Fairness-Aware Active Online Learning with Changing Environments

Sadaf MD Halim
Department of Computer Science
The University of Texas at Dallas
Richardson, Texas
sxh190015@utdallas.edu

Chen Zhao
Department of Computer Science
Baylor University
Waco, Texas
chen_zhao@baylor.edu

Xintao Wu
Department of Electrical Eng. & Comp. Sci.
University of Arkansas
Fayetteville, Arkansas
xintaowu@uark.edu

Latifur Khan
Department of Computer Science
The University of Texas at Dallas
Richardson, Texas
lkhan@utdallas.edu

Christan Earl Grant
Dept. of Computer & Info. Sci. & Eng.
University of Florida
Gainesville, Florida
christan@ufl.edu

Feng Chen
Department of Computer Science
The University of Texas at Dallas
Richardson, Texas
feng.chen@utdallas.edu

***Abstract*—In real-world applications, data-driven classifiers often grapple with a three-pronged challenge: data arrives in a continuous stream, most data in the wild are often unlabeled, and there is a critical need to maintain fairness in predictions across different sub-groups. Existing methods falter when addressing all these three factors concurrently. This work tackles this challenge by addressing a novel paradigm: Fairness-Aware Active Online Learning. We introduce a simple yet effective approach – FACTION, which actively selects the most crucial data points for labeling, going beyond traditional methods by considering both model uncertainty (epistemic uncertainty) and a newly introduced fairness notion derived from this very uncertainty. Additionally, FACTION leverages a system adept at identifying out-of-distribution samples within online learning environments. Extensive evaluations on real-world datasets, coupled with theoretical analysis, demonstrate FACTION's effectiveness in handling this complex challenge. Our model demonstrably outperforms relevant baselines adapted for this new setting.**

***Index Terms*—Group Fairness, Data Selection, Active Learning, Online Learning, Epistemic Uncertainty**

I. Introduction

Data-driven classifiers face multifaceted challenges in the real world. In dynamic environments, such as online learning systems, models must continuously adapt to evolving data streams while addressing distributional shifts and operating under resource constraints. Simultaneously, the prevalence of unlabeled data in the wild necessitates efficient strategies for data labeling. These challenges are further amplified by the growing societal emphasis on fairness, which requires equitable outcomes across diverse demographic groups. Addressing these issues in tandem is essential for developing robust and ethical machine learning systems.

Consider the example of a pedestrian detection system deployed in real-world urban settings. Live camera feeds generate continuous streams of data, with characteristics that vary depending on factors such as time, location, and population demographics. For instance, cameras near schools may capture a higher proportion of children during certain hours and adults at other times. This variability introduces significant distribution shifts. Moreover, the raw data is largely unlabeled, making manual annotation of all images both time and resource-intensive. Compounding these challenges is the need to ensure fairness in decision-making, as pedestrian detection models often exhibit biases against specific demographic groups, such as age and gender [6], [7]. This creates an urgent need for systems that can adapt to new data, select the most informative samples for labeling, and do so while mitigating bias.

Prior research addresses subsets of these challenges but rarely integrates them holistically. Fairness-aware learning frameworks [5] aim to reduce disparities in outcomes across demographic groups, yet often assume static datasets. Active Learning (AL) algorithms [1] prioritize the most informative samples for labeling to maximize model performance but overlook fairness considerations. Online learning frameworks [2], [4] dynamically adapt models to new data but are not inherently fairness-aware. Active Online Learning [2], [47] combines aspects of online learning and active learning but does not address fairness. To fill this critical gap, we propose a unified system called FACTION (**F**air **Acti**ve **On**line Learning) that integrates these concerns into a cohesive solution.

FACTION leverages uncertainty quantification to guide its active learning process. Notably, epistemic uncertainty is a form of uncertainty [9] that arises from a lack of relevant training data. This is crucial for active learning as it identifies areas where acquiring labels can significantly improve model performance. Importantly, samples that are out-of-distribution (OOD) with respect to the current model also often exhibit high epistemic uncertainty [45], [46], making it particularly effective in dynamic environments with distribution shifts. By focusing on epistemic uncertainty and leveraging recent advances in disentangling uncertainty types [45], [46], FACTION acquires highly informative samples to enhance performance.

Furthermore, FACTION incorporates a novel fairness metric into the sample selection process. This metric estimates the

fairness of a data sample by employing a fairness-sensitive density estimator based on features learned from a deep neural network. FACTION identifies discrepancies in the density distribution across different sensitive attributes within the same class. Extending the traditional active learning paradigm, which prioritizes the *most uncertain* samples, FACTION incorporates fairness by selecting the *most unfair* samples. A fairness constraint is incorporated into the loss function to further regularize the mode. Since we are concerned with treating diverse demographic groups equitably, FACTION focuses on group fairness. A detailed explanation of the fairness measures is provided in Section IV. Additionally, we provide theoretical bounds for FACTION, establishing guarantees on query complexity, regret, and fairness violation under certain key assumptions: (1) convexity of the domain, loss, and fairness functions, (2) Lipschitz continuity, and (3) bounded gradients. These are detailed in Section IV-G.

Challenges. Some key challenges for FACTION include:

- **Balancing Fairness and Performance**: Achieving optimal model performance while ensuring fairness is challenging, especially with evolving, unlabeled data.
- **Strategic Label Acquisition**: Fairness considerations require the AL process to balance data informativeness with fairness, making the label acquisition strategy complex.
- **Real-Time Adaptation**: Adapting to both shifting data distributions and evolving fairness requirements in real-time demands algorithms capable of quickly updating their strategies to maintain both performance and fairness.

Contributions. Our contributions can be summarized as:

- **The Fairness-Aware Active Online Learning Paradigm**: We introduce the practical and novel paradigm of Fairness-Aware Active Online Learning, that addresses real-world challenges.
- **Integrating Epistemic Uncertainty and Fairness Notions**: We create a new AL query function that integrates epistemic uncertainty and a new fairness notion.
- **An Effective and Practical System:** We introduce FACTION[1], a simple yet effective approach that involves selecting the most unfair samples and regularizing for fairness in the loss function.

By bridging fairness, active learning, and online adaptivity, FACTION addresses a key challenge in data-driven systems. Through a novel fairness-aware sample selection approach that remains robust to distribution shifts, FACTION contributes to the body of work on responsible and efficient data engineering.

II. Related Work

Fairness-Aware Active Online Learning. To our knowledge, prior work has not simultaneously addressed active learning, online learning, and fairness-awareness. However, significant progress has been made in various intersections of these areas. Traditional Active Learning [1] focuses on maximizing model accuracy by selecting the most informative labeled training instances. This often involves choosing samples with the highest uncertainty, such as those with the lowest model confidence [43]. Techniques like margin sampling, which compares the probabilities of the top two classes [42], and information entropy-based measures [41] are commonly employed. Recently, measures of epistemic uncertainty [44], [46] have also become effective heuristics for selecting unlabeled samples.

On the other hand, fairness in machine learning is another well-explored area of research, with various techniques for fair pre-processing [56], [55] and fairer feature engineering of data [54], as well as various fair learning algorithms [53], [52]. However, we are more specifically interested in fairness in more constrained settings, such as with dynamic and unlabeled data. Recently, incorporating fairness into Active Learning has become a significant research focus. Fair Active Learning (FAL) [33] integrates entropy and fairness measures to guide sample selection. This approach uses a metric called Expected Fairness to assess the impact of each sample on the model's fairness if included in the labeled set. This helps identify samples that most enhance model fairness, showing promising results. Another approach, FAL-CUR [34], combines Fair Clustering with uncertainty sampling, selecting the most uncertain and representative samples from clusters for labeling. D-FA^{2}L [12] introduces the idea of decoupled models for better fairness-aware active learning, leveraging disagreements between models to identify the most promising samples. Research has also explored improving minimax fairness in active learning [35], [37], aiming to minimize the largest group's error rate, which differs from traditional group fairness metrics like Demographic/Statistical Parity (DP) [32] and Equalized Odds (EO) [31], [51]. While minimax fairness is crucial in high-risk applications like healthcare, our work focuses on traditional group fairness concerns such as DP and EO, aiming to treat all population sub-groups equally.

Online Learning, extensively studied in prior research [3], [4], aims to develop an optimal model by adapting to new data in batches or individually, focusing on minimizing regret. A key challenge is managing distribution shifts in both data and label distributions [29], [30]. Active Online Learning adds complexity by handling unlabeled data, requiring the learner to selectively acquire labels within a specified budget and time frame. Algorithms must decide whether to query the label of each incoming sample, often using a probabilistic approach. For instance, QuFUR [2] estimates the uncertainty of each sample to determine the query probability, effectively managing hidden domain shifts. Similarly, another approach [47] uses a Bernoulli random variable to decide on querying, with the probability based on the prediction margin. Prior work has explored *fair* online learning using Blackwell approachability [59] to enforce demographic parity and group-wise calibration. Other approaches, such as fairness-aware online allocation in continuous-time [60], aim to balance fairness and efficiency, while FABBOO [61] addresses class imbalance and cumulative discrimination in streaming data. These methods offer theoretical guarantees for fairness in online learning. However, none incorporate an active strategy for acquiring unlabeled data, limiting their effectiveness in our paradigm.

[1]https://github.com/acesadaf/FACTION

Uncertainty. Quantifying uncertainty is a significant challenge in research. State-of-the-art methods to address this issue include Deep Ensembles [20], [23], feature space density characterization [13], [15], and Bayesian methods [44]. Epistemic and Aleatoric uncertainties are two fundamental types of uncertainty in predictive modeling [9], [21], [22], [46]. Epistemic uncertainty, or model uncertainty, arises from our lack of knowledge about the true underlying model that generated the data. It is associated with uncertainty in the model parameters and can be reduced with more data. Conversely, aleatoric uncertainty is inherent in the variability of the observed data, stemming from inherent stochasticity in the underlying processes, and cannot be eliminated with additional data.

III. Preliminaries

A. Group Fairness

Group fairness criteria usually focus on relationships between sensitive attributes and model outputs [24], [25]. This study considers a binary sensitive attribute (e.g., young vs. old), but can extend to multi-valued sensitive attributes. Given a data space $\mathcal{P} = \mathcal{X} \times \mathcal{S} \times \mathcal{Y} \times \mathcal{E}$, where $\mathcal{X} \in \mathbb{R}^d$ is the input feature space, $\mathcal{S} \in \{-1, 1\}$ is the sensitive space, $\mathcal{Y} \in \{0, 1\}$ is the binary output space, and $\mathcal{E} \in \mathbb{N}$ is the environment space. For a task $\mathcal{D} = \{(\mathbf{x}_i, s_i, y_i, e_i)\}_{i=1}^n \in \mathcal{P}$ in environment $e \in \mathcal{E}$, a detailed measure to ensure group fairness in class label prediction involves designing fair classifiers. This is done by regulating notions of fairness between sensitive subgroups $\{s_i = 1\}_{i=1}^{n_1}$ and $\{s_i = -1\}_{i=1}^{n_{-1}}$ where $n_1 + n_{-1} = n$, using methods like demographic parity [24], [27].

