DADI: DYNAMIC DISCOVERY OF FAIR INFORMATION WITH ADVERSARIAL REINFORCEMENT LEARNING

Michiel A. Bakker
Massachusetts Institute of Technology
MIT-IBM Watson AI Lab
bakker@mit.edu

Duy Patrick Tu
Massachusetts Institute of Technology
MIT-IBM Watson AI Lab
patrick2@mit.edu

Krishna P. Gummadi
Max Planck Institute for
Software Systems
gummadi@mpi-sws.org

Kush R. Varshney
IBM Research
MIT-IBM Watson AI Lab
krvarshn@us.ibm.com

Adrian Weller
University of Cambridge
Alan Turing Institute
aw665@cam.ac.uk

Alex 'Sandy' Pentland
Massachusetts Institute of Technology
MIT-IBM Watson AI Lab
pentland@mit.edu

ABSTRACT

We introduce a framework for dynamic adversarial discovery of information (DADI), motivated by a scenario where information (a feature set) is used by third parties with unknown objectives. We train a reinforcement learning agent to sequentially acquire a subset of the information while balancing accuracy and fairness of predictors downstream. Based on the set of already acquired features, the agent decides dynamically to either collect more information from the set of available features or to stop and predict using the information that is currently available. We attain group fairness (demographic parity) by rewarding the agent with an adversarial loss. Finally, we demonstrate empirically, using two real-world datasets, that we can trade-off fairness and predictive performance.

1 INTRODUCTION

Two parties are involved in information transfer: a *data owner* who has ownership over its own data or data it holds on behalf of others and a *data collector* who is tasked with collecting the most informative set of data, often to maximize the performance of some predictor downstream. Intentionally or otherwise, this process of data collection and prediction can lead to biases that favor one population subgroup over another. Numerous recent studies have shown that naively optimizing for accuracy can lead to unfair outcomes in high-stakes domains such as criminal justice, credit assessment, recruiting, and healthcare Kleinberg et al. (2017); Chalfin et al. (2016); Huang et al. (2007); Obermeyer et al. (2019).

Consequently, the data owner faces a critical decision: if it cannot trust the data collector, which information should it share to ensure fair decision making? While the optimal strategy to maximize predictive performance is to naively share all the data available, the data owner has to be more careful when it wants to ensure that the predictions downstream are fair. Removing the sensitive attribute is the most obvious strategy, but is ineffective when the attribute is redundantly encoded in other features Dwork et al. (2012). One line of work has proposed to learn a model that maps the original set of features to a 'fair' representation that can be shared safely with a data collector Edwards & Storkey (2016); Madras et al. (2018). These representations are made independent of the sensitive attribute using adversarial learning techniques in order to achieve demographic parity Edwards & Storkey (2016). At a high level, the idea is that if the representations from different subgroups are similar to each other, then any predictive model downstream will make decisions independent of the sensitive attribute. Though effective, a limitation of this strategy is that it only grants a data collector

access to an abstract representation of the data. For many applications, such as when information is used or audited by both human and machine decision makers, a data collector needs access to a set of attributes that describe an individual in a way that can be understood by a human Biran & Cotton (2017). If we consider the example of credit assessment, a bank not only collects data to assess initial creditworthiness but also wants to justify and explain the credit decision to an applicant, and store the applicant's information in a database to allow for audits and provide other services downstream.

For these reasons, we need a strategy that can select a set of features that are independent of the sensitive attribute at an individual level. For example, for individuals that live in Chicago, the most racially segregated city in America, zipcode will be highly correlated with race and using this feature can thus lead to racially biased predictions Logan (2014). In contrast, if an individual lives in Irvine, California, America's most racially integrated city, zipcode alone will not reveal an individual's race. Removing zipcode for all individuals is therefore an inefficient strategy for ensuring fairness. An effective feature selection policy can take these nuances into account at the individual level: first query and process the city attribute before deciding whether or not the zipcode feature can be collected safely. Therefore, we build on prior literature concerning active-feature value acquisition (AFA), developed for applications in which a decision maker trades-off accuracy with the acquisition costs of features. AFA allows one to predict which next feature should be selected based on the set of already collected features Krishnapuram et al. (2011).

