Pavan, R.; Razaviyayn, M.; and Beirami, A. 2022. A Stochastic Optimization Framework for Fair Risk Minimization. *Transactions on Machine Learning Research*. Expert Certification.

Lum, K.; and Johndrow, J. 2016. A statistical framework for fair predictive algorithms. *arXiv e-prints*, arXiv:1610.08077.

Ren, M.; Zeng, W.; Yang, B.; and Urtasun, R. 2018. Learning to Reweight Examples for Robust Deep Learning. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 4334–4343. PMLR.

Roh, Y.; Lee, K.; Whang, S.; and Suh, C. 2020. FR-Train: A Mutual Information-Based Approach to Fair and Robust Training. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 8147–8157. PMLR.

Roh, Y.; Lee, K.; Whang, S. E.; and Suh, C. 2021. Fair-Batch: Batch Selection for Model Fairness. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Shu, J.; Xie, Q.; Yi, L.; Zhao, Q.; Zhou, S.; Xu, Z.; and Meng, D. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32.

Tahir, A.; Cheng, L.; and Liu, H. 2023. Fairness through Aleatoric Uncertainty. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, 2372–2381. New York, NY, USA: Association for Computing Machinery.

Tiu, E.; Talius, E.; Patel, P.; Langlotz, C. P.; Ng, A. Y.; and Rajpurkar, P. 2022. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6: 1399–1406.

Wei, T.; Mei, B.; Lyu, J.; Zhang, R.; Zhou, F.; and Sun, Y. 2025. Personalized Bayesian Federated Learning with Wasserstein Barycenter Aggregation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Wick, M.; Tristan, J.-B.; et al. 2019. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32.

Wolf, L.; Hassner, T.; and Taigman, Y. 2011. Effective Unconstrained Face Recognition by Combining Multiple Descriptors and Learned Background Statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10): 1978–1990.

Xu, X.; Kim, M.; Lee, R.; Martinez, B.; and Hospedales, T. 2024. A Bayesian Approach to Data Point Selection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2015. Learning fair classifiers. *arXiv preprint arXiv:1507.05259*, 1(2).

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *International conference on machine learning*, 325–333. PMLR.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, 335–340. New York, NY, USA: Association for Computing Machinery.

Zhang, Z.; and Pfister, T. 2021. Learning Fast Sample Reweighting Without Reward Data. arXiv:2109.03216.

Zhang, Z.; Song, Y.; and Qi, H. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

A Theory: Full Assumptions, Lemmas, and Proofs

This appendix provides all technical details that underlie Sec. 5. Throughout we write $z = (\boldsymbol{\theta}, \mathbf{w})$, p_s for the group posterior, $p^\star$ for the barycenter, and D for the chosen discrepancy. We separate the analysis for $D \in \{W_2, \mathrm{MMD}_k, D_f\}$ and then present unified statements.

A.1 Assumptions

Assumption 1 (Loss regularity w.r.t. D) *For each $s \in S$, the group loss $\mathcal{L}_s : \Xi \to [0, \infty)$ is measurable and satisfies, for any distributions p, q on Ξ,*

$$\left| \mathbb{E}_p \mathcal{L}_s(z) - \mathbb{E}_q \mathcal{L}_s(z) \right| \leq C_s D(p, q),$$

with the following instantiations:

- **Wasserstein W_2:** *$\mathcal{L}_s$ is L_s–Lipschitz under the metric $d(\cdot, \cdot)$ inducing W_2; then $C_s = L_s$ by Kantorovich–Rubinstein type arguments one can prove W_1 then adapting to W_2.*
- **MMD:** *$\mathcal{L}_s \in \mathcal{H}_k$ with $\|\mathcal{L}_s\|_{\mathcal{H}_k} \leq C_s$; then $C_s = \|\mathcal{L}_s\|_{\mathcal{H}_k}$ and the MMD duality yields the bound.*
- **f-divergence:** *$\mathcal{L}_s \in [0, B_s]$; by Pinsker-type inequalities, $|\mathbb{E}_p f - \mathbb{E}_q f| \leq B_s \sqrt{2 c_f D_f(p \| q)}$ where c_f depends on the chosen f (e.g., $c_f = 1$ for KL divergence, $c_f = 2$ for JS divergence and $c_f = 1/(1+t)$ for χ^2 divergence).*

