The Backfire Effects of Fairness Constraints

Yi Sun [1] **Alfredo Cuesta-Infante** [2] **Kalyan Veeramachaneni** [1]

Abstract

Recently the fairness community has shifted from achieving one-shot fair decisions to striving for long-term fairness. In this work, we propose a metric to measure the long-term impact of a policy on the target variable distributions. We theoretically characterize the conditions under which threshold policies could lead to a backfire effect. We conduct experiments with a set of well-used fairness constraints on both synthetic and real-world datasets.

1. Introduction

The past decades have seen tremendous developments in machine learning algorithms. As machine learning algorithms have been deployed to fields such as loan application or recidivism prediction [5] [2], there are growing concerns about potential biases of those algorithms. Current solutions mitigating biases mostly focus on decision fairness, which ensures that decisions made by the algorithm is not disparate among population groups. This is based on the assumption that decisions will not affect the outcome distributions of the population. However, decisions made about individuals often create a feedback loop and nudge different segments of the population towards different distributions. Even decisions with good intentions may create unintended feedback loops. For example, a bank might take affirmative action and approve loans at a lower threshold for people coming from less-privileged socioeconomic groups. Yet if those people then have trouble paying back the loan later, the effect would further decrease their credit scores and credentials in the future.

There is a growing need to understand the long-term impact of deployed fairness concerned algorithms. In this paper, we focus on a set of threshold policies and investigate whether enforcing these policies have an equalized impact on different population groups if feedback loops of decisions are taken into consideration. In particular, we assume there is a target variable that measures the success probability of a positive outcome, such as loan repay probability. We consider a policy fair if it has a equal/comparable impact on shaping the distribution of the target variable of the groups. Our work characterizes when backfire effect occurs – where policies that appear neutral or fair result in a disproportionate impact on a protected group.

In this work, we make the following contributions:

- We first formally propose a metric to measure the impact of a threshold policy on the target variable distributions in terms of within-group and between-group segregation.
- We use Structural Causal Models (SCM) to theoretically characterize the conditions under which threshold policies could further entrench disparity of the target variable and lead to a backfire effect.We use both synthetic and real world dataset to illustrate the insight.

2. Background

Many of the datasets available in the fairness applications could have causal nature, where there are often causal dependencies of the target variable on the group attribute. In this section, we formulate the modeling assumptions in the lens of causal Markov Decision Process (MDP), where the MDP models the temporal transition of the underlying distribution and the Structural Causal Models (SCM) characterizes the causal relationship between variables at each state as a causal graph. This framework allows us to capture the causal relationships between variables in a dynamic setting.

At each time step, the state is a causal graph $\mathcal{G}$ with three endogenous nodes: (X, Y, Z), where $X \in \mathbb{R}^d$ is the set of features, $Z \in \{0, 1\}$ is the time-invariant sensitive attribute, and $Y \in [0, 1]$ is the target variable.

In the following we describe the transition in terms of the structural equations.

- **Initialization**: The process is initialized with time-invariant sensitive attribute Z, a set of observed features X^t, and the target variable Y^t. The target variable Y^t is a function of the features X^t, i.e., $Y^t = f_Y(X^t)$.
- **Action**: At time step t, a binary action $A^t \in \{0, 1\}$ is applied, which potentially depends on Y^t and sensitive

[1] MIT [2] University Rey Juan Carlos. Correspondence to: Yi Sun <yis@mit.edu>.

ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA, 2022. Copyright 2022 by the author(s).

Z, i.e, $A^t = f_A(Y^t, Z)$.

- **Outcome**: After applying the action, an binary outcome is observed. We use an auxiliary variable $O^t \sim Bernoulli(Y^t)$ to indicate the outcome variable, which is sampled from Bernoulli distribution with Y^t as the parameter.
- **Transition**: Based on the realized outcome O^t, the features X^t will be updated with based on action and outcome, where $X^{t+1} = f_X(X^t, A^t, O^t)$. The target variable Y^t will be updated accordingly.
- **Utility**: The decision maker's utility is a function of the realized outcome and the action, i.e., $\mathcal{U}^t = f_{\mathcal{U}}(O^t, A^t)$. Specifically, if a decision maker assigns a positive action ($A^t = 1$), the utility is increased by u_+ if the outcome is positive, and decreased by u_- if the outcome is negative.

3. The backfire effect of fairness policies

We first categorize the impact of a policy into within-group disparity and between-group disparity:

- **Within-group Disparity**: Within-group disparity happens when a policy could lead to further inequality and dichotomy within a population group.
- **Between-group Disparity**: Between-group disparity happens when a policy entrench the distribution disparity between two population groups.

We introduce a novel fairness metric to measure the impact of a policy on the distribution of the target variables. In particular, we consider a policy fair if it has equal or comparable effects on the target variable both in terms of within-group disparity and between-group disparity.

Definition 3.1 (Within-group disparity)**.** Let Y_z^t be the group z's target variable, and $\delta_z^t = g(Y_z^t) - g(Y_z^0)$ be the change in the aggregated value with respect to $t = 0$ for group z, where $g(\cdot)$ is an aggregation function.