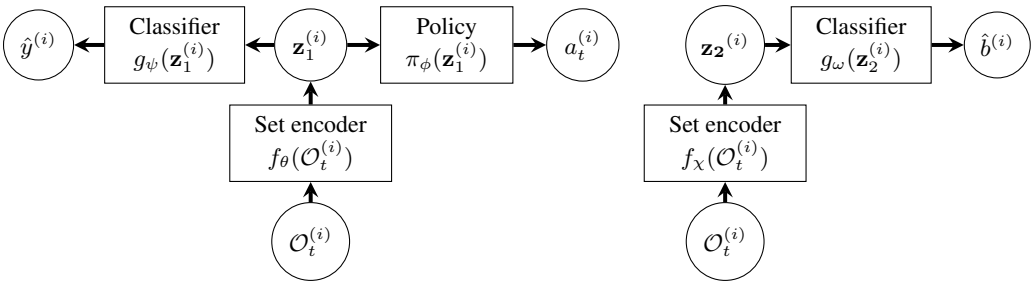
In this work, we develop an AFA method for the selection of personalized sets of features that ensure fair predictions downstream. Our contributions are as follows: to the best of our knowledge, we introduce the first framework for dynamic adversarial discovery of information (DADI) which we utilize to acquire feature sets that ensure fair decision making. In this framework, we formulate the feature acquisition task as a minimax optimization problem in which a reinforcement learning (RL) agent minimizes classification loss while maximizing the loss of an adversary. We actualize this with a joint framework that simultaneously trains a classifier, an adversary, and an RL agent using deep Q-learning. Finally, we demonstrate the effectiveness of our framework with two real-world public datasets.

2 Related Work

Fairness Recent years have seen a vast increase in academic work that seeks to define and obtain fairness in machine learning-based decision making systems. At a high level, this literature has focused on two families of definitions: *statistical* notions of fairness and *individual* notions of fairness Verma & Rubin (2018). Most of the literature, including this work, focuses on statistical or group definitions of fairness, in which we require parity of some statistical measure to hold across a small number of protected subgroups. In contrast, individual fairness definitions have no notion of protected subgroups, but instead formulate constraints that bind on pairs of individuals Dwork et al. (2012); Joseph et al. (2016). Both families of definitions have strengths and weaknesses; statistical notions are easy to verify but do not provide any guarantees to individuals, while individual notions do give individual guarantees but need a predefined distance metric to compare individuals which can be difficult to agree on in practice.

In this work, we focus on *demographic parity*, requiring parity of the positive classification rate across groups, i.e. $P(\hat{y} = 1 \mid b = 0) = P(\hat{y} = 1 \mid b = 1)$, where $\hat{y} \in \{0, 1\}$ is the binary prediction of a model that classifies feature set $\mathbf{x}$ and $b \in \{0, 1\}$ is the sensitive attribute. We refer to Verma & Rubin (2018) for a survey of fairness definitions. We demonstrate the effectiveness of our framework using demographic parity but note that alternative adversarial objectives have been introduced that could be combined with our framework to achieve other notions of fairness such as equal opportunity and equal odds Madras et al. (2018).

Adversarial training Adversarial training, introduced for deep generative modeling by Goodfellow et al. (2014), was first applied in the context of fairness by Edwards & Storkey (2016) to ensure that multiple distinct data distributions from different demographic subgroups are modeled as a single representation. The discriminator aims to distinguish between subgroups while an encoder aims to map each data distribution to a single representation to fool the discriminator. Subsequently, these representations can be safely shared with a data collector while ensuring that the representations are independent of the sensitive attribute and thus guaranteeing demographic parity for predictions



(a) Joint model for the classifier and policy.

(b) Model for the adversary.

Figure 1: Joint framework for dynamic adversarial discovery of information (DADI).

downstream. Beutel et al. (2017) further explore this approach in the context of demographically imbalanced data, Madras et al. (2018) extend this body of work by connecting multiple statistical notions of fairness to different adversarial objectives, and Adel et al. (2019) demonstrate a state-of-the-art method to learn fair representations using only one extra hidden layer in a neural network. Our setting is similar to fair representation learning in that we also guarantee that a data collector can safely use the information that is shared. However, where previous works aim to share an abstract representation of the data, our goal is to dynamically collect a subset of raw features that mitigates disparities while also allowing people to use the data, and interpret and justify downstream decisions.

Active feature-value acquisition Different from *active learning*, active feature-value acquisition (AFA) is concerned with feature-wise active learning for each instance. AFA systems are of great need in cost-sensitive applications where the data collector needs to balance an available information budget with predictive accuracy. Importantly, AFA is more efficient than population-level feature selection methods such as LASSO, and forward and backward selection as it considers differences across instances instead of requiring a single predictor Krishnapuram et al. (2011). A number of strategies have been proposed for determining which feature to select next, based on the features that are already collected Kanani & Melville (2008); Krishnapuram et al. (2011). A recent approach uses a variational autoencoder to represent the set of already acquired features and combines it with an acquisition function that maximizes expected information gain Ma et al. (2019). Though effective, these methods do not explicitly model a stopping criterion that determines when to stop collecting additional features and predict the final label. Hence, to effectively trade-off fairness and accuracy, we need a unified framework that jointly optimizes the acquisition strategy and the stopping criterion. We build on a method from Shim et al. (2018) and model the feature acquisition process as a Markov decision process (MDP) where the action space consists of the set of unselected set of features and an additional STOP action which, upon selection, terminates the acquisition process. Our framework differs in two key ways. First, to ensure fairness, we add a second network that predicts the sensitive attribute and change the reward function to balance low classification loss with a high adversarial loss. Second, we model the reward as the change of the combined loss at each timestep as opposed to only having a single classification reward at the end of the episode. This results in faster convergence and better policies.

