Dynamic Positive Reinforcement For Long-Term Fairness

Bhagyashree Puranik [1] **Upamanyu Madhow** [1] **Ramtin Pedarsani** [1]

Abstract

As AI-based decision-making becomes increasingly impactful on human society, the study of the influence of fairness-aware policies on the population becomes important. In this work, we propose a framework for sequential decision-making aimed at dynamically influencing long-term societal fairness, illustrated via the problem of selecting applicants from a pool consisting of two groups, one of which is under-represented. We consider a dynamic model for the composition of the applicant pool, where the admission of more applicants from a particular group positively reinforces more such candidates to participate in the selection process. Under such a model, we show the efficacy of the proposed *Fair-Greedy* selection policy which systematically trades greedy score maximization against fairness objectives. In addition to experimenting on synthetic data, we adapt static real-world datasets on law school candidates and credit lending to simulate the dynamics of the composition of the applicant pool.

1. Introduction

In this paper, we seek to develop a framework for sequential decision making aimed at influencing long-term societal fairness. Machine learning models are being increasingly applied in making critical decisions that affect humans, such as recidivism prediction (Dressel & Farid, 2018), mortgage lending (Berkovec et al., 2018), and recommendation systems (Yao & Huang, 2017). While the algorithms offer increased efficiency, speed, and scalability, they could introduce bias leading to the decisions being unfair towards certain groups of the population. There is a rich and rapidly growing literature on "fair" strategies that mitigate bias in algorithmic decision making, including label or data preprocessing and cost reweighting based on groups (Kamiran & Calders, 2012), addition of constraints that satisfy fairness criteria (Zafar et al., 2017), and learning representations that obfuscate group information (Zemel et al., 2013).

Modeling the long-term impacts of dynamic decision-making have been traditionally investigated using reinforcement learning frameworks via Markov Decision Processes (MDPs) and introducing fairness constraints in the reward functions (Wen et al., 2021; Ghalme et al., 2021; Jabbari et al., 2017; Chen et al., 2020; Patil et al., 2020; Joseph et al., 2018; Heidari & Krause, 2018; Gillen et al., 2018). The importance of introducing dynamics into notions of fairness is highlighted by (Liu et al., 2018), showing that static fairness criteria may lead to undesired long-term effects on minority groups. Prior works on the long-term effects of fairness such as (Zhang et al., 2019; 2020; Williams & Kolter, 2019; Mouzannar et al., 2019) have focused, either explicitly or implicitly, on the impact of decisions on the qualifications or score distributions of the different groups. We adopt an outlook complementary to the preceding body of work, seeking to influence the participation of under-represented groups in the selection process. Rather than studying the impact of fair policies, we provide a generic framework for achieving long-term fairness dynamically.

Our framework is motivated by real-world examples such as the following. Consider a company receiving applications every month, which wants to hire in an unbiased manner (e.g., by ultimately selecting equal numbers of male and female applicants). With the total intake fixed based on a budget, the company selects a certain proportion of candidates from each group. The hiring decisions affect the subsequent pool of applicants: admitting more candidates from a particular group might encourage more such candidates to apply, or successful candidates from a group might inspire other such candidates, providing positive feedback into the decision-making loop. Such a strategy could not only enhance diversity and equity, but also enable the company to learn more about a minority group so as to eventually have a richer pool of well-qualified applicants. Another motivating example is college admissions, where the goal may be to admit students with the best academic records, while accounting for socio-economic background and reducing bias based on sensitive attributes such as race or gender. Could one, for example, reverse the trend in the decrease in the proportion of women in STEM as documented in (Broad &

[1]Department of Electrical and Computer Engineering, University of California Santa Barbara, USA. Correspondence to: Bhagyashree Puranik <bpuranik@ucsb.edu>.

ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA, 2022.

McGee, 2014)? It reported that 18% of bachelor's degrees in computer science were awarded to women in 2010, down from 37% in 1985. We suggest here a structured framework for fair selection aimed at combating such systemic imbalances by encouraging a larger number of people from minority groups to participate in the selection process.

Contributions Based on a simple model for evolution of the composition of the applicant pool, we develop a framework for fair selection by formulating the problem as a Markov Decision Process (MDP) with two objectives – maximizing the utility by admitting candidates with the highest scores, and minimizing the disparity between the proportions of selected candidates from each group. We present two policies for fair selection: an optimal policy based on value iteration that maximizes the utility accumulated over multiple rounds, and a computationally simpler *Fair-Greedy (FG)* policy. We characterize the structure of the FG policy, show convergence and also prove that the applicant pool proportion approaches the target proportion that is desired by the system under identical score distributions across the groups. When the score distributions are distinct, we provide experimental evidence of convergence of the applicant pool proportion. We illustrate the efficacy of our approach with experiments with synthetic data, as well as with dynamic data created from the static datasets.

2. MDP formulation and Fair-Greedy Policy

Given that there are two groups u and v within the population, based on a binary valued sensitive attribute, we denote the total number of applicants in round t by N_t, out of which N_t^u belong to group u and $N_t^v = N_t - N_t^u$ belong to group v, based on a binary valued sensitive attribute. We wish to admit a fixed proportion $\bar{a}$ of the total applicants, leading to $A_t = \bar{a}N_t$ number of total applicants accepted in round t. We denote by A_t^u and $A_t^v = A_t - A_t^u$ the number of applicants selected in round t from groups u and v respectively.

Score distributions The qualification of an applicant is measured by the *score*, assumed to be an increasing function of the proficiency of a candidate. Let $\mathcal{P}_u$ and $\mathcal{P}_v$ denote the score distributions of the two groups. Thus the scores for groups u and v are $\{X_i^u\}_{i=1}^{N_t^u}$ and $\{X_i^v\}_{i=1}^{N_t^v}$, generated from $\mathcal{P}_u$ and $\mathcal{P}_v$ respectively. We denote the ordered scores by $\{X_{(i)}^u\}_{i=1}^{N_t^u}$ and $\{X_{(i)}^v\}_{i=1}^{N_t^v}$, where $X_{(i)}^u$ and $X_{(i)}^v$ denote the i^{th} largest scores out of N_t^u and N_t^v respectively.