Remark 1 *The constants C_s make explicit how the divergence choice affects tightness: W_2 gives linear (but potentially expensive) bounds, MMD/IPMs offer linear bounds and are kernel-amenable, while f-divergences yield $\sqrt{\cdot}$–type bounds.*

Assumption 2 (Cross–group compatibility) *There exists $K < \infty$ such that $|\mathcal{L}_s(z) - \mathcal{L}_{s'}(z)| \leq K$ for all z and $s, s' \in S$.*

Padding for unequal N_s. Let $\mathcal{P}_s : \mathbb{R}^{P+N_s} \to \mathbb{R}^{P+\bar{N}}$ (with $\bar{N} = \max_s N_s$) be the zero-padding operator used in the main text. We use the following fact.

Proposition 1 (Divergence preservation under padding) *Let p_s, q_s be distributions on $\mathbb{R}^{P+N_s}$, and $\bar{p}_s = (\mathcal{P}_s)_\# p_s$, $\bar{q}_s = (\mathcal{P}_s)_\# q_s$ their pushforwards. Then:*

- *$W_2(\bar{p}_s, \bar{q}_s) = W_2(p_s, q_s)$;*
- *$\mathrm{MMD}_k(\bar{p}_s, \bar{q}_s) = \mathrm{MMD}_k(p_s, q_s)$ for translation–invariant kernels;*
- *$D_f(\bar{p}_s \| \bar{q}_s) = D_f(p_s \| q_s)$ for any f–divergence.*

Proof. [Proof Sketch for Proposition 1] For Wasserstein distance, the optimal transport plan between padded distributions can be constructed from the optimal plan between original distributions, preserving the transport cost due to the isometry property.

For MMD with translation-invariant kernels, the kernel evaluations depend only on distances, which are preserved under padding.

For f-divergences, recall that $D_f(p \| q) = \int f\left(\frac{dp}{dq}\right) dq$ when $p \ll q$. The padded distributions have densities that factorize as:

$$\bar{p}_s(z) = p_s(\mathcal{Q}_s(z)) \cdot \mathbf{1}_{\mathrm{supp}(\mathcal{P}_s)}(z) \tag{10}$$

where the indicator function ensures the measure is supported on the padded subspace. The density ratio is preserved:

$$\frac{d\bar{p}_s}{d\bar{q}_s}(z) = \frac{p_s(\mathcal{Q}_s(z))}{q_s(\mathcal{Q}_s(z))} = \frac{dp_s}{dq_s}(\mathcal{Q}_s(z)) \tag{11}$$

on the support of $\mathcal{P}_s$. Therefore:

$$D_f(\bar{p}_s \| \bar{q}_s) = \int_{\mathrm{supp}(\mathcal{P}_s)} f\left(\frac{d\bar{p}_s}{d\bar{q}_s}\right) d\bar{q}_s \tag{12}$$

$$= \int_{\mathbb{R}^{P+N_s}} f\left(\frac{dp_s}{dq_s}\right) dq_s = D_f(p_s \| q_s) \tag{13}$$

Examples of preserved f-divergences include:

- KL divergence: $f(t) = t \log t$
- JS divergence: $f(t) = t \log t - (t+1) \log \frac{t+1}{2}$
- χ^2 divergence: $f(t) = (t-1)^2$
- α-divergence: $f(t) = \frac{t^\alpha - \alpha t + \alpha - 1}{\alpha(\alpha-1)}$

$\square$

Hence all our bounds proved in the common (padded) space numerically equal their counterparts in each group's native space.

Assumptions. We use the following discrepancy-specific regularity.

(A1) (**Loss regularity**) For each s, the loss $\mathcal{L}_s$ is bounded and satisfies a D–Lipschitz–type condition: for all distributions p, q,

$$\left| \mathbb{E}_p \mathcal{L}_s(z) - \mathbb{E}_q \mathcal{L}_s(z) \right| \leq C_s D(p, q),$$

where C_s depends on the choice of D:

- $C_s = L_s$ if $D = W_2$ and $\mathcal{L}_s$ is L_s–Lipschitz;
- $C_s = \|\mathcal{L}_s\|_{\mathcal{H}_k}$ if $D = \mathrm{MMD}_k$ and $\mathcal{L}_s \in \mathcal{H}_k$;
- $C_s = B_s \sqrt{2 c_f}$ if $D = D_f$ and $\mathcal{L}_s \in [0, B_s]$ (Pinsker-type).