Definition 1 (Notions of Fairness [24], [27], [28])**.** *A classifier $h : \mathbb{R}^d \times \Theta \to \mathbb{R}$ is fair when its predictions, $\mathbf{y} = \{\hat{y}_i\}_{i=1}^n$, are independent of the sensitive attribute $\mathbf{s} = \{s_i\}_{i=1}^n$. To get rid of the indicator function and relax the exact values, a linear approximated form of the difference between sensitive subgroups is defined [27],*

$$v(\mathcal{D}, \boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x},s,y,e)\sim\mathcal{P}}\Big[\frac{1}{\hat{p}_1(1-\hat{p}_1)}\Big(\frac{s+1}{2} - \hat{p}_1\Big)h(\mathbf{x}, \boldsymbol{\theta})\Big] \quad (1)$$

where $\hat{p}_1$ is an empirical estimate of pr_1. pr_1 is the proportion of samples in group $s = 1$ and correspondingly $1 - pr_1$ is the proportion of samples in group $s = -1$.

In eq. (1), when $\hat{p}_1 = \mathbb{P}_{(\mathbf{x},s,y,e)\in\mathcal{P}}(s = 1)$, the fairness notion $v(\mathcal{D}, \boldsymbol{\theta})$ is defined as the difference of demographic parity (DDP). Similarly, when $\hat{p}_1 = \mathbb{P}_{(\mathbf{x},s,y,e)\in\mathcal{P}}(y = 1, s = 1)$, $v(\mathcal{D}, \boldsymbol{\theta})$ is defined as the difference of equality of opportunity (DEO). Thus, parameters $\boldsymbol{\theta}$ are considered feasible if they strictly satisfy the fairness constraint $v(\mathcal{D}, \boldsymbol{\theta}) = 0$.

B. Active Learning (AL)

Let $\mathcal{D}^{\mathcal{U}}$ be a pool of unlabeled data, and $\mathcal{D}^{labeled}$ be a set of initially labeled instances. The goal of AL [1] is to iteratively select a subset $\mathcal{D}^{query} \subset \mathcal{D}^{\mathcal{U}}$ for annotation by an oracle. The learner has a budget $\mathcal{B}$, for the maximum number of labels it can query. Samples are typically selected by some measure of the model's uncertainty about an unlabeled sample, such as entropy or margin of the predicted probabilities [41]–[43]. An AL loop consists of these steps:

1) Train the model on $\mathcal{D}^{labeled}$.
2) Evaluate uncertainty on each data point in $\mathcal{D}^{\mathcal{U}}$.
3) Select $\mathcal{D}^{query}$ based on the most informative instances.
4) Query the oracle to obtain labels for $\mathcal{D}^{query}$.
5) Update $\mathcal{D}^{labeled}$ with the newly acquired labels.
6) Repeat Steps (1) to (5) until $|\mathcal{D}^{labeled}| = \mathcal{B}$.

For instance, in a digit recognition task with 10,000 handwritten samples ($\mathcal{D}^{\mathcal{U}} = 10000$) and a budget of 100 ($\mathcal{B} = 100$), Active Learning (AL) selects the 100 most uncertain samples (e.g., ambiguous strokes, poor contrast) for annotation.

C. Active Online Learning

Active online learning merges the active and online learning paradigms, processing unlabeled sequential data batches ($\mathcal{D}_t^{\mathcal{U}}$, or tasks) at each time step t. The learner uses an active selection strategy with a fixed budget $\mathcal{B}$ per task. The learner, faced with a loss function $f_t : \mathbb{R}^d \times \Theta \to \mathbb{R}$ at each time step, aims to adapt its model parameters to new information in real-time. The learner queries labels from $\mathcal{D}_t^{\mathcal{U}}$ until the budget is exhausted to form $\mathcal{D}_t^{labeled}$. This is added to the existing labeled task pool, $\mathcal{D}_t = \{\mathcal{D}_i^{labeled}\}_{i=1}^{t-1}$, and the learner learns from these samples and proceeds to the next task. The goal is to determine a sequence of model parameters $\{\boldsymbol{\theta}_t\}_{t=1}^T$ that maximizes performance within the given budget. For example, in a traffic monitoring system, active online learning prioritizes labeling uncertain pedestrian crossing events, such as cases with obstructed views, while continuously adapting to changing daylight and traffic patterns.

IV. Methodology

A. Fair Active Online Learning

We consider a sequential setting where a learner encounters tasks $\{\mathcal{D}_t^{\mathcal{U}}\}_{t=1}^T$ over time $t \in [T]$, with each task arriving unlabeled. Using an active selection strategy, the learner selects $\mathcal{B}$ samples from $\mathcal{D}_t^{\mathcal{U}}$ for labeling, forming $\mathcal{D}_t^{labeled}$. At each time step, the model parameters $\boldsymbol{\theta}_t$ are updated using the labeled task pool $\mathcal{D}_t = \{\mathcal{D}_i^{labeled}\}_{i=1}^{t-1}$, ensuring both performance and *fairness* i.e. the fair constraint $v_t(\mathcal{D}_t, \boldsymbol{\theta}_t) = 0$ is strictly satisfied and the loss $f_t(\mathcal{D}_t, \boldsymbol{\theta}_t)$ is minimized. For example, in pedestrian detection systems at busy intersections, the learner must adapt to changing weather and lighting conditions while ensuring consistent detection across age groups and demographics. The overall protocol is:

1) The learner selects parameters $\boldsymbol{\theta}_t$, using the existing labeled task pool, $\mathcal{D}_t$.
2) The learner receives the loss function f_t and fairness metric v_t.
3) The learner incurs an instantaneous loss $f_t(\mathcal{D}_t^{\mathcal{U}}, \boldsymbol{\theta}_t)$ and fairness estimation $v_t(\mathcal{D}_t^{\mathcal{U}}, \boldsymbol{\theta}_t)$.
4) The learner uses an intelligent querying strategy to acquire labels for the current task $\mathcal{D}_t^{\mathcal{U}}$, to get $\mathcal{D}_t^{labeled}$ of size $\mathcal{B}$, which is added to $\mathcal{D}_t$.
5) Advance to the next time step.

Regret. Assume $\mathcal{H} = \{h : \mathcal{X} \to [0, 1]\}$ is a fair hypothesis class. Consider a setting where $y_i = h^*(x_i) + \mathcal{E}_i$ for some fair classifier $h^* \in \mathcal{H}$, where $\mathcal{E}_i$ refers to a random noise $\mathcal{E}_t$, which are independent sub-Gaussian random variables with a mean of 0 and and a variance of η^2. Each task in $\{\mathcal{D}_t^{\mathcal{U}}\}_{t=1}^T$ provides samples. For the task at time t, we have, $\mathcal{D}_t^{\mathcal{U}} = \{x_i\}_{i=1}^{|\mathcal{D}_t^{\mathcal{U}}|}$. When a new task $\mathcal{D}_t^{\mathcal{U}}$ arrives, the learner incurs an instantaneous loss $f_t(\mathcal{D}_t^{\mathcal{U}}, \boldsymbol{\theta}_t)$. Assume that $f_t^*(\mathcal{D}_t^{\mathcal{U}})$ represents the best loss possible at each task t using the most optimal parameters for task t. We define the regret as:

$$R = \sum\nolimits_{t=1}^{T} (f_t(\mathcal{D}_t^{\mathcal{U}}, \boldsymbol{\theta}_t) - f_t^*(\mathcal{D}_t^{\mathcal{U}})) \tag{2}$$

While $\mathcal{D}_t^{\mathcal{U}}$ is unlabeled for the learner, we assume that labels are visible to the loss, f_t, only to compute and track the regret. The learner's objective at each time step t, is to query the unlabeled incoming dataset $\mathcal{D}_t^{\mathcal{U}}$, for labels within a budget $\mathcal{B}$, to minimize the overall regret R, while subject to fairness constraints $v_t(\mathcal{D}_t, \boldsymbol{\theta}_t) = 0$.

B. Fair Epistemic Uncertainty

As discussed in Sections I and II, Epistemic Uncertainty is an effective heuristic in AL and in OOD detection when data distributions shift. We first disentangle uncertainty into Epistemic and Aleatoric Uncertainty, removing the latter as it results from data ambiguity. Next, we derive a fairness metric based on epistemic uncertainty. To first estimate this uncertainty, we employ a deterministic neural network with a carefully regularized feature space, to extract feature vectors. Inspired by prior work [46], for our image experiments, we train a ResNet-18 architecture with spectral normalization, which prevents feature collapse in feature spaces by ensuring smoothness and sensitivity [19]. For tabular data, we use a simple MLP. The extracted feature vector is given by $z = r(x, \theta)$, where r uses weights θ to extract the feature representation of x at the final convolutional layer (for the ResNet) or a linear layer (for the MLP). Both architectures are detailed in Section V-A3.

In previous work, a density estimator, $G(z)$, is created using labeled training data. It uses class-specific components, $\{G_y(z)\}_{y=1}^C$, where C is the number of classes. Each component, $G_y(z)$, is fitted using feature vectors derived from samples belonging to that class. Here, $G_y(z)$ estimates the density of a test sample's feature representation, *z with respect to class y*, i.e., $g(z|y)$. The overall density is calculated by summing over the class-specific densities, weighted by the prior class probabilities. Prior work [46] showed that this density measures epistemic uncertainty, while aleatoric uncertainty is computed from softmax outputs. Epistemic uncertainty is low for familiar test samples and high for unseen ones. Consider a loan approval system: if the model has mostly seen young applicants, it would likely show high epistemic uncertainty (low density) when evaluating applications from older individuals. A density estimator, fitted with the mean and covariance of previously seen samples, will assign high density to similar (low uncertainty) samples and low density to unfamiliar ones. While previous work focuses on uncertainty, we address both uncertainty and fairness. Thus, we propose a new formulation for the density estimator:

1) For each class label y, and sensitive attribute value s, create a component $G_{y,s}(z)$ in the density estimator. If there are C classes and S possible sensitive values, this creates $C \times S$ components. The entire set of components is thus given by: $\{G_{y,s}(z) \mid \text{for all } y \in C \text{ and for all } s \in S\}$
2) Each component $G_{y,s}(z)$ returns the density of a feature vector z with respect to *both* the sensitive attribute s and the class label y. That is, $G_{y,s}(z)$ returns $g(z|y, s)$
3) The overall density is now calculated as:

$$g(z) = \sum\nolimits_y \sum\nolimits_s g(z \mid y, s) \cdot p(y, s) \tag{3}$$

where $p(y, s) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(y_i = y \,\text{and}\, s_i = s)$. Here, N is the total number of samples in the dataset, and $\mathbb{I}$ is the indicator function, returning 1 if the conditions are true and 0 otherwise. The overall density, $g(z)$, for a feature vector z measures the model's epistemic uncertainty regarding z. Here, $g(z)$ is calculated by summing over all components created from combinations of classes and sensitive attributes, as opposed to only classes. Low density implies high epistemic uncertainty regarding a sample.