Fairness-aware utility The goal is to optimize the *utility*, which comprises of two parts: a greedy term (to be maximized) which is the expected sum of scores of selected candidates, and a fair term (to be minimized) measuring disparity between groups based on a *target* proportion.

MDP formulation We define the MDP state $s_t \in [0, 1]$ as the proportion of applicants from group u out of the total, and the action $a_t \in [0, 1]$ as the proportion of selected candidates from group u out of the total selected candidates:

$$s_t = \frac{N_t^u}{N_t}, a_t = \frac{A_t^u}{A_t}.$$

We denote by $\bar{s} \in (0, 1)$ the long-term target of the proportion of group u among the selected applicants. For example, if group u is under-represented in the applicant pool, we may set $\bar{s}$ as the proportion of group u in society at large. Instead, if our long-term goal is to admit equal number from both groups, we set $\bar{s} = 0.5$. Note that formulating the states and actions as proportions of group u is sufficient since the proportion of applicants and admitted candidates from group v is naturally $1 - s_t$ and $1 - a_t$ respectively. The overall utility or reward is:

$$R(s_t, a_t) = R_{\mathcal{G}}(s_t, a_t) - \lambda L_{\mathcal{F}}(a_t), \tag{1}$$

where the *greedy* reward term is the expected sum of ordered scores of admitted candidates, given by:

$$\begin{aligned} R_{\mathcal{G}}(s_t, a_t) &= \frac{1}{A_t}\mathbb{E}\Bigg[\sum_{i=1}^{A_t^u} X_{(i)}^u + \sum_{i=1}^{A_t^v} X_{(i)}^v\Bigg] \\ &= \frac{1}{A_t}\mathbb{E}\Bigg[\sum_{i=1}^{a_t A_t} X_{(i)}^u + \sum_{i=1}^{(1-a_t)A_t} X_{(i)}^v\Bigg], \end{aligned}$$

and the *fairness* loss term is

$$L_{\mathcal{F}}(a_t) = (a_t - \bar{s})^2. \tag{2}$$

In (1), $\lambda \geq 0$ is a parameter used to control the weight given to the fairness objective relative to the greedy objective. The greedy objective promotes the admission of *good* candidates, while the fairness objective promotes fairness in selection proportion. Note that the fairness objective is balanced: it pushes the selection proportion towards $\bar{s}$ regardless of whether group u is under-represented or over-represented among the selected applicants.

Applicant pool evolution We model the positive reinforcement provided by our decision making as a set of transition probabilities $\mathcal{P}(s_{t+1}|s_t, a_t)$. We consider a model where the total number of applicants N_t to the system at round t can be any sequence of numbers and the number of applicants from group u to the system is sampled from a Poisson distribution based on the mean parameter and overall number of applicants (which is variable) as

$$N_t^u \sim Pois(\theta_t N_t), \tag{3}$$

where $Pois(\cdot)$ is the Poisson distribution with mean $\theta_t N_t$. Thus, θ_t is the mean proportion of group u in the applicant pool in round t. We consider the following simple model for positive reinforcement:

$$\theta_{t+1} = [\theta_t + \eta(a_t - s_t)]_{\mathcal{C}}, \tag{4}$$

where η is a step-size parameter and $[\cdot]_{\mathcal{C}}$ is the projection on the convex set $\mathcal{C} = [0, 1]$. Thus the update is such that when the admission rate a_t of the group u is higher than the application rate s_t, more applicants from the group are encouraged in future rounds, and vice versa. The state then evolves as

$$s_{t+1} = \frac{N^u_{t+1}}{N_{t+1}}.$$

The model for positive reinforcement is relevant to many real-world selection systems and is inspired by the social behavior that the successful admission of candidates from a particular group encourages more such candidates to apply to the institution. For instance, a large number of female college graduates in society serve as role-models, encouraging the future generations of women to go to college. However, if a particular program is known for admitting women at a rate smaller than the application rate, lesser women might consider the institution as worth applying to.

Optimal policy The optimal policy $\pi^*(s)$ for the preceding MDP can be found through dynamic programming, by constructing value functions (Bertsekas, 2007) and iteratively solving the Bellman equation. It is also known that the value iteration algorithm converges as long as the reward is bounded in magnitude (Bertsekas, 2007). However, analyzing the equilibrium state of the MDP under this optimal policy is intractable. We observe through simulations that the structure of the optimal policy $\pi^*(s)$ is similar to that of the simpler *Fair-Greedy* policy proposed next, and that the applicant pool evolution converges to an equilibrium point.

Fair-Greedy policy Finding an optimal policy is computationally expensive as the state space grows larger. We therefore propose a simple, yet effective, *Fair-Greedy* (FG) policy that optimizes the instantaneous overall utility in (1):

$$\pi^*_{FG}(s) = \arg\max_a R(s, a). \tag{5}$$

We provide insight into this policy by considering its performance for a large applicant pool (N_t large) with identical score distributions across the two groups. In this regime, we first prove that the greedy reward term is optimized when the admission proportion is the same as the applicant proportion. We then derive some key properties of the FG policy, and provide theoretical guarantees for the convergence of the applicant pool to the target proportion. We observe through simulations that when score distributions are non-identical, the applicant pool converges to an equilibrium point under the FG policy. Please refer to Appendix A for detailed proofs.