(A2) (**Cross–group compatibility**) There exists $K < \infty$ such that for any z and $s, s' \in S$, $|\mathcal{L}_s(z) - \mathcal{L}_{s'}(z)| \leq K$.

A.2 Discrepancy Transfer Bound (Theorem 1)

Proof. Let $\bar{R} = \sum_s \lambda_s R_s(\tilde{p}_s)$. Then

$$R(\tilde{p}^\star) - \bar{R} = \sum_s \lambda_s \left(\mathbb{E}_{\tilde{p}^\star} \mathcal{L}_s(z) - \mathbb{E}_{\tilde{p}_s} \mathcal{L}_s(z) \right).$$

By (A1) for each s,

$$\left| \mathbb{E}_{\tilde{p}^\star} \mathcal{L}_s(z) - \mathbb{E}_{\tilde{p}_s} \mathcal{L}_s(z) \right| \leq C_s D(\tilde{p}^\star, \tilde{p}_s),$$

and the claimed bound (7) follows by convexity of the absolute value and the triangle inequality.

The three concrete instantiations are immediate and routine from the three cases of Assumption 1. We present the proof for completeness.

Case A (Wasserstein): Define $\bar{R} = \sum_{s \in S} \lambda_s R(p_s)$ as the weighted average risk. Then:

$$R(\tilde{p}^\star) - \bar{R} = R(\tilde{p}^\star) - \sum_{s \in S} \lambda_s R_s(\tilde{p}_s) \tag{14}$$

$$= \sum_{s \in S} \lambda_s [R_s(\tilde{p}^\star) - R_s(\tilde{p}_s)] \tag{15}$$

$$= \sum_{s \in S} \lambda_s \left[\mathbb{E}_{z \sim \tilde{p}^\star}[\mathcal{L}_s(z)] - \mathbb{E}_{z \sim \tilde{p}_s}[\mathcal{L}_s(z)] \right] \tag{16}$$

Since $\mathcal{L}_s$ is L_s-Lipschitz, by the Kantorovich-Rubinstein duality:

$$|\mathbb{E}_{\tilde{p}^\star}[\mathcal{L}_s(z)] - \mathbb{E}_{\tilde{p}_s}[\mathcal{L}_s(z)]| \leq L_s \cdot W_2(\tilde{p}^\star, \tilde{p}_s) \tag{17}$$

Therefore:

$$\left| R(\tilde{p}^\star) - \bar{R} \right| \leq \sum_{s \in S} \lambda_s \left| \mathbb{E}_{\tilde{p}^\star}[\mathcal{L}_s(z)] - \mathbb{E}_{\tilde{p}_s}[\mathcal{L}_s(z)] \right| \tag{18}$$

$$\leq \sum_{s \in S} \lambda_s L_s \cdot W_2(\tilde{p}^\star, \tilde{p}_s) \tag{19}$$

Case B (MMD): For $\mathcal{L} \in \mathcal{H}_k$, using the reproducing property:

$$|\mathbb{E}_{\tilde{p}^\star}[\mathcal{L}_s(z)] - \mathbb{E}_{\tilde{p}_s}[\mathcal{L}_s(z)]| = |\langle \mathcal{L}_s, \mu_{\tilde{p}^\star} - \mu_{\tilde{p}_s} \rangle_{\mathcal{H}_k}| \tag{20}$$

$$\leq \|\mathcal{L}_s\|_{\mathcal{H}_k} \cdot \|\mu_{\tilde{p}^\star} - \mu_{\tilde{p}_s}\|_{\mathcal{H}_k} \tag{21}$$

$$= \|\mathcal{L}_s\|_{\mathcal{H}_k} \cdot \text{MMD}_k(\tilde{p}^\star, \tilde{p}_s) \tag{22}$$

Case C (f-divergence): Using Pinsker's inequality for KL divergence (similar bounds exist for other f-divergences):

$$\text{TV}(p_s, p^\star) \leq \sqrt{\frac{1}{2} D_{KL}(\tilde{p}_s \| \tilde{p}^\star)} \tag{23}$$

For bounded $\mathcal{L}_s$:

$$|\mathbb{E}_{\tilde{p}^\star}[\mathcal{L}_s(z)] - \mathbb{E}_{\tilde{p}_s}[\mathcal{L}_s(z)]| \leq 2B_s \cdot \text{TV}(\tilde{p}_s, p^\star) \tag{24}$$

$$\leq B_s \sqrt{2 D_{KL}(\tilde{p}_s \| \tilde{p}^\star)} \tag{25}$$

$\square$

A.3 Group Disparity Bound (Theorem 2)