The density estimator can be built using techniques like Gaussian Processes [17], Normalizing Flows [16], or Gaussian Discriminant Analysis (GDA) [18]. Like prior work [18], [46], we use GDA to create a Gaussian Mixture Model (GMM) as the density estimator $G(z)$. A Gaussian mixture component is created for each combination of class label and sensitive attribute by computing the mean and covariance from the feature vectors of all labeled training samples with the corresponding class label and sensitive attribute. The feature vector, z, is extracted from a sample x.

Our formulation for density estimation creates components based on *combinations of class labels and sensitive attributes*. This formulation helps create a new metric for assessing how *fair* samples are. We use densities with respect to each class/sensitive component to derive this measure. We propose that if a data sample is *fair*, then the difference in its calculated density with respect to two components having the *same class* but *different sensitive attribute values* should be low. Assuming two classes (binary classification), and two possible values for the sensitive attribute, we have:

$$\triangle g_0(z) = |g(z \mid y = 0, s = 1) - g(z \mid y = 0, s = -1)| \tag{4}$$
$$\triangle g_1(z) = |g(z \mid y = 1, s = 1) - g(z \mid y = 1, s = -1)| \tag{5}$$

Here, for a feature vector z derived from a sample, $\triangle g_0(z)$ is the absolute difference between the densities of two components with the same class $(y = 0)$ but different sensitive attributes. Similarly, $\triangle g_1(z)$ is computed for class $y = 1$. For fairness, if a sample belongs to class $y = 0$, it should have a low value for $\triangle g_0(z)$, indicating that it is equally likely to belong to either density component within this class, irrespective of the sensitive attribute. This implies that the

sample's sensitive attribute is independent of its class label. Conversely, unfair samples will have large values for $\triangle g_0(z)$ or $\triangle g_1(z)$. For example, for a sample in class $y = 0$, a large $\triangle g_0(z)$ suggests a significant preference for one component over another, i.e., for one sensitive group over the other.

This analysis can extend to a multi-class paradigm by using C different $\triangle g_c(z)$ values, $\{\triangle g_c(z)\}_{c=1}^{C}$, where C is the number of classes. However, we focus on binary classification with binary sensitive attributes, leaving extensions to future work. To conclude, these $\{\triangle g_c\}_{c=1}^{C}$ values for each class form our fairness notion based on epistemic uncertainty, where higher values are more unfair. Specifically, for feature vectors z_1 and z_2 from samples in class c:

$$\Delta g_c(z_1) > \Delta g_c(z_2) \implies z_1 \text{ is more unfair than } z_2$$

Fair Epistemic Uncertainty compared to other Group Fairness Metrics. Traditional group fairness metrics like Difference of Demographic Parity (DDP) or Equalized Odds Difference (EOD) [26] rely on correlations between predicted outcomes and sensitive attributes. In contrast, Fair Epistemic Uncertainty estimates fairness from the *feature space* by density estimation, directly leveraging the data. Our experiments validate the effectiveness of this feature-based approach.

C. Using Fair Epistemic Uncertainty for Fairness-Aware Active Online Learning

In Fair Active Online Learning, tasks arrive sequentially with unlabeled samples, i.e., $\{\mathcal{D}_t^{\mathcal{U}}\}_{t=1}^{T}$. At each time step t, the learner uses a Fair AL strategy to update the fair classifier $h_t : \mathbb{R}^d \times \Theta \to \mathbb{R}$. For each incoming task, the learner decides which samples to query for their labels to maximize information gain, ensure fairness, and adapt to environmental shifts. For each sample x, we propose a score $u(x)$ (explained shortly) to determine which samples to query:

$$u(x) = g(z) - \lambda \sum_{c=1}^{C} p_c^x * \triangle g_c(z) \tag{6}$$

where $z = r(x, \theta_{t-1})$ and r extracts the feature representation of x using parameters θ_{t-1} learned at the previous time step, C is the number of classes, and λ is a parameter controlling the trade-off between epistemic uncertainty, $g(z)$, and the fairness notions $\triangle g_c(z)$. Finally, p_c^x is the softmax probability of sample x belonging to class c, based on the most recent classifier, h_{t-1} trained by the learner in the previous time step. In this formulation for $u(x)$, we say *lower is better*, so the best sample to query would be one that *minimizes* $u(x)$.

We now explain our score, $u(x)$. The term $\sum_{c=1}^{C} p_c^x * \triangle g_c(z)$ is a weighted sum of all $\triangle g_c(z)$ values, where c is the class label. If the class labels were known, we would use only the $\triangle g_c(z)$ value corresponding to the class of x for fairness estimation (as discussed in Section IV-B), instead of summing over all $\triangle g_c(z)$ values as they would be irrelevant. But since this is *active* online learning, we do not know the class of an incoming sample. Thus, we use the softmax probability of the classifier h_{t-1} from the last time-step to estimate the most probable class for x. We then weigh each $\triangle g_c(z)$ value by multiplying with p_c^x, the probability of x belonging to class c.

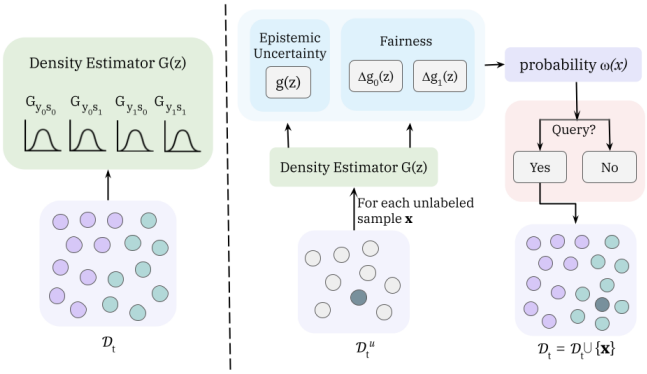


Fig. 1: Sample selection in FACTION: A Density Estimator, $G(z)$, is created from the available labeled training data. Next, each unlabeled example is scored using Epistemic Uncertainty and Fairness measures from the Density Estimator, which is converted to a probability. Finally, a Bernoulli trial using this probability determines whether to query for labels.

Traditional AL focuses on identifying samples with *high* uncertainty and learning from them by backpropagating their losses. Similarly, for fairness, we identify *highly unfair* samples and regularize them with a fairness constraint in the loss to enhance fair predictions on unseen inputs. Lower density, $g(z)$, indicates higher epistemic uncertainty, while a high value of $\sum_{c=1}^{C} p_c^x * \triangle g_c(z)$ indicates higher unfairness. To prioritize both, we negate the $\sum_{c=1}^{C} p_c^x * \triangle g_c(z)$ term in $u(x)$, and select samples with *low* $u(x)$ values, i.e., with high epistemic uncertainty *and* high unfairness.

As an example, consider a stop-and-frisk policy evaluation, where a model is trained to predict whether a stop leads to an arrest. Historical data shows racial disparities in stop decisions [62], with certain demographics overrepresented. A new stop instance involving an individual from an underrepresented demographic in a previously unobserved zip code would likely generate high epistemic uncertainty (low $g(z)$) due to a lack of similar samples in past training data. If this instance also produces large $\Delta g_c(z)$ values, it may suggest bias in how the model associates certain zip codes with racial groups. FACTION prioritizes querying labels for such cases, improving both environmental adaptation (new locations, demographics) and fairness (reducing racial bias in stop predictions).

The Role of Epistemic Uncertainty. Using the epistemic uncertainty, $g(z)$, in $u(x)$, offers several advantages: In the presence of an environment shift in a new task, $g(z)$ prioritizes samples that are most out-of-distribution (OOD) (featuring high epistemic uncertainty [45], [46]), facilitating quick adaptation to new environments. In the absence of a shift, it serves as an effective AL heuristic.

D. A Practical Design for Selection

We use a selection framework similar to prior work [2], [47], where the score $u(x)$ produces a *probability* of querying

for the label of sample x. Scores are normalized to a 0-1 range and then subtracted from 1, to prioritize lower-scoring samples by assigning higher probability. Thus, the probability, $\omega(x)$, is:

$$\omega(x) = 1 - Normalize(u(x)) \tag{7}$$

Next, starting from samples with the highest probability in a batch, we perform Bernoulli trials, denoted by $Bernoulli(p)$, where $p = min(\alpha * \omega(x), 1)$. Here α is a hyperparameter that controls the querying rate for samples. If $Bernoulli(p)$ returns 1, we query the label of a sample. In practice, if a sample x has a normalized score of 0.2 and α is set to 0.9, we have $\omega(x) = 1 - 0.2 = 0.8$, and it would be selected by a Bernoulli trial with probability $p = min(0.9 * 0.8, 1) = 0.72$. We conduct trials until the acquisition batch size for each AL iteration is reached. Fig. 1 shows the selection process.

With data arriving in batches, we normalize scores into probabilities using the maximum and minimum scores in each batch. This can extend to other settings not explored here, like samples arriving individually, where the normalization range can be updated incrementally with all gathered scores.

E. A Fairness-Regularized Loss Function

We utilize a simple fairness regularization in the loss function. On a dataset $\mathcal{D}_t$ whose labels have been acquired (as defined in Section IV-A), and with parameters θ_t at time t, we define the fairness loss:

$$\mathcal{L}_{fair} = [v(\mathcal{D}_t, \boldsymbol{\theta}_t)]_+ \tag{8}$$

where v is the general fairness notion from Section III-A, i.e., $v(\mathcal{D}, \theta)$ and $[\cdot]_+$ is the projection onto the non-negative space. In our experiments, we instantiate $v(\mathcal{D}, \theta)$ as the relaxed form of DDP, from Section III-A. In practice, strict constraints such as $v(\mathcal{D}, \theta) = 0$ are hard to enforce. Instead, we relax these constraints with an empirical constant ϵ, such that $\mathcal{L}_{fair} \leq \epsilon$. Thus, the total loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \mu * (\mathcal{L}_{fair} - \epsilon) \tag{9}$$

where $\mathcal{L}_{CE}$ refers to cross entropy loss, and the hyperparameter μ controls the trade-off between fairness and accuracy.