3 Adversarial Discovery of Fair Information

Problem setup The setup of our framework most follows the joint active feature acquisition and classification framework of (Shim et al., 2018). Let $(\mathbf{x}^{(i)}, y^{(i)}, b^{(i)}) \sim P$ be individual i in P represented by a d-dimensional feature vector $\mathbf{x}^{(i)} \subseteq \mathbb{R}^d$, a binary label $y^{(i)} \in \{0, 1\}$, and a binary sensitive attribute $b^{(i)} \in \{0, 1\}$. We acquire the features in sequential order starting with an empty set $\mathcal{O}_0 := \emptyset$ at time $t = 0$. At every later timestep t, we choose a subset of features from the unselected set of features, $\mathcal{S}_t^{(i)} \subseteq \{1, \ldots, d\} \setminus \mathcal{O}_{t-1}^{(i)}$. After each new acquisition step, the classifier has access to feature values in $\mathcal{O}_t^{(i)} := \mathcal{S}_t^{(i)} \cup \mathcal{O}_{t-1}^{(i)}$. We keep acquiring features up to time $T^{(i)}$ when we meet a stopping criterion. At that point, we will classify $\mathbf{x}^{(i)}$ using only the set of features in $\mathcal{O}_T^{(i)}$. Note that, in contrast to population-level feature selection methods like LASSO, AFA allows for personalized sets of features where each $\mathcal{O}_T^{(i)}$ is different. To learn a model that minimizes classification loss while maximizing the loss of the adversary we formulate the following optimization

problem

$$\min_{\psi,\theta,\phi}\max_{\omega,\chi}\frac{1}{|P|}\sum_{i\in P}\gamma\mathcal{L}_C\left(g_\psi(f_\theta(\mathcal{O}_{T,\phi}^{(i)})),y^{(i)}\right)-(1-\gamma)\mathcal{L}_A\left(g_\omega(f_\chi(\mathcal{O}_{T,\phi}^{(i)})),b^{(i)}\right), \tag{1}$$

where $\mathcal{L}_C$ and $\mathcal{L}_A$ are suitable losses for the label classifier and the adversary. The encoder f_θ feeds into a classifier g_ψ for the label prediction while f_χ and g_ω are the encoder and classifier for the sensitive attribute prediction. Hyperparameter $\gamma \in [0,1]$ specifies the desired balance between classification performance and fairness where high (low) values of γ correspond to accurate (fair) decisions.

Markov decision process We define a Markov decision process (MDP) to find the set of features $\mathcal{O}_T^{(i)}$ that minimizes the objective in equation 1. For each episode, the state at time t is represented by the set of selected features $\{x_j\}_{j\in\mathcal{O}_t}$. The size of the state space is 2^d, the powerset of the feature set. At each timestep t, the action space consists of the set of unselected features $\{1,\dots,d\}\setminus\mathcal{O}_{t-1}$ and an additional STOP action which, upon selection, stops the acquisition process after which the rewards are computed. The agent's cumulative reward, computed for individual (i), corresponds to

$$R(\mathcal{O}_T^{(i)}) = -\gamma\mathcal{L}_C\big(\hat{y}_T^{(i)},y^{(i)}\big)+(1-\gamma)\mathcal{L}_A\big(\hat{b}_T^{(i)},b^{(i)}\big) \tag{2}$$

where $\hat{y}_T^{(i)} = g_\psi(f_\theta(\mathcal{O}_T^{(i)}))$ and $\hat{b}_T^{(i)} = g_\omega(f_\chi(\mathcal{O}_T^{(i)}))$. The first reward encourages accurate classification and the second reward encourages independence between the feature set and the sensitive attribute. For both the classifier and adversary losses, $\mathcal{L}_C$ and $\mathcal{L}_A$, we use binary cross-entropy. When a policy π_ϕ^*, parametrized by ϕ, is optimal for this MDP given parameters ψ,θ,ω,χ, the policy π_ϕ^* is also the optimal solution to the objective in equation 1. We proof this in App A.

Finally, in the active feature acquisition framework in (Shim et al., 2018), the cumulative reward is computed and distributed only at the end of each episode. We observe, however, that distributing the rewards only at the end makes it difficult to converge to a good policy when the feature space is sufficiently high dimensional. Hence, to guide the policy, we reward the agent using the difference in combined loss at every later timestep $t > 1$ with $r(\mathcal{O}_t^{(i)}) = -\gamma\mathcal{L}_C\big(\hat{y}_t^{(i)},y^{(i)}\big)+(1-\gamma)\mathcal{L}_A\big(\hat{b}_t^{(i)},b^{(i)}\big)+\gamma\mathcal{L}_C\big(\hat{y}_{t-1}^{(i)},y^{(i)}\big)-(1-\gamma)\mathcal{L}_A\big(\hat{b}_{t-1}^{(i)},b^{(i)}\big)$ while for $t=1$ the reward equals the initial combined loss $r(\mathcal{O}_1^{(i)}) = -\gamma\mathcal{L}_C\big(\hat{y}_1^{(i)},y^{(i)}\big)+(1-\gamma)\mathcal{L}_A\big(\hat{b}_1^{(i)},b^{(i)}\big)$ resulting in equal cumulative reward $\sum_{t=1}^T r(\mathcal{O}_t^{(i)})$=$R(\mathcal{O}_T^{(i)})$.