Proof. For any s, s',

$$\mathbb{E}_{\tilde{p}_s} \mathcal{L}_s(z) - \mathbb{E}_{\tilde{p}_{s'}} \mathcal{L}_{s'}(z)$$
$$= \left(\mathbb{E}_{\tilde{p}_s} \mathcal{L}_s(z) - \mathbb{E}_{\tilde{p}^\star} \mathcal{L}_s(z) \right) + \left(\mathbb{E}_{\tilde{p}^\star} \mathcal{L}_s(z) - \mathbb{E}_{\tilde{p}^\star} \mathcal{L}_{s'}(z) \right)$$
$$+ \left(\mathbb{E}_{\tilde{p}^\star} \mathcal{L}_{s'}(z) - \mathbb{E}_{\tilde{p}_{s'}} \mathcal{L}_{s'}(z) \right).$$

The first and third terms are bounded via (A1) by $C_s D(\tilde{p}_s, \tilde{p}^\star)$ and $C_{s'} D(\tilde{p}_{s'}, \tilde{p}^\star)$. The middle term is bounded by K using (A2). Taking absolute values and further maximizing the middle term over s, s' and the sample path of $\tilde{p}^\star$ yields (8). $\square$

Now we formalize the condition for the elimination of $K_{\text{eff}}(t)$ in (8).

In order to eliminate $K_{\text{eff}}(t)$, we adopt the following *strong* assumption: the fairness-aware barycenter converges to a *single* parameter $z^\dagger$ that equalizes all group losses. This lets us turn the qualitative statement in Theorem 2 into an *asymptotically vanishing* bound with explicit rates that depend on the chosen discrepancy D.

Assumption 3 (Point–mass strong limit) *There exists $z^\dagger \in \Xi$ such that*

$$\mathcal{L}_s(z^\dagger) = \mathcal{L}_{s'}(z^\dagger) \qquad \forall s, s' \in S, \tag{26}$$

and the (empirical) barycenters produced by Fair-BADS satisfy

$$\tilde{p}_\star^{(t)} \xrightarrow{D} \delta_{z^\dagger} \quad and \quad D(\tilde{p}_s^{(t)}, \tilde{p}_\star^{(t)}) \to 0 \quad for \ all \ s \in S,$$

as $t \to \infty$.

We now prove that, under Assumption 3, the effective cross–group term in Theorem 2 *vanishes*. We present the result for the three divergences we consider (Wasserstein, MMD/IPM, and f-divergence). The Wasserstein case provides a *sup*-type (support-level) bound; for MMD and f-divergences, we obtain clean *expectation*-level bounds.[1]

A convenient empirical-particle inequality (Wasserstein). When $\tilde{p}_\star^{(t)}$ is represented by M equally weighted particles $\{z_\star^{(m)}\}_{m=1}^M$,

$$\max_{1 \leq m \leq M} \|z_\star^{(m)} - z^\dagger\| \leq \sqrt{M} W_2(\tilde{p}_\star^{(t)}, \delta_{z^\dagger}), \tag{27}$$

because $W_2^2 = \frac{1}{M} \sum_{m=1}^M \|z_\star^{(m)} - z^\dagger\|^2$ and hence $\max_m \|z_\star^{(m)} - z^\dagger\| \leq \sqrt{M} \left(\frac{1}{M} \sum_m \|z_\star^{(m)} - z^\dagger\|^2 \right)^{1/2}$.

Theorem 3 (Vanishing effective cross–group term) *Assume* (A1) *and Assumption 3. Define*

$$K_{\text{eff}}(t) := \sup_{z \in \text{supp}(\tilde{p}_\star^{(t)})} \max_{s, s' \in S} \left| \mathcal{L}_s(z) - \mathcal{L}_{s'}(z) \right|.$$

Then $K_{\text{eff}}(t) \to 0$ as $t \to \infty$. More precisely:

1. *(a) **Wasserstein case.** Suppose $D = W_2$ and each $\mathcal{L}_s$ is L_s-Lipschitz. Then for every t,*

$$K_{\text{eff}}(t) \leq 2 \max_{s \in S} L_s \sqrt{M} W_2(\tilde{p}_\star^{(t)}, \delta_{z^\dagger}), \tag{28}$$

and hence $K_{\text{eff}}(t) \to 0$ as soon as $W_2(\tilde{p}_\star^{(t)}, \delta_{z^\dagger}) \to 0$.