F. The Full Algorithm

Algorithm 1 summarizes FACTION. For each new task, we first record the learner's performance. Each incoming task, $\mathcal{D}_t^{\mathcal{U}}$, arrives unlabeled, with labels available only for calculating test metrics. Each task has a budget, $\mathcal{B}$. While $\mathcal{B}$ is not exhausted, we do as follows for each task: first, train the classifier on all labeled data available so far, i.e., $\mathcal{D}_t$, using $\mathcal{L}_{total}$ and learning rate γ_t. For simplicity, we keep γ_t constant in our experiments. Next, we fit a density estimator $G(z)$ as shown in Section IV-B. Using this, the score for each unlabeled sample in $\mathcal{D}_t^{\mathcal{U}}$ is calculated using Eq. 6, and is converted to a probability. Next, until the acquisition batch size for each AL iteration, $\mathcal{A}$, is reached, we conduct Bernoulli trials to decide on querying a sample's label, starting with the most probable sample. After acquiring $\mathcal{A}$ labels, we move to the next AL iteration within the same task, repeating the above steps. This continues until we exhaust the total task budget, $\mathcal{B}$. Then, we move to a new task.

Algorithm 1 FACTION

Input: Budget $\mathcal{B}$, Acquisition Batch Size $\mathcal{A}$, Labeled Samples $\mathcal{D}_t$, Query Parameter α, Classes $\mathcal{C}$, Sensitive Values S
Randomly initialize $\theta_{t-0} \in \Theta$,
for each task, $t \in [T]$ **do**
  Record the performance of θ_{t-1} on $\mathcal{D}_t^{\mathcal{U}}$
  $TaskBudget$ = $\mathcal{B}$
  while $TaskBudget > 0$ **do**
    Update model parameters to θ_{temp} by training
    on $\mathcal{D}_t$ using $\mathcal{L}_{total}$ (Eq. 9) and learning rate γ_t
    for each labeled sample, $l \in \mathcal{D}_t$ **do**
      Retrieve feature vector z_l using θ_{temp}
    end for
    Initialize density estimator $G(z)$ as a GMM
    for each $y \in \mathcal{C}$ **do**
      for each $s \in \mathcal{S}$ **do**
        Create component $G_{y,s}(z)$, in $G(z)$
        following Sec. IV-B using feature vector z
      end for
    end for
    $Probs \leftarrow [\,]$
    for each x in $\mathcal{D}_t^{\mathcal{U}}$ **do**
      Calculate $u(x)$ using $G(z)$ and Eq. 6
      $w(x) \leftarrow 1 - Normalize(u(x))$
      Add $(x, w(x))$ to $Probs$
    end for
    Sort Probs in descending order of $w(x)$
    $Acquired \leftarrow 0$
    while $Acquired < \mathcal{A}$ **do**
      for each $(x, w(x))$ in $Probs$ **do**
        $Query \leftarrow Bernoulli(min(\alpha * w(x), 1))$
        if $Query == 1$ **then**
          $\mathcal{D}_t \leftarrow \mathcal{D}_t \cup \{x\}$
          $\mathcal{D}_t^{\mathcal{U}} \leftarrow \mathcal{D}_t^{\mathcal{U}} - \{x\}$
          $Acquired \leftarrow Acquired + 1$
        end if
      end for
    end while
    $TaskBudget \leftarrow TaskBudget - \mathcal{A}$
  end while
  $\theta_t \leftarrow \theta_{temp}$
end for

G. Analysis

For the problem formulation in Section IV-A, we make the following standing assumptions on the loss and constraint functions.

Assumption 1 (Convexity)**.** *Domain Θ is convex and closed. The loss function f_t and the fair function v are convex.*

Assumption 2 ($F-$Lipschitz)**.** *There exists a positive constant F such that*

$$|f_t(\cdot, \boldsymbol{\theta}_1) - f_t(\cdot, \boldsymbol{\theta}_2)| \leq F, \quad ||v(\cdot, \boldsymbol{\theta}_1)|| \leq F,$$
$$\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta, \forall t \in [T]$$

Assumption 3 (Bounded gradient)**.** *The gradients $\nabla f_t(\boldsymbol{\theta})$ and $\nabla v(\boldsymbol{\theta})$ exist, and they are bounded by a positive constant H on Θ, i.e.,*

$$||\nabla f_t(\cdot, \boldsymbol{\theta})|| \leq H, \quad ||\nabla v(\cdot, \boldsymbol{\theta})|| \leq H, \quad \forall \boldsymbol{\theta} \in \Theta, \forall t \in [T]$$

Examples where these assumptions hold include logistic regression and L_2 regression over a bounded domain. Regarding constraints, a family of fairness notions, such as DDP and DEO introduced in Definition 1, are applicable [27]. Suppose Assumption 1-3 holds, Theorem 1 provides the bounds for the query complexity, regret, and violation of cumulative fairness. The detailed proof for Theorem 1 is present in Section VII.

Theorem 1 (Bounds for Query Complexity, Regret, and Violation of Cumulative Fairness)**.** *Under Assumptions 1-3, suppose the task sequence $\{\mathcal{D}_t\}_{t=1}^T$ can be partitioned into m disjoint nonempty subsets $\{\mathcal{I}_u\}_{u=1}^m$, where each $\mathcal{I}_u$ consists of a number of $\mathcal{D}_t$ characterized by a distinct environmental variation. For samples in the task sequence, they lie in an input space of dimension d. With the noise level $\tilde{\eta} = \max\{\eta, 1\}$ and parameter α,*

1) *We define q_t as the number of queries needed at $t \in [T]$ to satisfy the budget $\mathcal{B}$. The query complexity of Algorithm 1 is*

$$Q = \sum_{t=1}^{T} q_t = \mathcal{O}\Big(\sum_{u=1}^{m} \min\Big\{|\mathcal{I}_u|, \tilde{\eta}\sqrt{\alpha d|\mathcal{I}_u|}\Big\} + 1\Big)$$

2) *The learner's regret (Eq. 2) is bounded.*

$$R = \mathcal{O}\Big(\sum_{u=1}^{m} \max\Big\{\tilde{\eta}^2 d, \tilde{\eta}\sqrt{\frac{d|\mathcal{I}_u|}{\alpha}}\Big\}\Big)$$

3) *Given a non-increasing sequence $\{\gamma_t\} \subseteq (0, +\infty)$, with $\gamma_t = \frac{\gamma_0}{\sqrt{m|\mathcal{I}_u|}}$, where $\gamma_0 > 0$ is a constant, we have the bound of the violation of cumulative fairness that*

$$\begin{aligned} V &= \sum_{t=1}^{T} \Big|\Big|[v(\mathcal{D}_t, \boldsymbol{\theta}_t)]_+\Big|\Big| \\ &= \mathcal{O}\Big(\sum_{u=1}^{m} \max\Big\{\tilde{\eta}^2 d, \tilde{\eta}\sqrt{\frac{d\gamma_0\beta|\mathcal{I}_u|}{\alpha}}\Big\}\Big), \\ \text{where } \beta &= \frac{H^2\gamma_0}{\sqrt{m|\mathcal{I}_u|}} + \frac{F^2}{\gamma_0 m|\mathcal{I}_u|} + 2F \end{aligned}$$

Discussion. In the context of changing environments, Theorem 1 provides bounds for query complexity, loss regret, and violation of fairness. However, these bounds on R and V can be extended to a stationary environment, where R and V are required to grow sublinearly in T. In this scenario, the number of non-empty set m is set to 1 and $|\mathcal{I}_u| = T$. We thus derive $R = \mathcal{O}(\sqrt{T})$ and $V = \mathcal{O}(T^{\frac{1}{4}})$.

Limitations of Theoretical Analysis. While our theoretical analysis provides performance guarantees for FACTION, it has limitations. Specifically, Assumption 1 (convexity) may not hold in real-world scenarios, especially with deep networks and non-convex loss landscapes. However, FACTION performs well empirically despite this assumption. Future work could extend these guarantees by relaxing convexity constraints and leveraging advances in non-convex optimization.

H. Alternative Fairness Paradigms and Data Types

While FACTION focuses on group fairness, it could potentially extend to individual and counterfactual fairness. With an appropriate similarity metric, FACTION could enforce individual fairness by penalizing inconsistent treatment of similar samples. For counterfactual fairness, prior work [58] links it to demographic parity, which FACTION already optimizes. Extending our framework explicitly for counterfactual fairness would require modifying our density estimator (Equations 3–5) to compare samples with their counterfactual variants, a challenge that depends on effective counterfactual image generation—a promising direction for future research.

Extensions to other data types. We show that FACTION generalizes to image and tabular data. Since its selection relies on feature representations, extensions to other types of data such as text requires constructing a robust feature space, after which FACTION's density-based uncertainty and fairness estimation can apply naturally.

V. Experiments

A. Experimental Setting

1) Datasets: We evaluate FACTION on 5 datasets.

The *Rotated Colored MNIST* [38] extends RotatedMNIST with 10,000 digits (0-9) rotated by $\{0, 15, 30, 45\}$ degrees, with the rotation angles representing different environments. For binary classification, digits 0-4 are labeled as 0, and 5-9 as 1. Following prior fairness works [48]–[50], we use digit color (green/red) as the sensitive attribute. We create deliberate label-color correlations for each rotation with coefficients $\{0.9, 0.8, 0.7, 0.6\}$, where 0.5 indicates unbiased assignment and 0.9 indicates high bias. Data from each rotation angle is split into 3 equal subsets (tasks), resulting in 12 sequential tasks.

The *CelebA* [39] dataset contains over 200,000 celebrity images. Following prior work [14], [36], we use binary attributes—Young, Male, Attractiveness, and Smiling. To create 4 environments, we combine Young and Smiling. Data from each environment is split into 3 subsets, forming 12 sequential tasks. Similar to prior work [14], Male is the sensitive attribute (gender), and Attractiveness is the class label.

FairFace [40] consists of 108,501 face images across 7 racial groups. Each race defines an environment, with data split into three subsets per race, forming 21 sequential tasks. Gender is the sensitive attribute, and age (binary: 1 for over 50, 0 otherwise) is the target label.

The FFHQ-Features dataset: This dataset is an annotated variant of FFHQ [10], providing images of individuals with annotations for facial expression, age, and gender. We use the four most common facial expressions as environments (additional details in Section V-A3). We set age as a binary classification target (greater than 50 or less), and gender as the sensitive attribute. Each environment is split into 3 tasks, giving a total of 12 tasks.