Definition 3.2 (Between-group disparity)**.** We define the between-group disparity at time step t as

$$\Delta^t = |\delta_{z=0}^t - \delta_{z=1}^t|$$

We say that a policy has a backfire effect if $\Delta^T \geq \Delta^0$,i.e., the policy increases the disparity from time $t = 0$. We use figure 1 to illustrate the backfire effect in terms of within-group disparity and between-group disparity.

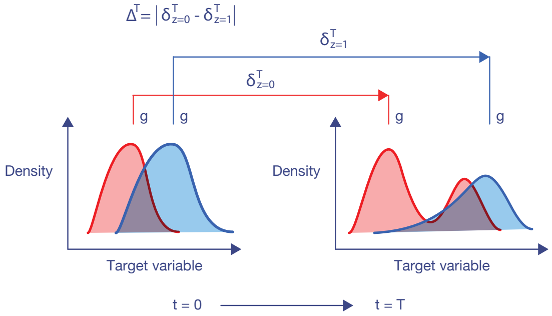


Figure 1: An illustration of backfire effects of a policy. $\delta_{z=0}$ and $\delta_{z=1}$ measures the impact of the policy on the red group and blue group respectively. The policy leads to within-group disparity for the red group as well as between-group disparity.

The choice of the aggregation function In previous work [1][4][6], the analysis has only been focused on the group average, where g is the mean function. We extend the analysis by allowing g to be a distance function that measures the shift from distribution Y^0 to distribution Y^t (F-divergences for example) to characterize the distribution shift in a more fine-grained way. One interesting choice with real-world implication is Gini-coefficient, which measures income inequality within a population group.

3.1. The impact of threshold policies

In this section, we restrict our attention to the family of threshold policies, where $A^t = \mathbb{1}(Y^t \geq \tau^t)$ for some threshold τ^t. We analyze theoretically when threshold policies would lead to disparity between groups. Let

$$X_+ = f_X(X^t, A^t = 1, O^t = 1) - X^t$$
$$X_- = X^t - f_X(X^t, A^t = 1, O^t = 0)$$

be the change in feature value for positive outcome and negative outcome respectively.

Theorem 3.3. *The within-group disparity at time T after applying thresholds $\tau^1, .., \tau^T$ is*

$$\delta_z(\tau^{1:T}) = \sum_{t=0}^{T} g((1 - F_{Y_z^t}(\tau^t))\nabla f_Y(X^t)[Y_z^t(X_+ + X_-) - X_-])$$

and the between-group disparity is

$$\Delta^T = |\delta_0(\tau_0^{1:T}) - \delta_1(\tau_1^{1:T})|$$

where $F_{Y_z^t}$ is the CDF of Y_z^t, and $\tau_0^{1:T}$ and $\tau_1^{1:T}$ are thresholds applied on group 0 and 1 respectively.

The theorem shows that the disparity depends on the initial distribution, and the feature contribution direction $\nabla f_Y(X^t)$. With all other things fixed, when threshold τ increases, $1 - F_{Y_z^t}(\tau^t)$ decreases, and the disparity decreases.

4. Case Studies

Next we use two case studies to empirically illustrate the impact of threshold policies. In both cases, the initial group distribution is time-invariant and sampled from $Z^0 \sim Bernoulli(p_0)$ where p_0 is the probability the an individual comes from group z_0.

Loan Application Example The loan application example was first proposed by [3] to study the one-step feedback effect of fairness constraints. We first frame the loan application example in the format of a dynamic SCM. The variables in the SCM are as follows: $Z \in \{0, 1\}$ is the binary sensitive attribute, $X \in [c_{min}, c_{max}]$ is the credit score , $A \in \{0, 1\}$ is the binary loan approval/rejection decision, and $Y \in [0, 1]$ is the probability of repaying. The initial distribution of Z, X^0, Y^0 is estimated from the dataset. Since the data is not sequential in nature, we use a synthetic structural equation for the feature update function f_X.

$$X^{t+1} = \begin{cases} \min\{X^t + X_+, c_{max}\} & \text{if } O^t = 1, A^t = 1 \\ \max\{c_{min}, X^t - X_-\} & \text{if } O^t = 0, A^t = 1 \end{cases}$$

Synthetic Gaussian In the second example, we extend previous simulation work [1] where the feature variable is only 1-dim. The initial feature distribution X^0 is sampled from a group-specific 2-dim Gaussian distribution. The target variable is the sigmoid of a linear transformation of the feature vectors with weight vector M. The i-th feature positively contribute to the target variable ($[\nabla f_Y(X)]_i > 0$) if i-th component in M is positive.

$$X^0 \sim \mathcal{N}_d(\mu_z, \Sigma_z)$$

$$Y^t = \frac{1}{1 + e^{-X^t \cdot M}}$$

$$X^{t+1} = \begin{cases} X^t + X_+ & \text{if } O^t = 1, A^t = 1 \\ X^t - X_- & \text{if } O^t = 0, A^t = 1 \end{cases}$$