Theorem 2.1. *If the score distributions $\mathcal{P}_u$ and $\mathcal{P}_v$ of the two groups are identical, under the regime of large N_t, the greedy reward $R_{\mathcal{G}}(s_t, a_t)$ is optimized by the action:*

$$a^*_{\mathcal{G}} = \arg\max_{a_t} R_{\mathcal{G}}(s_t, a_t) = s_t. \tag{6}$$

Theorem 2.2. *For identical score distributions across the groups, the Fair-Greedy policy satisfies the following: (a) $s_t < \pi^*_{FG}(s_t) < \bar{s}$, if $s_t < \bar{s}$; (b) $\bar{s} < \pi^*_{FG}(s_t) < s_t$, if $s_t > \bar{s}$; (c) $\pi^*_{FG}(s_t) = \bar{s}$, if $s_t = \bar{s}$. Furthermore, if the step-size η_t decays with time and satisfies the conditions (i) $\sum_t \eta_t = \infty$ and (ii) $\sum_t \eta_t^2 < \infty$, the applicant pool proportion converges to the target proportion $\bar{s}$. This implies that the admission or action at equilibrium also approaches the societal or target proportion, in the asymptotic regime that the total applicants in every round are large.*

3. Experimental evaluation

FG policy on synthetic data: We begin by evaluating our framework with synthetic Gaussian datasets. In the first experiment, we set the target proportion $\bar{s} = 0.4$ and the selection rate $\bar{a} = 0.3$ (i.e., we aim to select 30% of the candidates who have applied). We assume identical Gaussian score distributions for the groups with means $\mu_u = \mu_v = 5$ and variances $\sigma_u^2 = \sigma_v^2 = 1$. The step-size is fixed as $\eta = 0.05$. Figure 1(a) shows the convergence of the applicant pool to the target proportion of 40% as guaranteed by our analysis. The framework is capable of handling an inversion in the majority and minority proportions as supported by the evolutions shown from two distinct initial applicant mean proportion parameters $\theta_0 = 0.1$ and $\theta_0 = 0.9$. We report on the dynamics for the proportion of applicants and admitted candidates for individual sample paths in which the number of applicants is randomly drawn as in (3). We do not smooth over multiple sample paths in such figures because our objective is to highlight the convergence of the mean parameter θ_t over each sample path. Note that tuning of the hyperparameter λ is not required when score distributions are identical (here we set $\lambda = 2$). As long as $\lambda > 0$, the applicant pool converges to the target proportion, with only the rate of convergence increasing with λ, as we depict in Figure 1(b). Next, we focus on a setting where the underprivileged class u has larger variance, but slightly smaller mean ($\sigma_u^2 = 1.5$, $\mu_u = 4.9$). We set $\bar{s} = 0.4$, and consider a more selective process, with $\bar{a} = 0.1$. From Figure 1(c), we note that the applicant mean and also the group admission converges to a proportion larger than $\bar{s}$. This is due to the fact that as the admission rate gets selective, the greedy part of the reward is optimized by an action that admits more from the group with longer tail (larger variance). This is also evident in Figure 1(d), where we observe that for smaller values of λ, i.e., when more weight is assigned to the greedy reward, the mean parameter θ_t converges to larger values. However with enough weight being given to fairness, the applicant pool still converges to the desired ratio.

FG policy on real-world datasets: We simulate the dynamics by considering the following: (i) the law school (LS) (Wightman, 1998) bar exam dataset found at (git, 2018), applying our framework for selecting candidates who

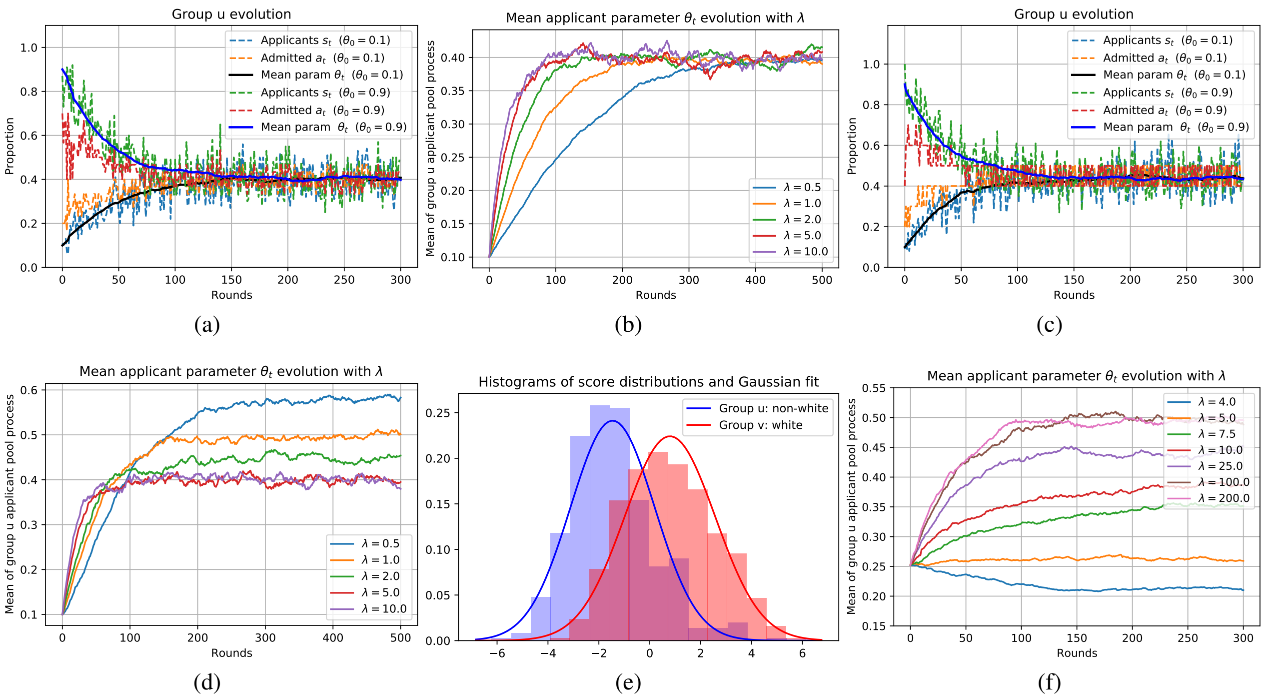

Figure 1: (a) FG policy under identical score distribution across groups, showing convergence from distinct initial mean parameters $\theta_0 = 0.1, 0.9$. (b) Applicant pool converges to target under identical score distributions. (c) FG policy under selective system, lower mean and larger variance for group u. Shows convergence from $\theta_0 = 0.1, 0.9$. (d) Applicant pool convergence for the selective system under FG policy. (e) Histograms and Gaussian fit for score distributions of the law school dataset (f) Applicant pool evolution with $\theta_0 = 0.25$, with varying λ for the law school dataset.