The *New York Stop-and-Frisk* [11] dataset tracks whether pedestrians stopped on suspicion of weapon possession actually had a weapon. We consider race (black/non-black) as

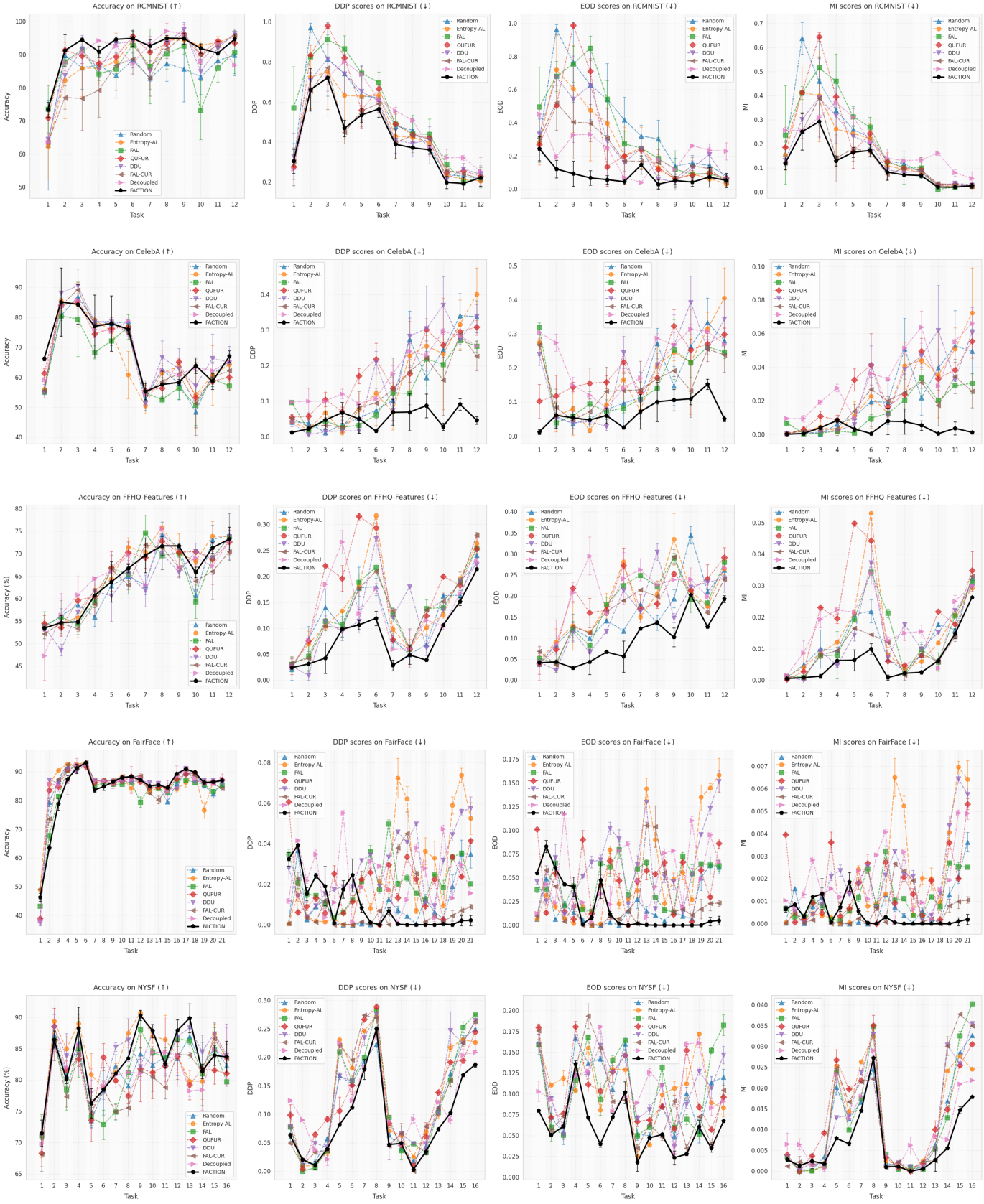


Fig. 2: Results on RCMNIST, CelebA, FFHQ-Features, FairFace and NYSF. Higher is better for accuracy, lower is better for fairness metrics. FACTION provides superior fairness compared to baselines on most tasks while maintaining strong accuracy.

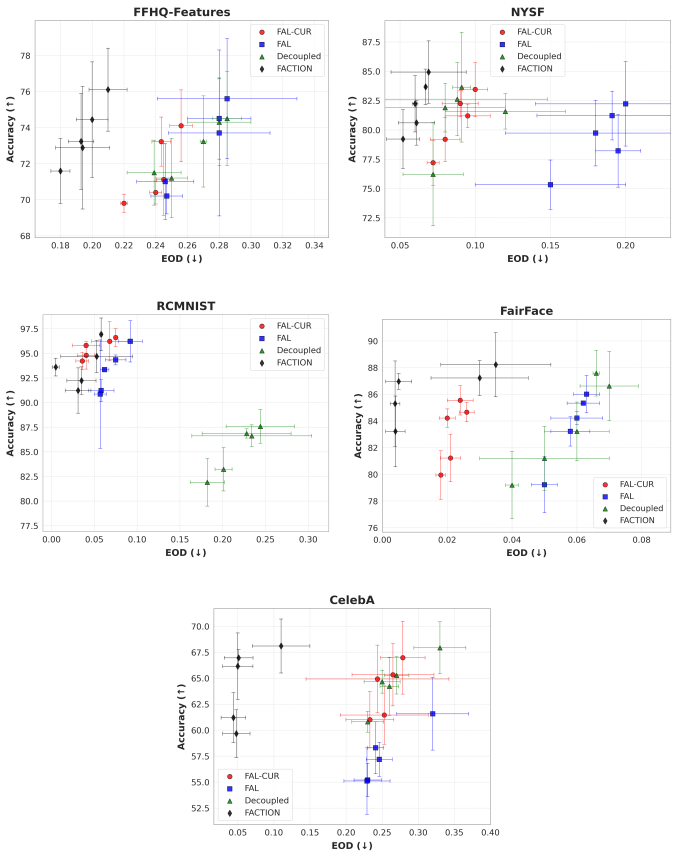


Fig. 3: Trade-offs for the fairness-aware models. As accuracy improves (higher is better), the EOD score tends to worsen (lower is better). Points near the top-left are preferred.

the sensitive attribute. The data is divided by geographic area to create different distributions, with each area's data further grouped by yearly quarters (Jan-Mar, Apr-Jun, etc.) to introduce temporal shifts. This creates 16 tasks (more details in Section V-A3). Our classification target is whether an individual was frisked.

Fairness Metrics. We use 3 popular group fairness metrics, Difference of Demographic Parity (DDP) [52], [27], Equalized Odds Difference (EOD) [51], and Mutual Information (MI) [33], [53], where lower absolute value is better for all.

2) Baselines: To our knowledge, no prior work directly addresses Fair Active Online Learning. Thus, we adapted seven baselines from various perspectives. For Fair AL, **FAL** [33] introduces a notion called "Expected Fairness". **FAL-CUR** [34] is another fairness-aware active learning method, using fair clustering and a representative score for each sample to ensure fairness. **D-FA^{2}L** [12] employs decoupled models for fairness-aware AL, leveraging disagreements between them to identify promising data points. We refer to D-FA^{2}L as **Decoupled** in all results for clarity. We adapt FAL, FAL-CUR, and Decoupled to the online setting by applying them sequentially at each time step. **QuFUR** [2] addresses *active online learning*, balancing regret and the number of labels to query. **DDU** [46] (Deep Deterministic Uncertainty) leverages epistemic uncertainty to achieve strong performance in AL and OOD detection. Entropy-based Active Learning **(Entropy-AL)** [1] is a classical AL approach using Shannon entropy for selection. We also include a naive **Random** baseline, selecting samples at random. Finally, we have our approach, **FACTION**.

3) Hyperparameters and Additional Details: FACTION's hyperparameters were tuned within the following ranges: λ (Eq. 6): $\{0.0001, 0.001, 0.01, 0.1, 1, 5, 10, 100\}$; μ (Eq. 9): $\{0.1, 0.3, 0.5, 0.7, 0.9, 1, 1.2, 1.4, 1.8, 2, 2.4, 2.8, 3\}$; ϵ (Eq. 9): $\{0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.5\}$; and α (Algorithm 1, line 29): $\{0.1, 0.5, 1, 3, 5, 10\}$. The acquisition batch size, $\mathcal{A}$, for each iteration in AL, is a parameter common to all baselines and FACTION. Given the large number of tasks in each dataset (up to 21 in FairFace), we perform batch AL ($\mathcal{A} > 1$) as it is significantly faster. For all methods, we set $\mathcal{A} = 50$. All methods were warm-started with an initial labeled set of 100 randomly selected samples. When testing at each time step with an incoming task, $\mathcal{D}_t^{\mathcal{U}}$, the full dataset is used for evaluation. Experiments were repeated five times, and the mean and standard deviation are reported.

Model Architecture For all experiments and all methods in Figures 2, 3, 4 and 5, we use a standard ResNet-18 with spectral normalization for images and a simple two-layer MLP (hidden dimension 512, ReLU activation) for tabular data. Features are extracted from the first linear layer in the MLP.

Additional Dataset Details. For all datasets, each task must contain more unlabeled samples than the AL budget $\mathcal{B}$ (200, consistent with prior work in Section V-B), for AL to be meaningful. Most tasks across all datasets have over 10 times more samples than the budget. However, for some exceptions, we removed environments with insufficient samples, such as specific quarters of the year in "Staten-Island" (NYSF dataset) and certain emotions like "Contempt" (FFHQ-Features), where task sample counts fell below the budget.