Generalized framework The generalized framework in Fig. 1 consists of two parts: the first part in Fig. 1(a) seeks to learn a representation of the set of observed features $\mathbf{z}_1^{(i)} = f_\theta(\mathcal{O}_t^{(i)})$ capable of classifying the label $\hat{y}^{(i)} = g_\psi(\mathbf{z}_1^{(i)})$ and estimating the optimal next action $a_t^{(i)} = \pi_\phi(\mathbf{z}_1^{(i)})$. The model has two heads that share the same encoder leading to improved performance over a model with two separate encoders. In parallel, the second network in Fig. 1(b) seeks to learn a related but separate representation $\mathbf{z}_2^{(i)} = f_\chi(\mathcal{O}_t^{(i)})$, which is fed to a classifier g_ω that predicts the sensitive attribute $\hat{b}^{(i)} = g_\psi(\mathbf{z}_2^{(i)})$. While in adversarial representation learning the adversarial loss is backpropagated directly through a gradient reversal layer to update the encoder (Goodfellow et al., 2014; Edwards & Storkey, 2016), our agent learns to fool the adversary by selecting the set of features that maximize the adversarial classification loss.

We realize f_θ, g_ψ, f_χ, g_ω and π_ϕ as neural networks parametrized by θ, ψ, χ, ω, and ϕ, which are optimized using alternating gradient descent steps. To facilitate encoding of partially observed feature sets, we adopt a feature-level set encoder (Vinyals et al., 2015) which maps the set of features to an order-invariant set representation $\mathbf{z}^{(i)}$. The final set embedding $\mathbf{z}_1^{(i)}$ is fed to both the classifier and the policy network while a second independent set embedding $\mathbf{z}_2^{(i)}$ is fed to the adversary. We refer to App B for details on the set encoding process and to App C for implementation details.

4 EXPERIMENTS

DADI seeks to select the subset of features that can be used by data collectors with the assurance that their trained classifiers are both fair and accurate. We use demographic disparity

$|P(\hat{y} = 1 \mid b = 0) - P(\hat{y} = 1 \mid b = 1)|$ as a measure for the degree of unfairness as strict demographic parity is hard to enforce in practice. The performance of the classifier is measured using the area under the receiver operating characteristic curve (AUC) to account for imbalanced label distributions. For all three datasets (the two real-world dataset and the synthetic dataset experiment in App D) we use one-hot encoding for categorical features and standardize numerical features. For the mapping to actions, we combine multiple one-hot encoded binary features that stem from the same categorical feature into a single action (e.g. the binary features *marital=divorced, marital=married* and *marital=single* correspond to a single action that acquires these features simultaneously). We use 6-fold cross validation with a random $70\%/13.3\%/16.7\%$ train/validation/test split.

We evaluate DADI empirically on two real-world datasets. The Adult Income Dataset from the UCI Machine Learning Repository (Lichman et al., 2013) is an often used benchmark dataset for fair classification that comprises 14 demographic and occupational attributes, which translates, after preprocessing and expanding the categorical features, to 98 continuous and binary features and 14 actions for 48,842 individuals. The task corresponds to classifying whether a person's income is above \$50,000 (25% are above). Rows with missing values are omitted resulting in a dataset with 45,222 samples. We use gender as the sensitive attribute, listed as male or female. The Mexican Poverty dataset is extracted from the Mexican household survey 2016, which contains ground-truth household poverty levels and 99 attributes, related to household information such as the number of rooms or the type of heating system (Ibarrarán et al., 2017). The dataset is used in (Noriega-Campero et al., 2019) for fair feature selection, motivated by the real-world example of fair distribution of social programs. The dataset comprises a sample of 70,305 households in Mexico, with 183 continuous and one-hot encoded binary features and 99 actions, allowing us to demonstrate that our method is effective even for high-dimensional datasets. Classification is binary according to the country's official poverty line, with 36% of the households having the label poor. The considered sensitive attribute describes whether the head of the household is a senior citizen.