Dataset	μ_u	μ_v	σ_u^2	σ_v^2
LS bar study	-1.46	0.79	2.73	3.16
German credit	0.32	0.85	1.93	2.06

Table 1: Gaussian score distribution parameters

are likely to be successful in the bar exam, based on features such as LSAT scores, undergraduate GPA, law school GPA and others, with race as the sensitive attribute; (ii) the German credit dataset (Dua & Graff, 2017) with gender as the sensitive attribute, where the motivation is to encourage higher levels of participation of women in the financial lending system. From the raw datasets, we calculate the score distributions by fitting a logistic regression based predictor. The histograms of the scores of the two groups resemble the Gaussian distribution. We fit a Gaussian for each of the histograms, to obtain the mean and variance parameters of the score distributions $\mathcal{P}_u$ and $\mathcal{P}_v$, listed in Table 2.

In the experiments with LS dataset we set selection rate $\bar{a} = 0.3$, and target proportion $\bar{s} = 0.5$. Figure 1(e) depicts the distinct group-wise score distributions for the LS dataset, and Figure 1(f) shows the convergence of the applicant pool by plotting the mean parameter θ_t for various settings for the hyperparameter λ. The initial mean for the applicant proportion is set to $\theta_0 = 0.25$, based on the proportion of non-white samples in the dataset. Since the mean of scores of the underprivileged group is smaller and the application is not very selective, the utility is maximized by admitting more from the privileged group. However, as the importance of fairness is increased through λ, the applicant pool and admission rate both approach 50%. We defer the plots for the German dataset and other experimental details to Appendix B, but remark that since the score distributions for the German dataset are *closer*, the value of λ required to achieve similar fairness target is smaller than that for the law school dataset.

4. Conclusion

In this paper, we propose a framework for fair selection of applicants to a system, and study the long-term effects of decisions on the dynamics of the applicant pool. Our results indicate the potential of achieving long-term fairness objectives through positive reinforcement via decision making. We hope that this work stimulates the collaboration between machine learning researchers and social scientists required for these ideas to make real-world impact. A key future direction is to devise and conduct experiments for measuring, understanding and shaping the evolution dynamics posited in our framework.

Acknowledgements

This work was supported by the Army Research Office under grant W911NF-19-1-0053, and by the National Science Foundation under grants CCF 1909320 and CCF 2224263.

References

Law school bar study dataset, 2018. URL https://github.com/algowatchpenn/GerryFair/blob/master/dataset/lawschool.csv.

Berkovec, J. A., Canner, G. B., Gabriel, S. A., and Hannan, T. H. Mortgage discrimination and fha loan performance. In *Mortgage Lending, Racial Discrimination, and Federal Policy*, pp. 289–305. Routledge, 2018.

Bertsekas, D. *Dynamic programming and optimal control: Volume II*, volume 2. Athena scientific, 2007.

Broad, S. and McGee, M. Recruiting women into computer science and information systems. *Association Supporting Computer Users in Education*, 2014.

Chen, Y., Cuellar, A., Luo, H., Modi, J., Nemlekar, H., and Nikolaidis, S. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, pp. 181–190. PMLR, 2020.

Dressel, J. and Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Ghalme, G., Nair, V., Patil, V., and Zhou, Y. State-visitation fairness in average-reward mdps. *arXiv preprint arXiv:2102.07120*, 2021.

Gillen, S., Jung, C., Kearns, M., and Roth, A. Online learning with an unknown fairness metric. *Advances in neural information processing systems*, 31, 2018.

Heidari, H. and Krause, A. Preventing disparate treatment in sequential decision making. In *IJCAI*, pp. 2248–2254, 2018.

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. Fairness in reinforcement learning. In *International conference on machine learning*, pp. 1617–1626. PMLR, 2017.

Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 158–163, 2018.

Kamiran, F. and Calders, T. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR, 2018.

Mouzannar, H., Ohannessian, M. I., and Srebro, N. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 359–368, 2019.

Patil, V., Ghalme, G., Nair, V., and Narahari, Y. Achieving fairness in the stochastic multi-armed bandit problem. In *AAAI*, pp. 5379–5386, 2020.

Wen, M., Bastani, O., and Topcu, U. Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics*, pp. 1144–1152. PMLR, 2021.

Wightman, L. F. Lsac national longitudinal bar passage study. lsac research report series. 1998.

Williams, J. and Kolter, J. Z. Dynamic modeling and equilibria in fair decision making. *arXiv preprint arXiv:1911.06837*, 2019.

Yao, S. and Huang, B. Beyond parity: Fairness objectives for collaborative filtering. *Advances in neural information processing systems*, 30, 2017.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017.

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

Zhang, X., Khaliligarekani, M., Tekin, C., and Liu, M. Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhang, X., Tu, R., Liu, Y., Liu, M., Kjellstrom, H., Zhang, K., and Zhang, C. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33:18457–18469, 2020.

A. Proofs

We restate and prove the theorems on the optimality of greedy reward and the convergence of the applicant pool under identical score distributions in this section.