B. Results

In our experiments, performance is sequentially evaluated on each task as it arrives. As unlabeled data batches from each task arrive, the learner acquires samples within a limited budget, $\mathcal{B}$ to learn efficiently, while adapting to the new environment and maintaining fairness. We set $\mathcal{B}$ = 200, in line with our baselines using identical [33] or similar [34] values. We report accuracy and 3 fairness metrics in Figure 2. Here, each row of images represents different metrics evaluated on tasks of a single dataset. FACTION, represented by the black line in all figures, exhibits superior performance compared to baselines across all datasets. Our algorithm is able to achieve better fairness results across metrics on a large majority of the tasks in each dataset, without sacrificing accuracy compared to baselines. In terms of accuracy, competitors like QuFUR and DDU show comparable performance, adapting to new tasks without significant drops in performance unlike some other baselines. Guided by an epistemic uncertainty based selection system, our model is able to match these competitors in accuracy and adapt quickly. However, FACTION far outstrips them in fairness metrics as they are naturally not fairness-

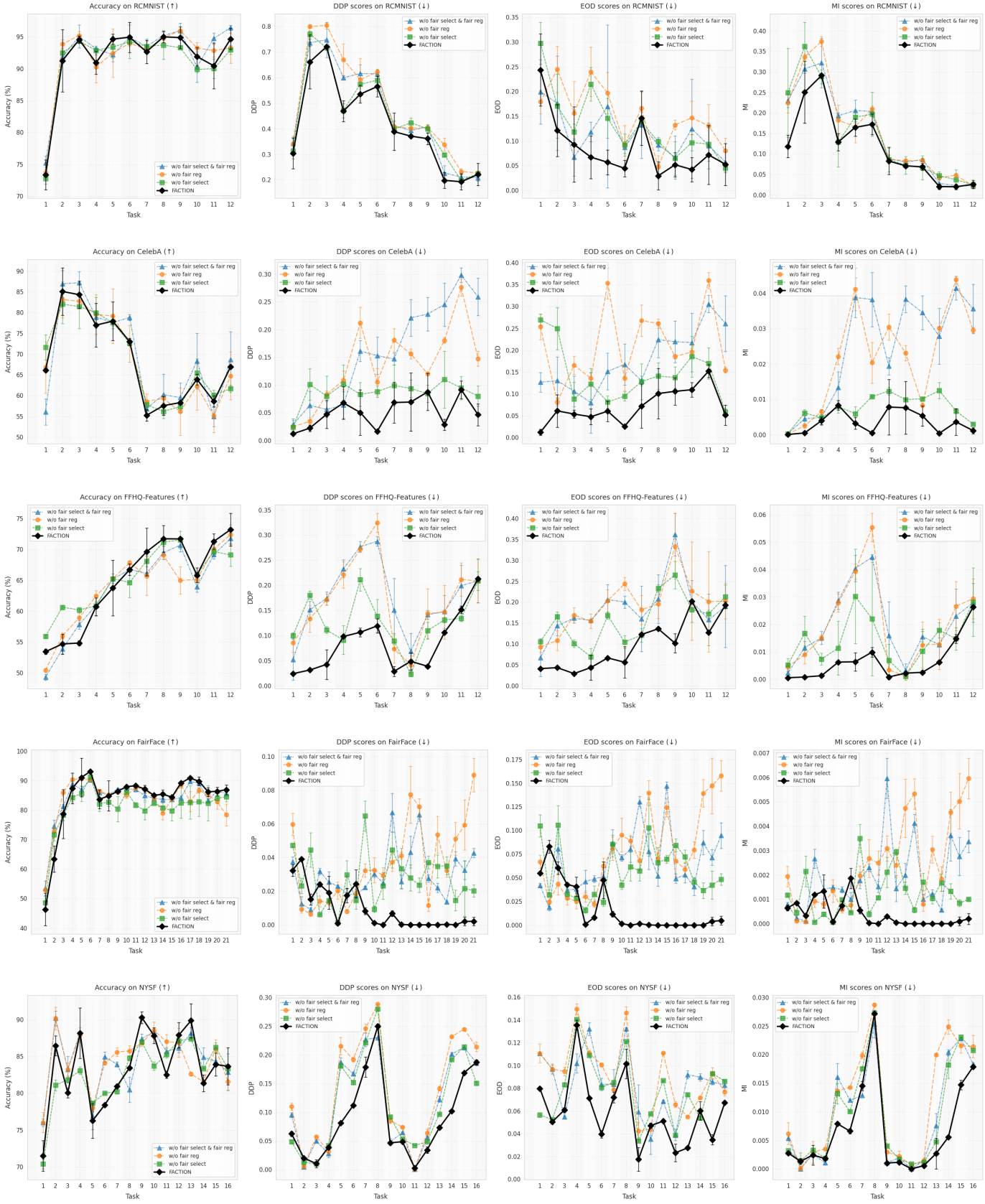


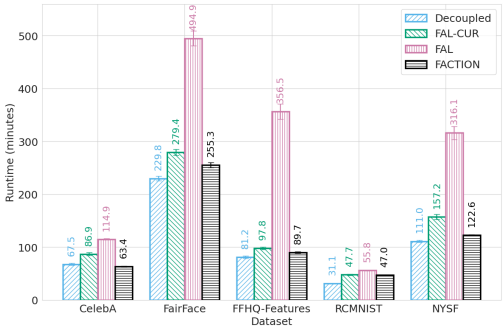
Fig. 4: Ablation experiments on all datasets. Simplified variants exhibit inferior fairness performance.

aware. On the other hand, fairness-aware baselines like FAL-CUR often outperform the non-fairness aware baselines in terms of fairness. Nonetheless, in most cases, FACTION is significantly more fair. We think this is because the fairness-aware baselines only look for fair samples in the selection step, but do not regularize for fairness when learning from them. If the dataset is inherently biased, it is unlikely that the learner will find enough *unbiased* samples to learn from. Thus, we instead identify *unfair* or *biased* samples first through our approach using $\triangle g_c(z)$ values from Section IV-B, and then learn from them through a simple fairness regularizer in the loss function, to make fairer predictions at test time. This approach often leads to large differences in fairness, such as in CelebA, where FACTION achieves a DDP score of 0.05 in the final task, when no competing baseline is able to do better than 0.20. Finally, the fairness-aware baselines often suffer from large accuracy drops on new tasks due to environment shifts, unlike FACTION.

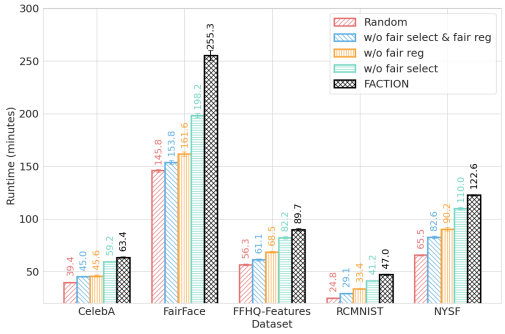
Sensitivity Analysis. We analyze fairness-accuracy trade-offs among fairness-aware methods (FACTION, FAL, FAL-CUR, and Decoupled) across all datasets. Fig. 3 shows how varying fairness parameters affects performance, with preferred models positioned at top-left (higher accuracy, lower EOD). We vary each method's key fairness parameter (for space constraints, please refer to the respective papers for parameter details): FAL's l {64, 96, 128, 196, 256}, FAL-CUR's β {0.3, 0.4, 0.5, 0.6, 0.7}, Decoupled's threshold α {0.1, 0.2, 0.4, 0.6, 0.8}, and FACTION's μ {0.3, 0.5, 0.7, 1.4, 2.8} which directly controls the fairness-accuracy tradeoff through regularization strength. Results show mean and standard deviation from 5 runs per configuration.

Ablation. We conduct ablation studies on all datasets, as shown in Figure 4. Here, "w/o Fair Select" removes fairness criteria from the selection step,"w/o Fair Reg" removes the Fairness Regularizer from the loss function, and "w/o Fair select & Fair Reg" removes both. We see that all variants exhibit inferior fairness.

Runtimes. We measured empirical runtimes to investigate computational costs of fairness methods. Figure 5a shows average total runtimes across 5 runs on a Tesla V100 GPU with Intel(R) Xeon(R) E5-2680 CPU. FACTION runs faster than FAL and FAL-CUR, with FAL's expected fairness calculation being very expensive. While FACTION is slightly slower than Decoupled (which uses a simpler model disagreement-based selection approach), it achieves significantly better fairness results on all datasets (Figure 2). Figure 5b compares FACTION with simplified variants, including "Random" (everything removed), "w/o fair select & fair reg" (only epistemic uncertainty), and other variants are defined similarly to the ablation. Evidently, runtimes increase as we add components, but remain reasonable, with the full system requiring less than twice the runtime of random selection on all datasets. Table I shows the runtime and performance of FACTION variants on NYSF, reporting the mean across 16 tasks. The full FACTION system significantly improves fairness while maintaining competitive accuracy. Compared to its non-fairness-aware variant (w/o fair select & fair reg), it sacrifices just over 1% accuracy for substantial gains in DDP (0.118 vs 0.089, 24.5% improvement), EOD (0.084 vs 0.059, 29.8% improvement), and MI (0.009 vs 0.006, 33.3% improvement).



(a) Runtimes of all fairness-aware models.



(b) Runtimes of FACTION with ablated variants.

Fig. 5: Runtime comparisons for FACTION.

Model	Runtime(m)	Acc(↑) / DDP(↓) / EOD(↓) / MI(↓)
Random	65.2	81.44 / 0.114 / 0.101 / 0.011
w/o fair sel. & fair reg	82.6	**84.51** / 0.118 / 0.084 / 0.009
w/o fair reg	90.2	84.50 / 0.138 / 0.091 / 0.012
w/o fair select	110.0	82.73 / 0.110 / 0.078 / 0.010
FACTION	122.6	83.41 / **0.089** / **0.059** / **0.006**

TABLE I: FACTION compared to its ablated variants on runtime and performance (mean across all tasks) in NYSF.

Additional Results. To assess FACTION's generality, we test it with Wide ResNet-50 (WRN-50) [57] on the CelebA dataset, applying it to both FACTION and all baselines. Figure 6 shows that FACTION consistently improves fairness while maintaining competitive accuracy with other methods.

VI. Conclusion

Our research proposes FACTION, to effectively handle the unique challenges of the practical and novel Fairness-Aware Active Online Learning paradigm. Through extensive experiments on both images and tabular data, and with robust theoretical guarantees, we demonstrate that FACTION adapts to changing data distributions while intelligently querying for a limited budget of labels and ensuring fairness in its predictions. This is achieved by a lightweight combination of epistemic uncertainty, a novel fairness notion and simple regularization, which enables FACTION to perform fair, effective and efficient sample selection in dynamic environments.

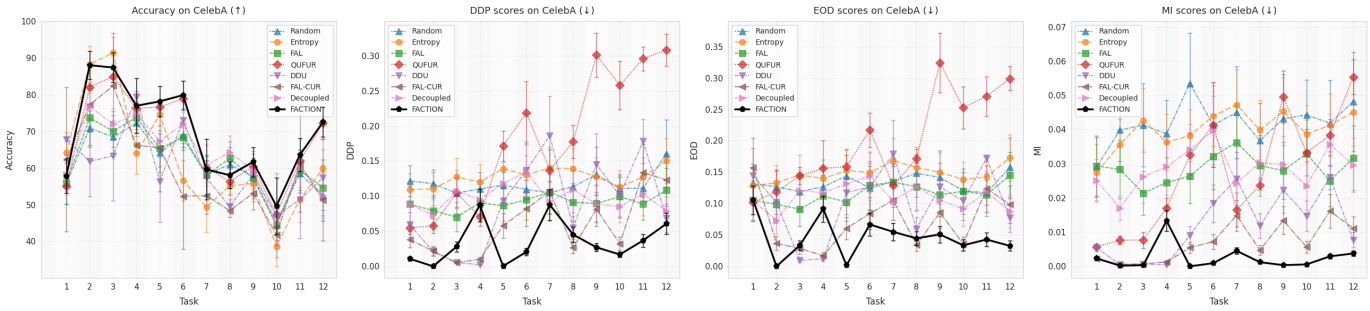


Fig. 6: Performance of all methods on the CelebA dataset using the WRN-50 model.