2. *(b) **MMD/IPM case(expectation level).** Suppose $D = \text{MMD}_k$, each $\mathcal{L}_s \in \mathcal{H}_k$ with $\|\mathcal{L}_s\|_{\mathcal{H}_k} \leq C_s$. Then*

$$\max_{s, s' \in S} \left| \mathbb{E}_{\tilde{p}^{(t)}_\star}[\mathcal{L}_s] - \mathbb{E}_{\tilde{p}^{(t)}_\star}[\mathcal{L}_{s'}] \right| \leq 2 C_{\max} \text{MMD}_k(\tilde{p}_\star^{(t)}, \delta_{z^\dagger}) \tag{29}$$

where $C_{\max} = \max_{s \in S} C_s$, and thus the expected cross-group gap vanishes as $\text{MMD}_k(\tilde{p}_\star^{(t)}, \delta_{z^\dagger}) \to 0$.

---

[1]One can turn the expectation bounds for MMD/f-divergence into support-type statements under additional smoothness/uniform-continuity assumptions; we do not pursue these technicalities here.

3. **(c) f-divergence case (expectation level).** *Suppose $D = D_f$, each $\mathcal{L}_s \in [0, B_s]$, and let c_f be the Pinsker-type constant for D_f (e.g., $c_f = 1$ for KL). Then*

$$\max_{s,s' \in S} \left| \mathbb{E}_{\tilde{p}_\star^{(t)}}[\mathcal{L}_s] - \mathbb{E}_{\tilde{p}_\star^{(t)}}[\mathcal{L}_{s'}] \right| \leq 2 \max_{s \in S} B_s \sqrt{2 c_f D_f(\tilde{p}_\star^{(t)} \| \delta_{z^\dagger})}, \tag{30}$$

and hence the expected *cross-group gap vanishes whenever $D_f(\tilde{p}_\star^{(t)} \| \delta_{z^\dagger}) \to 0$.*

A.4 Fairness–Aware SVGD: ELBO Improvement Bound

Fairness–aware SVGD improves the ELBO while aligning to $p^\star$. We also analyze one SVGD step for a fixed group s. Let $q_s^{(\ell)}$ be the empirical particle distribution and consider the transport $t(z) = z + \varepsilon \phi_{\text{fair}}(z)$ with

$$\phi_{\text{fair}}(z) = \mathbb{E}_{z' \sim q_s^{(\ell)}} \left[k(z', z) \nabla_{z'} \log p_{\text{fair}}(z') + \nabla_{z'} k(z', z) \right],$$

$$\log p_{\text{fair}} = \log p_s + \log p^\star.$$

Let $F(q) = \text{ELBO}(q) = -\text{KL}(q \| p_s) + \text{const}$. Following standard techniques, a first–order Taylor expansion plus a change of variables argument yields

$$F(q_s^{(\ell+1)}) - F(q_s^{(\ell)}) = \varepsilon \mathbb{E}_{q_s^{(\ell)}} \left[\text{tr}\big(\mathcal{A}_{p_s} \phi_{\text{fair}}(z)\big) \right] + O(\varepsilon^2),$$

where $\mathcal{A}_p \phi = \nabla \log p^\top \phi + \nabla \cdot \phi$ is the Stein operator. Let $\phi_s^\star$ be the KSD-optimal direction using $\log p_s$ alone. Decompose

$$\phi_{\text{fair}} = \phi_s^\star + \Delta, \quad \Delta(z) = \mathbb{E}_{z' \sim q_s^{(\ell)}} \left[k(z', z) \nabla_{z'} \log p^\star(z') \right].$$

By Cauchy–Schwarz in the RKHS,

$$\left| \mathbb{E}_{q_s^{(\ell)}}[\text{tr}(\mathcal{A}_{p_s} \Delta(z))] \right|$$
$$\leq \left\| \nabla \log p_s - \nabla \log p^\star \right\|_{L^2(q_s^{(\ell)})} \|k\|_{\mathcal{H}} \cdot \text{KSD}(q_s^{(\ell)}, p_s), \tag{31}$$

which we denote by $C_{\text{KSD}} \cdot \text{KSD}(q_s^{(\ell)}, p_s)$. Since $\phi_s^\star$ maximizes the linear functional defining the KSD,

$$\mathbb{E}_{q_s^{(\ell)}} \left[\text{tr}(\mathcal{A}_{p_s} \phi_s^\star(z)) \right] = \text{KSD}(q_s^{(\ell)}, p_s).$$