Theorem A.1. *If the score distributions $\mathcal{P}_u$ and $\mathcal{P}_v$ of the two groups are identical, the greedy reward $R_{\mathcal{G}}(s_t, a_t)$ is optimized by the action:*

$$a_{\mathcal{G}}^* = \arg\max_{a_t} R_{\mathcal{G}}(s_t, a_t) = s_t. \tag{7}$$

Proof. Recall that the greedy reward is given by:

$$R_{\mathcal{G}}(s_t, a_t) = \frac{1}{A_t}\mathbb{E}\Bigg[\sum_{i=1}^{A_t^u} X_{(i)}^u + \sum_{i=1}^{A_t^v} X_{(i)}^v\Bigg] \tag{8}$$

Since we assume the space of actions as $a_t \in [0, 1]$, the number of admitted candidates from each group, more formally, are $A_t^u = \lfloor a_t A_t \rfloor$ and $A_t^v = \lfloor (1 - a_t)A_t \rfloor$. For simplicity of presentation, we omit the 'floor' without loss of generality of our results since we are interested in the regime that N_t is large. Therefore, we write:

$$R_{\mathcal{G}}(s_t, a_t) \quad = \quad a_t\mathbb{E}\left[\frac{\sum_{i=1}^{a_t A_t} X_{(i)}^u}{a_t A_t}\right] + (1 - a_t)\mathbb{E}\left[\frac{\sum_{i=1}^{(1-a_t)A_t} X_{(i)}^v}{(1 - a_t)A_t}\right]$$

By the law of large numbers, the collection of score variables $\{X_i^u\}_{i=1}^{N_t^u}$ and $\{X_i^v\}_{i=1}^{N_t^v}$ converge to their respective distributions $\mathcal{P}_u$ and $\mathcal{P}_v$ as N_t increases. Choosing the top $A_t^u = a_t A_t$ candidates out of N_t^u (similarly top A_t^v out of N_t^v) is equivalent to setting a threshold t_u (similarly, t_v) and admitting all candidates with scores above the threshold. This holds for generic score distributions and they need not necessarily be identical across the groups. Thus for large N_t, the average score of the admitted candidates from each group approaches its expected value as:

$$\lim_{N_t \to \infty} \frac{\sum_{i=1}^{a_t A_t} X_{(i)}^u}{a_t A_t} \quad = \quad \mathbb{E}[X^u | X^u \geq t_u] \tag{9}$$

$$\lim_{N_t \to \infty} \frac{\sum_{i=1}^{(1-a_t)A_t} X_{(i)}^v}{(1 - a_t)A_t} \quad = \quad \mathbb{E}[X^v | X^v \geq t_v] \tag{10}$$

Rewriting the greedy reward in terms of the above conditional expectations leads to the following equation:

$$R_{\mathcal{G}}(s_t, a_t) = a_t \frac{\int_{t_u}^{\infty} u\mathcal{P}_u(u)du}{\int_{t_u}^{\infty} \mathcal{P}_u(u)du} + (1 - a_t)\frac{\int_{t_v}^{\infty} v\mathcal{P}_v(v)dv}{\int_{t_v}^{\infty} \mathcal{P}_v(v)dv} \tag{11}$$

with the additional constraint being that the thresholds t_u and t_v are such that the total number of admitted candidates is equal to $A_t = \bar{a}N_t$. Note that t_u and t_v depend on the current state s_t and action a_t.

Since the acceptance is decided by a group-wise threshold, the fraction of applicants from a group who are admitted is precisely determined by the area under its score distribution beyond the threshold. Formalizing the above, for large N_t, we have:

$$\int_{t_u}^{\infty} \mathcal{P}_u(u)du = 1 - F_u(t_u) = \frac{a_t A_t}{s_t N_t}$$

$$\int_{t_v}^{\infty} \mathcal{P}_v(v)dv = 1 - F_v(t_v) = \frac{(1 - a_t)A_t}{(1 - s_t)N_t}.$$

and the constraint on the total number of candidates admitted can now be expressed through the following equivalent statements:

$$a_t A_t + (1 - a_t)A_t \quad = \quad \bar{a}N_t$$

$$s_t N_t(1 - F_u(t_u)) + (1 - s_t)N_t(1 - F_v(t_v)) \quad = \quad \bar{a}N_t,$$

and finally, we have:

$$s_t N_t \int_{t_u}^{\infty} \mathcal{P}_u(u)du + (1-s_t)N_t \int_{t_v}^{\infty} \mathcal{P}_v(v)dv = \bar{a}N_t. \tag{12}$$

Let us now consider the maximization of the greedy reward. Given state s_t, and generic distributions $\mathcal{P}_u$ and $\mathcal{P}_v$, we need to set the thresholds t_u and t_v for the respective groups such that the sum of scores of all admitted candidates is maximized. We show by contradiction that to maximize the greedy reward, we require $t_u = t_v$.

Assume a pair of thresholds (t_u, t_v) that result in the maximization of the greedy reward, and $t_u < t_v$. Let us denote the expected sum of scores of the admitted candidates by $S(t_u, t_v)$, which is the optimum. One can construct thresholds $t'_u = t_u + \epsilon_1$ and $t'_v = t_v - \epsilon_2$ (where $\epsilon_1, \epsilon_2 > 0$, infinitesimally small for large N_t), such that we admit one more candidate from group v (as a result of the decreased threshold) and one less from group u (as a result of the increased threshold) as compared to the case with thresholds (t_u, t_v). As long as $t'_v > t'_u$, we have $S(t'_u, t'_v) > S(t_u, t_v)$, which contradicts the assumption that (t_u, t_v) maximize the greedy reward. Similarly, if we begin with a pair of optimal (t_u, t_v) such that $t_u > t_v$, we can construct thresholds $t'_u = t_u - \epsilon_3$ and $t'_v = t_v + \epsilon_4$, so that we admit one more candidate from group u and one less from group v. As long as $t'_u > t'_v$, we arrive at the contradiction $S(t'_u, t'_v) > S(t_u, t_v)$. Thus the greedy reward is optimized when thresholds across the groups are equal, irrespective of the nature of $\mathcal{P}_u$ and $\mathcal{P}_v$.