Fig. 2 shows the results for both datasets. First, we observe that increasing γ, i.e., increasing the relative weight of the classification reward, leads to an increase in both performance and disparity. Naturally, as the adversarial reward becomes less important, the agent will have a stronger incentive to maximize accuracy resulting in the collection of more features. This, in turn, leads to higher disparity. Importantly, however, we observe that while the AUC increases drastically from the start, demographic parity only increases drastically for larger values of γ, allowing for policies that achieve good predictive performance with minimal disparity loss. This conclusion is supported by the graph in the rightmost frame of Fig. 2 where we show the Pareto front along the AUC-disparity trade-off. These results are encouraging as a data collector can strongly mitigate disparity while still making accurate decisions. Finally, we refer to App. E for a detailed analysis on the acquired features for the Adult dataset and to a synthetic loan approval example that demonstrates our method's ability to select personalized fair feature sets.

5 Conclusion and Future Work

A number of recent works have focused on adversarial learning of fair representations. However, the methods underlying these works are ineffective when the data owner is required to share raw features, a key aspect in many use cases where features are collected for both human and machine decision making. To tackle this problem, we propose DADI in which we frame the data owner's choice as a reinforcement learning problem where an agent selects a subset of features while an adversary critiques potentially unfair sets.

Importantly, however, our framework is more generally applicable in settings where a data owner may wish to guard itself against a naive or malicious data collector by sharing only a subset of features. First, by changing the adversarial objective function, (Madras et al., 2018) demonstrate that one can achieve other notions of fairness such as equal opportunity and equal odds. Second, several recent works have formulated adversarial objectives to attain private data representations (Yang et al., 2018; Phan et al., 2019). These objectives could be adopted using DADI to automate dynamic discovery of private information which could be further extended by encoding features in different levels of precision (such as age by year or age by decade), allowing the agent to select the level of precision that maximizes accuracy while minimizing privacy risk. Moreover, adding monetary acquisition costs of features as a penalty at each collection step would allow our agent to holistically trade-off accuracy, information costs, and fairness or privacy (Shim et al., 2018).

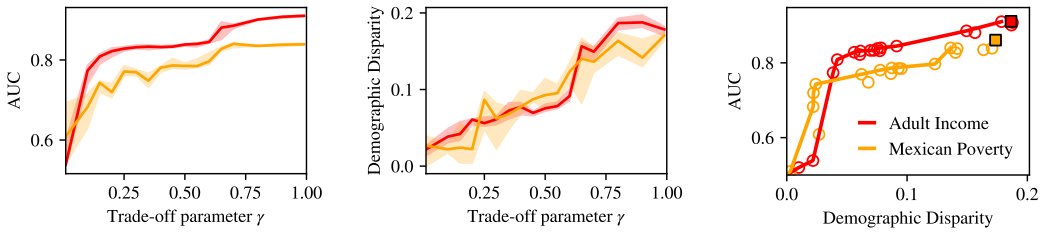


Figure 2: DADI for mitigating demographic disparity across subgroups in the Adult and Mexico datasets. AUC (left) and disparity (center) are given for 18 different values of trade-off parameter γ. The lines are plotted using the mean with 95% confidence intervals computed using 6-fold CV. The rightmost figure shows the Pareto front along the AUC-disparity trade-off, computed using different values of γ (circles). The black outlined squares correspond to the baseline unfair classifiers for which we use the pretrained classifier with access to the full feature set.

REFERENCES

Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. One-network adversarial fairness. In *AAAI Conf. Artif. Intell.*, 2019.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 Workshop on Explainable AI (XAI)*, volume 8, 2017.

Aaron Chalfin, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. Productivity and selection of human capital with machine learning. *American Economic Review*, 106(5), 2016.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innov. Theoret. Comp. Sci. Conf.*, pp. 214–226, 2012.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *Int. Conf. Learn. Represent.*, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Adv. Neur. Inf. Proc. Syst.*, pp. 2672–2680, 2014.

Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang. Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, 33(4):847–856, 2007.

Pablo Ibarrarán, Nadin Medellín, Ferdinando Regalia, Marco Stampini, Sandro Parodi, Luis Tejerina, Pedro Cueva, and Madiery Vásquez. *How Conditional Cash Transfers Work*. Number 8159 in IDB Publications (Books). Inter-American Development Bank, 2017. ISBN ARRAY(0x47b15000). URL https://ideas.repec.org/b/idb/idbbks/8159.html.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Adv. Neur. Inf. Proc. Syst.*, pp. 325–333, 2016.

Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. In *Int. Conf. Mach. Learn.*, pp. 2444–2453, 2018.

Pallika Kanani and Prem Melville. Prediction-time active feature-value acquisition for cost-effective customer targeting. *Adv. Neur. Inf. Proc. Syst.*, 2008.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *Innov. Theoret. Comp. Sci. Conf.*, 2017.

Balaji Krishnapuram, Shipeng Yu, and R Bharat Rao. *Cost-sensitive Machine Learning*. CRC Press, 2011.

Moshe Lichman et al. Uci machine learning repository, 2013.