VII. Proofs of Theoretical Results

Before presenting the proof of Theorem 1, two Lemmas are provided.

Lemma 1. *In the setting of Theorem 1, with probability* $1-\frac{\delta}{2}$, *we have* $f_t(\mathcal{D}_t) - f_t^*(\mathcal{D}_t) = \mathcal{O}(\omega(\mathcal{D}_t)), \forall t \in [T]$.

Lemma 2. *Let* $\mathbf{a}_1, \cdots, \mathbf{a}_k$ *be* k *vectors in* $\mathbb{R}^p$. *For* $i \in [k]$, *define* $N_i = \lambda I + \sum_{j=1}^{i=1} \mathbf{a}_j \mathbf{a}_j^T$. *Then for any* $S \subseteq [k]$, $\sum_{i\in S} \omega(\mathbf{a}_i) \leq \ln \frac{\det(\lambda I + \sum_{i\in S} \mathbf{a}_i \mathbf{a}_i^T)}{\det(\lambda I)}$.

Lemma 1 upper bounds the regret with the sum of score estimates $\omega(\mathcal{D}_t)$. Lemma 2 bounds the sum of estimated scores using $\omega(\cdot)$ for k queried samples in an environment. The proofs of Lemma 1 and 2 are omitted, as they are extended from [2]. We now give the proof of Theorem 1.

Proof. Let $p_t = \min\{1, \alpha\omega(\mathbf{x}_t)\}$ be the learner's query probability at time t. It is easy to see that $\mathbb{E}_{t-1}[q_t|\mathbf{x}_t, H_{t-1}] = p_t$, where $q_t = \text{Bernoulli}(\min\{1, \alpha\omega(\mathbf{x}_t)\})$ and $H_{t-1} = \{\mathbf{x}_{1:t}, h_{1:t}, \xi_{1:t}\}$. For simplicity, we denote $\mathbb{E}_{t-1} := \mathbb{E}_{t-1}[q_t|\mathbf{x}_t, H_{t-1}]$. Let random variable $Z_t = q_t\omega(\mathbf{x}_t)$. We have the following simple facts:

$$Z_t \leq \tilde{\eta}^2; \quad \mathbb{E}_{t-1} Z_t = p_t\omega(\mathbf{x}_t);$$
$$\mathbb{E}_{t-1} Z_t^2 \leq \tilde{\eta}^2 \mathbb{E}_{t-1} Z_t \leq \tilde{\eta}^2 p_t \omega(\mathbf{x}_t)$$

For every $u \in [m]$, define event

$$F_u = \Big\{ \Big| \sum_{t\in\mathcal{I}_u} p_t\omega(\mathbf{x}_t) - \sum_{t\in\mathcal{I}_u} q_t\omega(\mathbf{x}_t) \Big| \tag{10}$$
$$\leq \mathcal{O}\Big(\tilde{\eta}\sqrt{\sum_{t\in\mathcal{I}_u} p_t\omega(\mathbf{x}_t) \ln\frac{T}{\delta}} + \ln\frac{T}{\delta}\Big)\Big\} \tag{11}$$

Applying Freedman's inequality to $\{Z_t\}_{t\in\mathcal{I}_u}$, we have $\mathbb{P}(F_u) \geq 1 - \frac{\delta}{4m}$. Similarly,

$$G = \Big\{ \Big| \sum_{t\in\mathcal{I}_u} p_t - \sum_{t\in\mathcal{I}_u} q_t \Big| \leq \mathcal{O}\Big(\tilde{\eta}\sqrt{\sum_{t\in\mathcal{I}_u} p_t \ln\frac{T}{\delta}} + \ln\frac{T}{\delta}\Big)\Big\} \tag{12}$$

Applying Freedman's inequality to $\{q_t\}_{t\in\mathcal{I}_u}$, we have $\mathbb{P}(G) \geq 1 - \frac{\delta}{4}$.

By the definition of F_u, solving for $\sum_{t\in\mathcal{I}_t} p_t\omega(\mathbf{x}_t)$ in Eq.(10), we have

$$\sum_{t\in\mathcal{I}_t} p_t\omega(\mathbf{x}_t) = \mathcal{O}\Big(\sum_{t\in\mathcal{I}_t} q_t\omega(\mathbf{x}_t) + \tilde{\eta}^2\Big)$$

Using Lemma 2 with $\{\mathbf{a}_i\}_{i=1}^k = \{\mathbf{x}_t\}_{t\in\mathcal{Q}_t}$ where $\mathcal{Q}_t$ is the set of labeled samples seen up to time $t-1$, and $S = \mathcal{I}_u \cap \mathcal{Q}_T$, we have

$$\sum_{t\in\mathcal{I}_u} q_t\omega(\mathbf{x}_t) \leq \tilde{\eta}^2 \ln\det\Big(I + C^2 \sum_{t\in\mathcal{I}_u\cap\mathcal{Q}_T} \mathbf{x}_t\mathbf{x}_t^T\Big)$$
$$\leq 2\tilde{\eta}^2 d \ln(1 + C^2\frac{|\mathcal{I}_u|}{d}) = \mathcal{O}(\tilde{\eta}^2 d)$$

By combining with $\sum_{t\in\mathcal{I}_t} p_t\omega(\mathbf{x}_t)$, we have $\sum_{t\in\mathcal{I}_t} p_t\omega(\mathbf{x}_t) = \mathcal{O}(\tilde{\eta}^2 d)$.

We divide the samples in environment u into high and low risk subsets with index sets $\mathcal{I}_{u,+}$ and $\mathcal{I}_{u,-}$. For simplicity, we omit the subscript u hereafter.

$$\mathcal{I}_+ = \{t \in \mathcal{I}_u : \alpha\omega(\mathbf{x}_t) > 1\}, \mathcal{I}_- = \mathcal{I}\backslash\mathcal{I}_+$$

We consider bounding the regrets and the query complexities in these two sets:

(1) For every $t \in \mathcal{I}_+$, $p_t = 1$, label y_t is queried,

$$\sum_{t\in\mathcal{I}_+} \omega(\mathbf{x}_t) = \sum_{t\in\mathcal{I}_+} q_t\omega(\mathbf{x}_t) \leq \sum_{t\in\mathcal{I}_u} q_t\omega(\mathbf{x}_t) = \mathcal{O}(\tilde{\eta}^2 d)$$

Since for every t in $\mathcal{I}_-$, $\omega(\mathbf{x}_t) > \frac{1}{\alpha}$, we have $\sum_{t\in\mathcal{I}_+} \omega(\mathbf{x}_t) > \frac{|\mathcal{I}_+|}{\alpha}$. This implies that $\sum_{t\in\mathcal{I}_+} p_t = |\mathcal{I}_+| = \mathcal{O}(\alpha\tilde{\eta}^2 d)$.

(2) For every $t \in \mathcal{I}_-$, $p_t = \alpha\omega(\mathbf{x}_t)$. Therefore,

$$\sum_{t\in\mathcal{I}_-} \alpha\omega(\mathbf{x}_t)^2 = \sum_{t\in\mathcal{I}_-} p_t\omega(\mathbf{x}_t) \leq \sum_{t\in\mathcal{I}_u} p_t\omega(\mathbf{x}_t) = \mathcal{O}(\tilde{\eta}^2 d)$$

By Cauchy-Schwarz, and the fact that $|\mathcal{I}_-| \leq |\mathcal{I}_u|$, we have

$$\sum_{t\in\mathcal{I}_-} \omega(\mathbf{x}_t) \leq \sqrt{|\mathcal{I}_-|(\sum_{t\in\mathcal{I}} \omega(\mathbf{x}_t)^2)} = \mathcal{O}(\tilde{\eta}\sqrt{d\frac{|\mathcal{I}_u|}{\alpha}})$$

Consequently, $\sum_{t\in\mathcal{I}_-} p_t = \sum_{t\in\mathcal{I}_-} \alpha\omega(\mathbf{x}_t) \leq \mathcal{O}(\tilde{\eta}\sqrt{\alpha d|\mathcal{I}_u|})$. Summing over the two cases, we have

$$\sum_{t\in\mathcal{I}_u} p_t \leq \mathcal{O}(\alpha\tilde{\eta}^d + \tilde{\eta}\sqrt{\alpha d|\mathcal{I}_u|})$$
$$\sum_{t\in\mathcal{I}_u} \omega(\mathbf{x}_t) \leq \mathcal{O}(\tilde{\eta}^2 d + \tilde{\eta}\sqrt{d\frac{|\mathcal{I}_u|}{\alpha}})$$

If $\alpha \leq \frac{|\mathcal{I}_u|}{\tilde{\eta}^2 d}$, we have $\alpha\tilde{\eta}^2 d \leq \tilde{\eta}\sqrt{\alpha d|\mathcal{I}_u|}$, otherwise we use the trivial bound $\sum_{t\in\mathcal{I}_u} p_t \leq |\mathcal{I}_u|$. The above bounds can be simplified to

$$\sum_{t\in\mathcal{I}_u} p_t \leq \mathcal{O}(\min\{|\mathcal{I}_u|, \tilde{\eta}\sqrt{\alpha d|\mathcal{I}_u|}\}) \tag{13}$$
$$\sum_{t\in\mathcal{I}_u} \omega(\mathbf{x}_t) \leq \mathcal{O}(\max\{\tilde{\eta}^2 d, \tilde{\eta}\sqrt{d\frac{|\mathcal{I}_u|}{\alpha}}\}) \tag{14}$$

For the query complexity, from the definition of event G, applying AM-GM inequality on Eq.(12), we have

$$Q = \sum_{t=1}^T q_t = \mathcal{O}\Big(\sum_{t=1}^T p_t + 1\Big) = \mathcal{O}\Big(\sum_{u=1}^m \min\Big\{|\mathcal{I}_u|, \tilde{\eta}\sqrt{\alpha d|\mathcal{I}_u|}\Big\} + 1\Big)$$

Similarly, the regret guarantee is derived by using the definition of event E, Lemma 1, and Eq. (14).