Thus, for arbitrary score distributions, the action that maximizes the greedy reward is such that:

$$\begin{aligned} t_u &= t_v \\ \implies F_u^{-1}\left(1 - \frac{a_t A_t}{s_t N_t}\right) &= F_v^{-1}\left(1 - \frac{(1-a_t)A_t}{(1-s_t)N_t}\right) \end{aligned} \tag{13}$$

If $\mathcal{P}_u$ and $\mathcal{P}_v$ are identical, the arguments of the inverse CDFs in (13) need to be equal. Thus the optimal action should be such that:

$$\begin{aligned} 1 - \frac{a_t A_t}{s_t N_t} &= 1 - \frac{(1-a_t)A_t}{(1-s_t)N_t} \\ \implies a_t &= s_t. \end{aligned}$$

Thus, the greedy reward is maximized by choosing the admission proportion of group u to be same as the applicant proportion of group u:

$$a^*_{\mathcal{G}} = s_t.$$

□

Employing theorem A.1, we arrive at the the following theorem which informs us about the convergence of the applicant pool and characterizes the FG policy.

Theorem A.2. *For identical score distributions across the groups, the Fair-Greedy policy satisfies the following properties:*

$$\begin{aligned} s_t &< \pi^*_{FG}(s_t) < \bar{s}, \text{ if } s_t < \bar{s} \\ \bar{s} &< \pi^*_{FG}(s_t) < s_t, \text{ if } s_t > \bar{s} \\ & \pi^*_{FG}(s_t) = \bar{s}, \text{ if } s_t = \bar{s} \end{aligned}$$

Furthermore, if the step-size η_t decays with time and satisfies the conditions (i) $\sum_t \eta_t = \infty$ and (ii) $\sum_t \eta_t^2 < \infty$, the applicant pool proportion converges to the target proportion $\bar{s}$. This implies that the admission or action at equilibrium also approaches the societal or target proportion, in the asymptotic regime that the total applicants in every round are large.

Proof. Under the FG policy, $a_t = \pi^*_{FG}(s_t)$. The applicant pool update for the mean parameter is:

$$\theta_{t+1} = [\theta_t + \eta(\pi^*_{FG}(s_t) - s_t)]_{\mathcal{C}}. \tag{14}$$

The fairness loss in (2) is minimized when the admission proportion is same as the target, formalized as:

$$a^*_{\mathcal{F}} = \arg\min_{a_t} L_{\mathcal{F}}(a_t) = \bar{s}$$

The overall reward $R(s_t, a_t)$ is a sum of the greedy reward and fairness loss (scaled by λ). The fairness loss is convex (hence $-L_{\mathcal{F}}(a_t)$ is concave) in a_t. It can be seen that the greedy reward monotonically decreases in either directions around $a_t = s_t$, and in addition it possesses continuity in a_t. When at state s_t, suppose the optimal action a^* of the FG policy is such that $a^* < s_t$, when $s_t < \bar{s}$. Then by continuity and since the greedy reward is maximized at s_t, $\exists$ some $a' > s_t$, such that $R_{\mathcal{G}}(s_t, a') \geq R_{\mathcal{G}}(s_t, a^*)$, and moreover has a smaller fairness loss, i.e., $L_{\mathcal{F}}(a') < L_{\mathcal{F}}(a^*)$, which violates the optimality of a^*. Thus the optimal action for the FG policy must be $a^* > s_t$, if $s_t < \bar{s}$. Similar arguments hold if $s_t > \bar{s}$, and here we can show that the optimal action must be such that $a^* < s_t$. Hence, it follows that the optimal action for overall utility lies between the optimal actions for greedy and fairness terms:

$$s_t < \pi^*_{FG}(s_t) < \bar{s}, \text{ if } s_t < \bar{s} \tag{15}$$

$$\bar{s} < \pi^*_{FG}(s_t) < s_t, \text{ if } s_t > \bar{s} \tag{16}$$

$$\pi^*_{FG}(s_t) = \bar{s}, \text{ if } s_t = \bar{s} \tag{17}$$

Now we show the convergence of the applicant pool to its equilibrium. Let us consider a step-size that decays with time such that $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$. Consider the case when $s_t < \bar{s}$, where we have: $s_t < \pi^*_{FG}(s_t) < \bar{s}$. From (14), we can see that the mean proportion parameter θ_{t+1} increases. Similarly, when $s_t > \bar{s}$, it follows that $\bar{s} < \pi^*_{FG}(s_t) < s_t$, and the mean proportion parameter decreases. Note that the target proportion is a fixed point of the FG policy, i.e., $\pi^*_{FG}(\bar{s}) = \bar{s}$. Due to the above characterization of $\pi^*_{FG}(s_t)$ and the model for the update of the applicant pool, the mean parameter θ_t grows or reduces in the direction of $\bar{s}$. Hence, as the step-size is decaying, one can show that the mean parameter θ_t converges to $\bar{s}$ (see Lemma A.3 for details). Moreover, the variance of the number of group u applicants is $var(N^u_t) = \theta_t N_t$ due to the Poisson distribution. Thus, the state $s_t = N^u_t / N_t$ has variance $O(1/N_t)$. Consequently, in the asymptotic regime that N_t is large, using Chebyshev's inequality one can show that s_t also converges to θ_t in probability. This implies that the applicant proportion approaches $\bar{s}$, which completes the proof. □

Lemma A.3. *If the step-size η_t decays with time and satisfies the conditions (i) $\sum_t \eta_t = \infty$ and (ii) $\sum_t \eta_t^2 < \infty$, the mean of the applicant pool proportion for group u converges to the target proportion $\bar{s}$ under the FG policy, when the score distributions across the groups are identical.*

Proof. We wish to show that $\theta_t \to \bar{s}$ as $t \to \infty$. Let $d_t = \frac{1}{2}(\theta_t - \bar{s})^2$. Fix an $\epsilon > 0$. We need to show that there exists some $t_0(\epsilon)$ such that when $t \geq t_0(\epsilon)$,

$$d_{t+1} \leq d_t - \gamma_t, \text{ if } d_t \geq \epsilon \tag{18}$$

$$d_{t+1} < c\epsilon, \text{ if } d_t < \epsilon \tag{19}$$

where c is a positive constant. Moreover $\gamma_t > 0$ and $\sum_t \gamma_t = \infty$. If the above hold, then eventually for some $t = t_1(\epsilon) \geq t_0(\epsilon)$, one has $d_t < \epsilon$. But due to (18) and (19) $d_t < c\epsilon$ for all $t > t_1(\epsilon)$. Since ϵ is arbitrary, $\theta_t \to \bar{s}$ as $t \to \infty$.