John Logan. *Diversity and disparities: America enters a new century*. Russell Sage Foundation, 2014.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez-Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. In *Int. Conf. Mach. Learn.*, pp. 4234–4243, 2019.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Int. Conf. Mach. Learn.*, pp. 3381–3390, 2018.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Int. Conf. Mach. Learn.*, pp. 1928–1937, 2016.

Alejandro Noriega-Campero, Michiel Bakker, Bernardo Garcia-Bulle, and Alex Pentland. Active fairness in algorithmic decision making. *AAAI/ACM Conf. Artif. Intell. Ethics Soc.*, 2019.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. ISSN 0036-8075. doi: 10.1126/science.aax2342. URL https://science.sciencemag.org/content/366/6464/447.

NhatHai Phan, Ruoming Jin, My T Thai, Han Hu, and Dejing Dou. Preserving differential privacy in adversarial learning with provable robustness. *arXiv preprint arXiv:1903.09822*, 2019.

Hajin Shim, Sung Ju Hwang, and Eunho Yang. Joint active feature acquisition and classification with variable-size set encoding. In *Adv. Neur. Inf. Proc. Syst.*, pp. 1368–1378, 2018.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *AAAI Conf. Artif. Intell.*, 2016.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *IEEE/ACM Int. Workshop Softw. Fairness*, pp. 1–7, 2018.

Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. *Int. Conf. Learn. Represent.*, 2015.

Tsung-Yen Yang, Christopher Brinton, Prateek Mittal, Mung Chiang, and Andrew Lan. Learning informative and private representations via generative adversarial networks. In *IEEE Int. Conf. Big Data*, pp. 1534–1543, 2018.

A REWARD FUNCTION FOR THE OPTIMAL POLICY

Here we show that, under the optimal policy, maximizing the agent's reward function is equivalent to optimizing the optimization objective (1) in the main text

$$\underset{\psi,\theta,\omega,\chi}{\text{minimize}} \frac{1}{|P|} \sum_{i \in P} \gamma \mathcal{L}_C \left(g_\psi(f_\theta(\mathcal{O}_T^{(i)})), y^{(i)} \right) - (1-\gamma) \mathcal{L}_A \left(g_\omega(f_\chi(\mathcal{O}_T^{(i)})), b^{(i)} \right) \tag{3}$$

Theorem 1. *Consider a policy π_ϕ^*, parametrized by ϕ, that is optimal for this MDP given parameters $\psi, \theta, \omega, \chi$. In that case, policy π_ϕ^* is the optimal solution to the objective in equation 1.*

Proof. Maximizing the cumulative reward in equation 2 over the population P yields

$$\arg\max_\phi \frac{1}{|P|} \sum_{i \in P} -\gamma \mathcal{L}_C \big(g_\psi(f_\theta(\mathcal{O}_{T,\phi}^{(i)}), y^{(i)}) \big) + (1-\gamma) \mathcal{L}_A \big(g_\omega(f_\chi(\mathcal{O}_{T,\phi}^{(i)}), b^{(i)}) \big) \tag{4}$$

$$= \arg\min_\phi \frac{1}{|P|} \sum_{i \in P} \gamma \mathcal{L}_C \big(g_\psi(f_\theta(\mathcal{O}_{T,\phi}^{(i)}), y^{(i)}) \big) - (1-\gamma) \mathcal{L}_A \big(g_\omega(f_\chi(\mathcal{O}_{T,\phi}^{(i)}), b^{(i)}) \big), \tag{5}$$

which is equivalent to the minimization objective in equation 3. □

B SET ENCODER

A set encoder is used to encode arbitrary sets of features. The set encoder was introduced as part of the sequence-to-sequence framework in Vinyals et al. (2015), while the authors in Shim et al. (2018) adopt it for active feature-value acquisition. The set encoder has two parts: a *reading block* and a *processing block*. First, each feature is represented by a vector $\mathbf{u}_j = [x_j \ \mathcal{I}(j)]$ where x_j is the feature-value and $\mathcal{I}(j)$ is a one-hot vector with 1 at position j and zeros elsewhere, allowing the network to incorporate coordinate information. The reading block embeds each vector $\mathbf{u}_j$ onto a memory vector $\mathbf{m}_j$ using a neural network with a shared set of parameters across all features $j \in \{1, \dots, d\}$. The processing block reads the memory (so all memory vectors) into an initial reading vector $\mathbf{r}_0 = \frac{1}{N}\sum_j \mathbf{m}_j$ at processing step 0. This vector $\mathbf{r}_0$ is padded with zeros and fed to an LSTM to compute an initial query vector $\mathbf{q}_0$. At each consecutive time step t an attention weight for each memory vector $\mathbf{m}_i$ is computed using

$$a_{i,t} = \frac{\exp\left(\mathbf{m}_i^T \mathbf{q}_t\right)}{\sum_j \exp\left(\mathbf{m}_j^T \mathbf{q}_t\right)} \tag{6}$$