Furthermore, with the Lemma 1 presented in [8], it yields $||\mu||^2 \leq \beta m|\mathcal{I}_u|$, where $\beta = \frac{H^2\gamma_0}{\sqrt{m|\mathcal{I}_u|}} + \frac{F^2}{\gamma_0 m|\mathcal{I}_u|} + 2F$. Together this inequality with $\gamma_t = \frac{\gamma_0}{\sqrt{m|\mathcal{I}_u|}}, \forall t \in [T]$, we have

$$V \leq \frac{\tilde{\eta}\sqrt{m|\mathcal{I}_u|\gamma_0}}{\alpha m|\mathcal{I}_u|}||\mu|| \leq \tilde{\eta}\sqrt{\frac{d\gamma_0\beta|\mathcal{I}_u|}{\alpha}}$$

which yields the bounds for the violation of the long-term constraints. □

Acknowledgment

The research reported herein was supported in part by NSF grant number 2147375 and NIST grant number 60NANB24D143. Any opinions, findings, conclusions, and recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF or NIST.

REFERENCES

[1] B. Settles, "Active learning literature survey," Univ. Wisconsin-Madison Dept. Comput. Sci., 2009.

[2] Y. Chen, H. Luo, T. Ma, and C. Zhang, "Active online learning with hidden shifting domains," in *Proc. Int. Conf. Artif. Intell. Stat.*, pp. 2053–2061, 2021.

[3] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2165–2176, 2004.

[4] A. Daniely, A. Gonen, and S. Shalev-Shwartz, "Strongly adaptive online learning," in *Proc. Int. Conf. Mach. Learn.*, pp. 1405–1411, 2015.

[5] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *Proc. IEEE Int. Conf. Data Min. Workshops*, pp. 643–650, 2011.

[6] X. Li *et al.*, "Dark-skin individuals are at more risk on the street: Unmasking fairness issues of autonomous driving systems," *arXiv preprint arXiv:2308.02935*, 2023.

[7] M. Brandao, "Age and gender bias in pedestrian detection algorithms," *arXiv preprint arXiv:1906.10490*, 2019.

[8] X. Yi, X. Li, T. Yang, L. Xie, T. Chai, and K. Johansson, "Regret and cumulative constraint violation analysis for online convex optimization with long term constraints," in *Proc. Int. Conf. Mach. Learn.*, pp. 11998–12008, 2021.

[9] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine Learning*, vol. 110, no. 3, pp. 457–506, 2021.

[10] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4217–4228, Dec. 2021.

[11] P. W. Koh *et al.*, "WILDS: A Benchmark of in-the-Wild Distribution Shifts," in *Proc. ICML*, 2021.

[12] Y. Cao and C. Lan, "Fairness-Aware Active Learning for Decoupled Model," in *Proc. 2022 Int. Joint Conf. Neural Networks (IJCNN)*, pp. 1–9, 2022.

[13] J. Liu *et al.*, "Simple and principled uncertainty estimation with deterministic deep learning via distance awareness," in *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 7498–7512, 2020.

[14] S. Park, J. Lee, P. Lee, S. Hwang, D. Kim, and H. Byun, "Fair contrastive learning for facial attribute classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 10389–10398, 2022.

[15] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 21464–21475, 2020.

[16] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. Int. Conf. Mach. Learn.*, pp. 1530–1538, 2015.

[17] V. Dutordoir, H. Salimbeni, J. Hensman, and M. Deisenroth, "Gaussian process conditional density estimation," in *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[18] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[19] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral Normalization for Generative Adversarial Networks," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[20] Y. Wen, D. Tran, and J. Ba, "Batchensemble: An alternative approach to efficient ensemble and lifelong learning," *arXiv preprint arXiv:2002.06715*, 2020.

[21] L. Wimmer, Y. Sale, P. Hofman, B. Bischl, and E. Hüllermeier, "Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures?" in *Proc. Uncertainty in Artif. Intell.*, pp. 2282–2292, 2023.

[22] L. Smith and Y. Gal, "Understanding measures of uncertainty for adversarial example detection," *arXiv preprint arXiv:1803.08533*, 2018.

[23] M. Dusenberry *et al.*, "Efficient and scalable Bayesian neural nets with rank-1 factors," in *Proc. Int. Conf. Mach. Learn.*, pp. 2782–2792, 2020.

[24] Y. Wu, L. Zhang, and X. Wu, "On Convexity and Bounds of Fairness-aware Classification," in *WWW*, 2019.

[25] C. Zhao and F. Chen, "Rank-Based Multi-task Learning for Fair Regression," in *IEEE Int. Conf. Data Mining (ICDM)*, 2019.

[26] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Comput. Surveys*, vol. 55, no. 3, pp. 1–44, 2022.

[27] M. Lohaus, M. Perrot, and U. Von Luxburg, "Too Relaxed to Be Fair," in *Proc. ICML*, 2020.

[28] C. Zhao *et al.*, "Adaptive Fairness-Aware Online Meta-Learning for Changing Environments," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Min.*, pp. 2565–2575, 2022.

[29] R. Wu *et al.*, "Online adaptation to label distribution shift," in *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 11340–11351, 2021.

[30] A. Li, A. Boyd, P. Smyth, and S. Mandt, "Detecting and Adapting to Irregular Distribution Shifts in Bayesian Online Learning," in *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 6816–6828, 2021.

[31] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro, "Learning non-discriminatory predictors," in *Conf. Learn. Theory*, pp. 1920–1953, 2017.

[32] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proc. Int. Conf. Mach. Learn.*, pp. 325–333, 2013.

[33] H. Anahideh, A. Asudeh, and S. Thirumuruganathan, "Fair active learning," *Expert Syst. Appl.*, vol. 199, p. 116981, 2022.

[34] R. M. Fajri, A. Saxena, Y. Pei, and M. Pechenizkiy, "Fal-cur: Fair active learning using uncertainty and representativeness on fair clustering," *Expert Syst. Appl.*, vol. 242, p. 122842, 2024.

[35] J. D. Abernethy *et al.*, "Active Sampling for Min-Max Fairness," in *Proc. Int. Conf. Mach. Learn.*, pp. 53–65, 2022.

[36] Z. Wang *et al.*, "Towards fairness in visual recognition: Effective strategies for bias mitigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 8919–8928, 2020.

[37] S. Shekhar, G. Fields, M. Ghavamzadeh, and T. Javidi, "Adaptive sampling for minimax fair classification," in *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 24535–24544, 2021.

[38] M. Ghifary, W. B. Kleijn, M. Zhang, and D. Balduzzi, "Domain generalization for object recognition with multi-task autoencoders," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2551–2559, 2015.

[39] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015.

[40] K. Karkkainen and J. Joo, "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pp. 1548–1558, 2021.

[41] P. Ren *et al.*, "A Survey of Deep Active Learning," *ACM Comput. Surveys*, vol. 54, no. 9, art. no. 180, pp. 1–40, Oct. 2022.

[42] T. Scheffer, C. Decomain, and S. Wrobel, "Active Hidden Markov Models for Information Extraction," in *Advances in Intelligent Data Analysis*, Berlin, Heidelberg: Springer, 2001, pp. 309–318.

[43] D. D. Lewis and J. Catlett, "Heterogeneous Uncertainty Sampling for Supervised Learning," in *Machine Learning Proceedings 1994*, San Francisco, CA: Morgan Kaufmann, 1994, pp. 148–156.

[44] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *Proc. Int. Conf. Mach. Learn.*, pp. 1183–1192, 2017.

[45] A. Kirsch, J. Mukhoti, J. van Amersfoort, P. H. S. Torr, and Y. Gal, "On Pitfalls in OoD Detection: Entropy Considered Harmful," in *Uncertainty and Robustness in Deep Learning Workshop*, ICML, 2021.

[46] J. Mukhoti, A. Kirsch, J. van Amersfoort, P. H. S. Torr, and Y. Gal, "Deep Deterministic Uncertainty: A New Simple Baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 24384–24394, 2023.

[47] J. Lu, P. Zhao, and S. C. Hoi, "Online passive-aggressive active learning," *Mach. Learn.*, vol. 103, pp. 141–183, 2016.

[48] T. Li, Z. Li, A. Li, M. Du, A. Liu, Q. Guo, G. Meng, and Y. Liu, "Fairness via Group Contribution Matching," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, pp. 436–445, 2023.

[49] R. Chen *et al.*, "Fast model debias with machine unlearning," in *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[50] V. Piratla, P. Netrapalli, and S. Sarawagi, "Focus on the Common Good: Group Distributional Robustness Follows," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2022.

[51] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.

[52] K. Padh, D. Antognini, E. Lejal-Glaude, B. Faltings, and C. Musat, "Addressing Fairness in Classification with a Model-Agnostic Multi-Objective Algorithm," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 161, pp. 8215–8224, 2021.

[53] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Proc. Mach. Learn. Knowl. Discov. Databases: Eur. Conf., ECML PKDD*, Bristol, UK, Sep. 24–28, 2012, Part II, pp. 35–50. Springer, 2012.

[54] R. Salazar, F. Neutatz, and Z. Abedjan, “Automated feature engineering for algorithmic fairness,” *Proc. VLDB Endowment*, vol. 14, no. 9, pp. 1694–1702, 2021.

[55] Y. Lin, S. Gupta, and H. V. Jagadish, “Mitigating Subgroup Unfairness in Machine Learning Classifiers: A Data-Driven Approach,” in *Proc. IEEE 40th Int. Conf. Data Eng. (ICDE)*, pp. 2151–2163, 2024, doi: 10.1109/ICDE60146.2024.00171.

[56] S. Guha, F. A. Khan, J. Stoyanovich, and S. Schelter, “Automated Data Cleaning Can Hurt Fairness in Machine Learning-based Decision Making,” in *Proc. IEEE 39th Int. Conf. Data Eng. (ICDE)*, pp. 3747–3754, 2023, doi: 10.1109/ICDE55515.2023.00303.

[57] S. Zagoruyko and N. Komodakis, “Wide residual networks,” *arXiv preprint arXiv:1605.07146*, 2016.

[58] L. Rosenblatt and RT Witter. “Counterfactual fairness is basically demographic parity.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 12. 2023.

[59] E. Chzhen, C. Giraud, and G. Stoltz, “A Unified Approach to Fair Online Learning via Blackwell Approachability,” in Advances in Neural Information Processing Systems, vol. 34, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), pp. 18280–18292, Curran Associates, Inc., 2021.

[60] S. Cayci, S. Gupta, and A. Eryilmaz. ”Group-fair online allocation in continuous time.” Advances in Neural Information Processing Systems 33 (2020): 13750-13761.

[61] V. Iosifidis and E. Ntoutsi. ”FABBOO - Online Fairness-Aware Learning Under Class Imbalance.” International Conference on Discovery Science. Cham: Springer International Publishing, 2020.

[62] NYCLU, “Long-Awaited Stop-and-Frisk Data Raises Questions About Racial Profiling and Overly Aggressive Policing,” 2024. [Online]. Available: https://www.nyclu.org/press-release/long-awaited-stop-and-frisk-data-raises-questions-about-racial-profiling-and-overly. [Accessed: Feb. 25, 2025].