We first show that (19) holds.

$$\begin{aligned}
d_{t+1} &= \frac{1}{2}(\theta_{t+1} - \bar{s})^2 \\
&= \frac{1}{2}([\theta_t - \eta_t(s_t - a_t)]_{\mathcal{C}} - \bar{s})^2 \\
&\leq \frac{1}{2}(\theta_t - \eta_t(s_t - a_t) - \bar{s})^2 \\
&= d_t + \eta_t(\bar{s} - \theta_t)(s_t - a_t) + \frac{1}{2}\eta_t^2(s_t - a_t)^2 \\
&\leq d_t + \eta_t(\bar{s} - \theta_t)(s_t - a_t) + \frac{1}{2}\eta_t^2 \\
&\leq d_t + \frac{\eta_t}{2}((\bar{s} - \theta_t)^2 + 1) + \frac{1}{2}\eta_t^2
\end{aligned}$$

Since η_t is arbitrarily small, if $d_t < \epsilon$, we have:

$$d_{t+1} < c\epsilon. \tag{20}$$

When $d_t \geq \epsilon$, we want to first show that

$$(\bar{s} - \theta_t)(\theta_t - a_t) \leq -\delta(\epsilon) \tag{21}$$

where $\delta(\epsilon) > 0$. If this holds, we have,

$$d_{t+1} \leq d_t - \eta_t \delta(\epsilon) + \frac{1}{2}\eta_t^2. \tag{22}$$

Let us denote $\gamma_t = \eta_t \delta(\epsilon) - \frac{1}{2}\eta_t^2$. Since $\eta_t \to 0$, there exists some $t_2(\epsilon)$ such that $\gamma_t > 0$ for $t > t_2(\epsilon)$. Moreover, due to conditions on step size, we have $\sum_t \gamma_t = \infty$.

Next, we will account for the stochasticity of s_t. We have $s_t - a_t = \theta_t + (s_t - \theta_t) - a_t$. Denoting $z_t = s_t - \theta_t$, we have

$$d_{t+1} \leq d_t + \eta_t(\bar{s} - \theta_t)(\theta_t + z_t - a_t) + \frac{1}{2}\eta_t^2. \tag{23}$$

z_t is a zero-mean random variable. Also $E[z_t^2] = var(s_t) = \theta_t / N_t$, which is bounded. Therefore $v_t := \sum_{m=0}^{t} \eta_m z_m$ is a martingale, and $E[v_t^2]$ is also bounded. This implies, by the martingale convergence theorem, that v_t converges to a finite random variable. Therefore, we have $\sum_{m=t}^{\infty} \eta_m z_m \to 0$. Since $\mid \theta_t - \bar{s} \mid$ is bounded, the effect of noise z_t is asymptotically negligible.

What remains to be shown is (21). In the regime of large number of applicants N_t, we can see that the state s_t is equal to its mean θ_t with probability approaching one, through the Chebyshev inequality.When $d_t \geq \epsilon$, since s_t is equal to θ_t, we need to consider only the cases (i) $s_t > \bar{s}$ and (ii) $s_t < \bar{s}$. Under both these cases, we have $(\bar{s} - \theta_t)(\theta_t - a_t) < 0$ due to the structure of the FG policy in (15) and (16), when the score distributions across the groups are identical. □

B. Experimental details

B.1. Evaluation on synthetic data

Optimal policy based on value iteration Let us first consider the MDP setting from Section 2, where the policy learnt is the optimal policy based on value iteration maximizing the accumulated utilities. Consider the case where the two groups have identical score distributions. This may often be the case in real-world scenarios when there is no inherent reason for the sensitive attribute to influence the scores or proficiency of a candidate. Let the score distributions be Gaussian with means $\mu_u = \mu_v = 5$ and variances $\sigma_u^2 = \sigma_v^2 = 1$. In this experiment, we set $\bar{s} = 0.4$ and the admission rate is fixed to $\bar{a} = 0.3$, or in other words, the selector aims to admit only 30% of the total applied candidates. The other parameter values used for this experiment are $\gamma = 0.99$, $\lambda = 1.5$, a fixed step-size of $\eta = 0.05$. Figure 2 shows how the proportion of applicants, admitted candidates and mean parameter θ_t vary for group u. We see in the figure that beginning the process from different initial states $\theta_0 = 0.1, 0.9$, we observe convergence of the applicant pool proportion for group u. The optimal policy under the evolution model considered has resulted in close to 40% of the applicants belonging to group u, and also approximately the same proportion of the admitted candidates are from group u. However, here the hyperparameter λ needs to be tuned to achieve a desired fairness target.

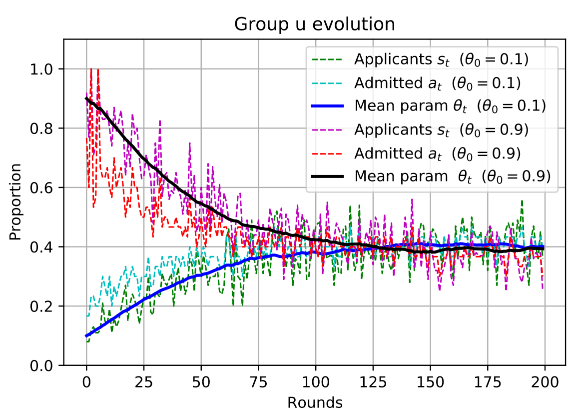

Figure 2: Optimal policy under identical score distributions across the groups.