where $\mathbf{m}_i^T \mathbf{q}_t$ is the dot product of the memory and query vectors. Using the attention vector $\mathbf{a}_t$, we update the reading vector $\mathbf{r}_t = \sum_i a_{i,t}\mathbf{m}_i$ which we concatenate with the query vector and feed to the LSTM to compute the next query vector $\mathbf{q}_{t+1}$ =LSTM($[\mathbf{q}_t \ \mathbf{r}_t]$). In turn, this new query vector is used to compute the new attention vector $\mathbf{a}_t$, We repeat this process for a fixed number of processing steps to achieve a final readout vector $\mathbf{r}_T$, which is subsequently fed to the classifiers and policy network. We refer to Vinyals et al. (2015) for a more detailed description over the encoder and experiments for different number of processing steps. Note that the attention mechanism guarantees that the final readout vector $\mathbf{r}_T$ retrieved from processing is invariant to different permutations of the features in the set.

C ARCHITECTURE AND TRAINING DETAILS

Architecture We use two separate encoders f_θ and f_χ with the same architecture but different parameters. The encoders consist of a memory block, a neural network with two hidden layers of 64-64 units that maps each feature value and its coordinate information to a 32-dimensional real-valued memory vector, and a processing block, an LSTM with 32 hidden units that performs 5 processing steps over the memory to obtain a final read vector. Both classifiers g_ψ and g_ω and the policy network π_ϕ are realized as neural networks with two hidden layers of 64-64 units. The networks share the same architecture for both real-world datasets and use rectified linear units (ReLUs) as activation functions. For the synthetic dataset, we use 16 units for the LSTM and 32-32 units for the encoder, the label classifier, the adversary, and the policy.

Pretraining In the first training phase, we train the encoders f_θ and f_χ, and classifiers g_ψ and g_ω with both the full set of features and randomly missing features. To obtain the partially missing feature sets, we drop each feature with probability $p \sim U(0, 1)$, sampled once for instance to encourage different degrees of sparsity. We train the models using the Adam optimizer with binary cross-entropy loss for 2,000 iterations, a batch size of 64 and a learning rate of 0.0005. In each batch, half of the samples have randomly missing features and half contain the full feature set.

Joint Training In the second training phase, all networks f_θ, f_χ, g_ψ, g_ω, and π_ϕ are trained jointly for 20,000 iterations ($\approx$ 5 million steps). We apply a synchronous variant of n-step Q-Learning (Mnih et al., 2016) where multiple agents run in parallel and collect n-step experiences $(s_t, a_t, r_t, , ..., s_{t+n}, a_{t+n}, r_{t+n})$ using ϵ-greedy exploration, 4 steps, and a discount factor of 1. We decrease ϵ linearly in the first 10,000 iterations from 1 to 0.1. We train with 64 agents in parallel, one agent for one respective instance in the batch. After collecting a running history of n-step experiences, f_θ, f_χ, π_ϕ, g_ψ, and g_ω are jointly updated. The policy network π_ϕ and encoder f_θ are updated using gradient descent by backpropagating the squared loss $(Q(s_t, a_t) - R)^2$. To avoid overestimation and improve stability, we use double Q-learning with $Q(s_{t+n}, \arg\max_a Q(s_{t+n}, a; \phi); \phi')$ as Q estimate while the target Q network $\pi_{\phi'}$ is updated every 100 iterations (Van Hasselt et al., 2016). Q-values corresponding to actions of already acquired features are excluded when taking the argmax to prevent the agent from selecting the same feature twice. The classifiers g_ψ and g_ω, together with encoders f_θ and f_χ, are trained by gradient descent to minimize the (group-normalized)

cross-entropy loss. Each state $\{x_j\}_{j \in \mathcal{O}_t}$ in the history of experiences represents a partial feature set that is used, in combination with the ground-truth label and sensitive attribute, to update the classifier and adversary at each timestep. All networks are trained using the Adam optimizer with learning rate 0.0005.

D SYNTHETIC LOAN APPROVAL EXAMPLE

To illustrate our method's ability to select features that are predictive of the sensitive attribute for only some of the individuals, we generate a synthetic example for loan approval based on (Kallus & Zhou, 2018). Suppose there are two population subgroups such as white and non-white where half of the population is in $b = 0$ and half in $b = 1$. We generate $100,000$ datapoints that are randomly assigned to a subgroup with equal probability. For each individual, we generate a feature vector $X = (X_1, \ldots, X_{10})$, consisting of

- Four fair features $X_1, \ldots, X_4 \overset{iid}{\sim} \mathcal{N}(0, 1)$.
- Four unfair features where the mean is subgroup-dependent: for $b = 0$, $X_5, \ldots, X_8 \overset{iid}{\sim} \mathcal{N}(-1/2, 1)$ and for $b = 1$, $X_5, \ldots, X_8 \overset{iid}{\sim} \mathcal{N}(1/2, 1)$.
- A categorical fair feature with three categories $X_9 \overset{iid}{\sim} \mathcal{U}\{0, 1, 2\}$. X_9 is one-hot encoded into three binary dummy variables that indicate the category.
- A conditionally fair feature where the degree of unfairness is determined by X_9. For $b = 0$, $X_{10} \overset{iid}{\sim} \mathcal{N}(-X_9/2, 1)$ and for $b = 1$, $X_{10} \overset{iid}{\sim} \mathcal{N}(X_9/2, 1)$. When $X_9 = 0$ the feature is fair, while for $X_9 = 1$ and $X_9 = 2$ the feature is unfair.