Therefore, for sufficiently small ε (dropping $O(\varepsilon^2)$) we have

$$F(q_s^{(\ell+1)}) - F(q_s^{(\ell)}) \geq \varepsilon \big(1 - C_{\text{KSD}}\big) \text{KSD}(q_s^{(\ell)}, p_s)$$

where $C_{\text{KSD}} = \left\| \nabla \log p_s - \nabla \log p^\star \right\|_{L^2(q_s^{(\ell)})} \|k\|_{\mathcal{H}}$ and the condition $C_{\text{KSD}} < 1$ can be ensured in practice by (i) annealing the strength of the barycenter term early on; (ii) adaptively tuning the kernel bandwidth; or (iii) updating $p^\star$ frequently so it stays close to each p_s.

Remark 2 *To explicitly control the interaction between posterior inference and fairness alignment, one can replace $\log p_{fair}(z) = \log p_s(z) + \log p^\star(z)$ by $\log p_{fair}^{(\lambda)}(z) = \log p_s(z) + \lambda \log p^\star(z)$, which leads to the bound*

$$F\big(q_s^{(\ell+1)}\big) - F\big(q_s^{(\ell)}\big) \gtrsim \epsilon \big(1 - \lambda C_{\text{KSD}}\big) \text{KSD}(q_s^{(\ell)}, p_s),$$

for a constant C_{KSD} that depends on the (squared–)RKHS norm of the kernel and the score misalignment $\|\nabla \log \tilde{p}_s - \nabla \log p^\star\|_{L^2(q_s^{(\ell)})}$. Hence, by shrinking λ we can always guarantee a strictly positive ascent step (monotone ELBO increase) even when the fairness score term is temporarily antagonistic to the likelihood term. When the two score fields agree ($C_{\text{KSD}} \downarrow 0$), the fairness term no longer slows the ELBO ascent.

B Algorithm

We provide the pseudocode of the proposed method in Algorithm 1 to clearly outline the key steps of Fair-BADS. The algorithm maintains group-specific particle approximations of the joint posterior over model parameters θ and sample weights $\mathbf{w}$. For each group, particles are updated via SVGD, where the update direction is informed by both the group posterior and a shared central distribution that encourages fairness across groups. The central distribution is iteratively computed over group-specific particles and serves as a soft alignment target. At each iteration, particles are updated by combining gradients from the group likelihood and the central prior, resulting in a fairness-aware inference process. This formulation ensures that each group's learning signal is preserved while ensuring inter-group consistency, which is critical for achieving fair data selection.

Algorithm 1: Fair-BADS

Require: Training data $\{(\mathbf{x}_i, y_i, s_i)\}_{i=1}^N$, meta data $\mathcal{D}_m$, group indices S, particle count M, learning rate ϵ, discrepancy measure D.

1: Initialize group-specific particles $\{z_s^{(m)}\}_{m=1}^M$ for each $s \in S$.
2: Initialize global central particles $\{\bar{z}^{(m)}\}_{m=1}^M$.
3: **for** each training iteration **do**
4:     **for** each group $s \in S$ **do**
5:         Extract $\mathcal{D}_t^s \subset \mathcal{D}_t$.
6:         **for** each particle $z_s^{(m)} = (\theta_s^{(m)}, \mathbf{w}_s^{(m)})$ **do**
7:             Compute group posterior gradient $\nabla_z \log p_s(z)\big|_{z=z_s^{(m)}}$ from Eq. (4).
8:         **end for**
9:         Estimate gradient $\nabla_z \log p^\star(z)\big|_{z=z_s^{(m)}}$ from central distribution.
10:         Combine gradients: $\nabla_z \log p_{\text{fair}}(z) = \nabla_z \log \tilde{p}_s(z) + \nabla_z \log \tilde{p}^\star(z)$.
11:         Update $\{z_s^{(m)}\}_{m=1}^M$ via SVGD.
12:     **end for**
13:     Compute central distribution $\bar{\mathbf{Z}}$ across all groups as in Section 4.2.
14:     Update KDE reference using central particles.
15: **end for**