B.2. Evaluation on dynamically adapted real-world datasets

The law school bar exam dataset consists of data collected by a Law School Admission Council survey across law schools in the United States. The predictions indicate whether or not a candidate would pass the bar exam based on features such as LSAT scores, undergraduate GPA, law school GPA, race, sex, family income, age and so on. We consider race as the sensitive attribute, and though originally there are 8 distinct races in the dataset, we group the samples by combining samples corresponding to all others except 'white', giving rise to binary groups 'white' and 'non-white'. The German credit dataset consists of 1000 instances, with 20 features (both numeric and qualitative), such as credit history, account history, employment status, age, gender and so on. This is typically used to assess the risk of lending loans to people, i.e., to determine if granting credit is risky or not. We consider gender as the binary valued sensitive attribute, labeling women as group u and men as group v. The dataset is imbalanced – about 31% of the instances belong to group u.

After pre-processing the datasets to suit our usage, our first step is to learn a score distribution that measures the proficiency of every sample. To achieve this, we fit a predictor based on logistic regression that uses the features and labels to fit scores, which are the derived as the product of the model coefficients and the features. The histograms of the scores of the two groups reveal that they are Gaussian in nature. We fit a Gaussian distribution for the group-wise scores, to obtain the mean and variance parameters of the score distributions $\mathcal{P}_u$ and $\mathcal{P}_v$.

The histograms and the Gaussian fit for the score distributions for the law school and German credit dataset are depicted in Figures 1(e) and 3 respectively. For the law school dataset the parameters of scores are $\mu_u = -1.46$, $\sigma_u^2 = 2.73$, $\mu_v = 0.79$, $\sigma_v^2 = 3.16$. For the German credit dataset the score distributions are closer with parameters $\mu_u = 0.32$, $\sigma_u^2 = 1.93$, $\mu_v = 0.85$, $\sigma_v^2 = 2.06$.

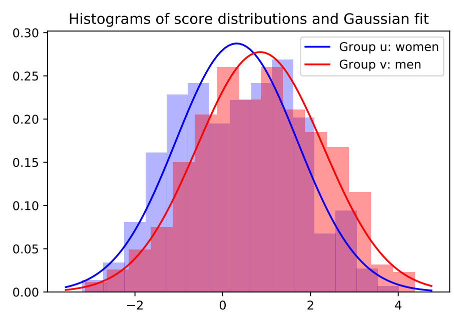

Figure 3: Histograms and Gaussian fit for score distributions of German credit dataset

We will now simulate the dynamics of the application process, under the FG policy, by sampling from these distributions with initial state of the applicant process θ_0 determined by the number of instances of respective groups, which is 0.25 for the law school and 0.31 for the German credit datasets respectively. The variation of the applicant pool for different values of hyperparameter λ are shown for the datasets in Figures 1(f) and 4 respectively. The evolution step size used in these simulations is $\eta = 0.025$, admission rate is set to $\bar{a} = 0.3$ and the target proportion is set to $\bar{s} = 0.5$, which is equivalent to demographic parity, i.e., admitting same number proportion of candidates from both groups. In both the figures, we observe that when the greedy reward is favored (lower values of λ), the applicant pool in fact converges to a point lesser than the target, while it approaches the target as λ increases. This means that for maximizing the utility, more samples need to be admitted from group v, due to the nature of their score distributions, when less importance is allotted to fairness objective. The tuning of the hyperparameter λ to achieve desired level of applicant pool proportion depends on the order statistics of $\mathcal{P}_u$ and $\mathcal{P}_v$. The step-size parameter η can be set appropriately based on how quickly we wish to achieve convergence.

These experiments with real-world datasets indicate that scores which are fit after learning predictors based on logistic regression are distributed like Gaussians. Once we have the parameters of the scores, the application of the FG policy and the applicant pool evolution follows.

It is interesting to examine how the score distributions change when we approach fairness through unawareness, that is, by omitting the sensitive attributes while learning the logistic regression based predictor. Note that we learn a single predictor based on all samples and then distinguish the scores based on the sensitive attribute. Table 2 lists the score parameters when the predictor is learnt with or without the inclusion of the sensitive attribute for the law school bar study and the

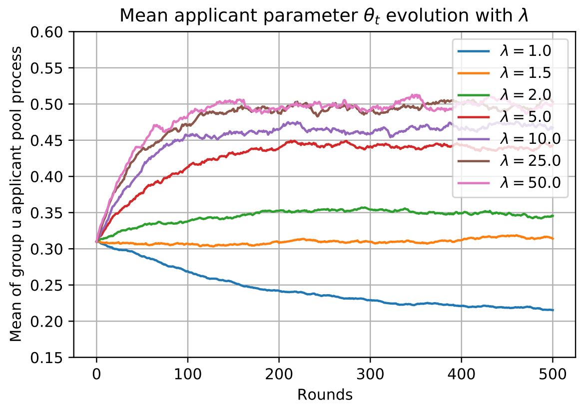

Figure 4: German credit dataset: applicant pool convergence with initial mean proportion parameter $\theta_0 = 0.31$, as λ is varied.

German credit datasets. For the law school dataset, we observe that the score distributions are not very different, although the difference between the means of minority and majority groups has decreased slightly when the sensitive attribute is dropped during the learning. For the German credit dataset, the distributions are significantly closer when the sensitive attribute is omitted, and there is a clear drop in the difference between the group means.

Dataset	Sensitive attribute	μ_u	μ_v	σ_u^2	σ_v^2
LS bar study	included	-1.46	0.79	2.73	3.16
LS bar study	excluded	-1.33	0.76	2.85	3.23
German credit	included	0.32	0.85	1.93	2.06
German credit	excluded	0.62	0.84	2.03	2.14

Table 2: Gaussian score distribution parameters for different datasets