The label $Y \in \{0, 1\}$, indicating whether or not a individual will pay back the loan if approved, is logistic in X with $P(Y = 1 \mid X, B) = \sigma(\beta^T X/10)$ where $\sigma(t) = 1/(1 + e^{-t})$ and β is a vector with alternating 1 and -1.

To make the task more challenging, features X_4 an X_8 are unobserved while the other features are used to train the policy, the classifier, and the adversary. The trade-off parameter is set to $\gamma = 0.5$ such that the classifier and adversary losses contribute equally to the reward. In the top left frame of Fig. 3, we show the predictiveness of the full feature set and of each feature in isolation by comparing the AUC of the classifier and adversary when only the corresponding feature is fed to the set encoder. As expected, in isolation, the fair and categorical features are independent of the sensitive attribute ($AUC = 0.5$) while all features are predictive for the label y ($AUC > 0.5$).

The probability that a feature is selected for individuals in each of the three categories of X_9 is shown in the bottom left frame in Fig. 3. First, the features that are independent of the sensitive attribute are always selected, while the three unfair features are never selected. The selection probability of the conditionally fair feature is highest when $X_9 = 0$ (when feature X_{10} is fair and thus safe to share) and lowest for $X_9 = 2$ (when feature X_{10} is unfair and should not be shared). This demonstrates that the agent, in contrast to population-level feature selection methods, can exploit conditional dependencies in the dataset to perform fair feature selection at the individual level.

E FEATURE-LEVEL ANALYSIS FOR THE ADULT INCOME DATASET

In the two graphs on the right of Fig. 3 we see, for three different fairness-accuracy trade-offs ($\gamma \in \{0.1, 0.5, 0.9\}$), an overview of the probability that each feature is selected as well as the AUC of the classifier and adversary when only a single feature is fed to the encoder. As expected, features that are predictive of the label but not of the sensitive attribute such as 'education-num', 'age' and 'capital-gain' are selected frequently for all three values of γ. However, features for which the classifier and adversary AUC is comparable ('marital-status', 'occupation', hours-per-week', 'native country', and 'race') are selected only when the classifier loss is weighted more heavily in the reward ($\gamma = 0.9$). Finally, the 'relationship' feature which reveals the gender of an individual's partner, is never selected as this feature alone allows one to predict the sensitive attribute with an AUC comparable to when we use all features (AUC $= 0.86$ when only using 'relationship' while AUC $= 0.87$ when using all features).

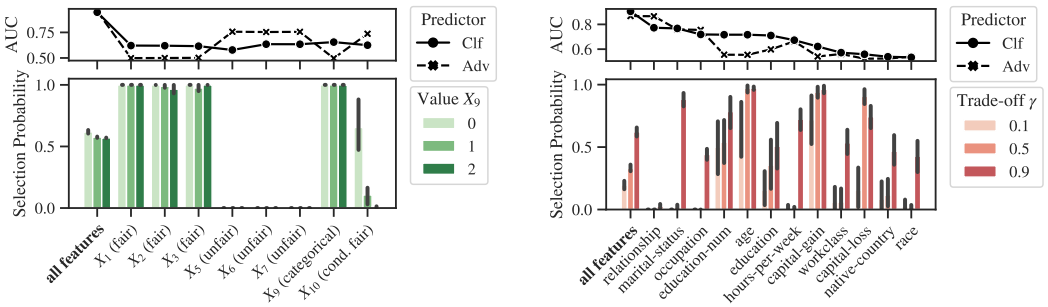


Figure 3: AUC (top) of the classifier and adversary and selection probability (bottom) for each feature in the Synthetic (left) and Adult Income (right) datasets. The AUC for each feature, averaged using 6-fold CV, is computed by feeding only the corresponding feature to the pretrained classifier and the adversary. Note that, if the feature is categorical, all features that stem from the original feature are used. The full feature set is used for the values in the 'all features' column. The selection probability, averaged using 6-fold CV with 95% confidence intervals, corresponds to the fraction of test set examples for which the feature is queried. For the synthetic dataset, the shaded subgroups represent the three values of the categorical variable ($X_9 \in \{0, 1, 2\}$) while, for the Adult Income dataset, the subgroups represent three different fairness-accuracy trade-offs ($\gamma \in \{0.1, 0.5, 0.9\}$).