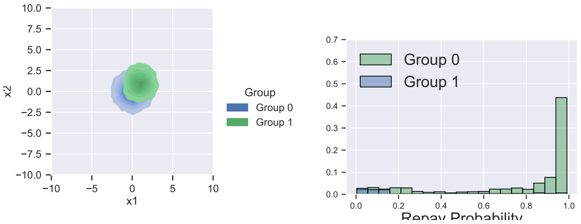


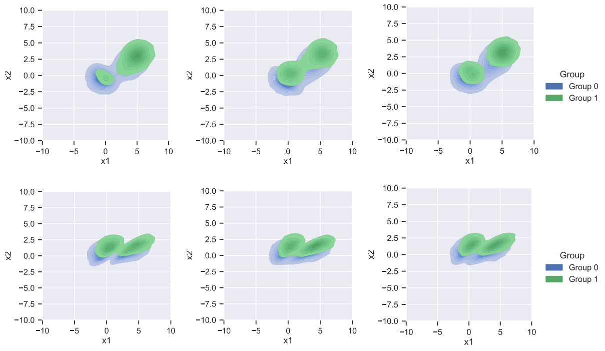
Figure 2: Left: Initial feature distribution of the synthetic Gaussian example. Right: Initial target variable distribution of loan application example.

4.1. Candidate Threshold Policies

The goal of the decision maker is to determine a threshold that maximizes its utility function subject to some fairness constraints. Here we list a few commonly used fairness constraints used in simulation.

Max Utility Max utility (`MaxUtil`) policy maximizes the expected utility without constraints.

Demographic Parity Demographic parity (`DemoPar`) policy maximizes the expected utility subject to the demographic parity constraints, which requires that a policy issues loans to the same percentage of applicants in both groups ($\mathbb{E}[A^t = 1|Z = 0] = \mathbb{E}[A^t = 1|Z = 1]$).



(a) `MaxUtil`. (b) `DemoPar`. (c) `EqOpp`.

Figure 3: Synthetic Gaussian Example. Top: $M_1 = [1, 1]$. Bottow: $M_2 = [1, -1]$.

Equalized Opportunity Equalized opportunity (`EqOpp`) policy maximizes the expected utility subject to the equalized odds constraints, which requires that both groups have equalized false positive rates, i.e. ,$\mathbb{E}[A^t = 1|Y^t = 0, Z = 0] = \mathbb{E}[A^t = 1|Y^t = 0, Z = 1]$.

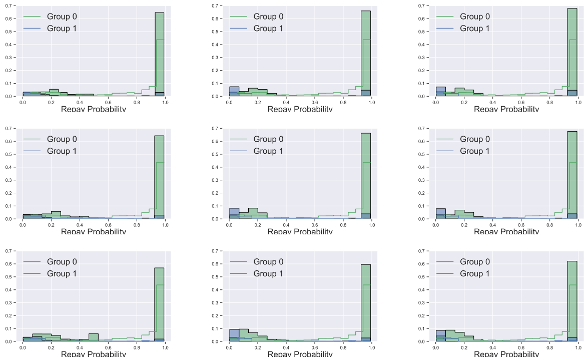
It can be seen that `MaxUtil`, `DemParity` and `EqOpp` belong to the family of threshold policies.

4.2. Simulation result

Loan Application We leverage the loan application example to examine the effects of cost ratio q. We define the **cost ratio** $q = \frac{X_-}{X_+ + X_-}$ as the fraction of change in the features X^t for a negative outcome, with respect to the full range it could change. We categorize the simulation settings into three regimes based on the value of the cost ratio: (1) forgiving setting: $q < \frac{1}{2}(X^+ = 150, X^- = 75)$; (2) neutral setting: $q = \frac{1}{2}(X^+ = X^- = 75)$; (3) harsh setting: $q > \frac{1}{2}(X^+ = 75, X^- = 150)$.

In figure 4, we plot the final distribution of the target variable Y under three different settings. Compared to the initial distribution shown in figure 2, repeatedly enforcing a policy change the shape of the distributions in a way that could not simply captured by the group mean. All policies create dichotomy and within group disparity on the target variable distribution where “the rich gets richer”.

In figure 5 and figure 6, we plot the within-group disparity and between-group disparity as a function of the cost ratio q using mean and gini-coefficient as the aggregation function respectively. When the cost ratio q is lower, `EqOpp` results in higher between-group disparity when using mean as a metric. As cost ratio increases, `DemoPar` results in the



(a) `MaxUtil`. (b) `DemoPar`. (c) `EqOpp`.

Figure 4: Loan application example. The unfilled bar indicate the initial distribution and the filled bars indicate the final distribution. Top row: forgiving setting. Middle row: neutral setting. Bottom row: harsh setting.

	`MaxUtil`	`DemoPar`	`EqOpp`
Mean	0.011	0.005	**0.024**
Median	0.007	**0.039**	0.008
KL-divergence	0.034	**0.414**	0.301
Wasserstein	0.003	**0.079**	0.074

Table 1: Disparity Δ^T when measured using different g function (forgiving setting). The bold number indicates the policy that results in the largest disparity. Using different aggregation function g leads to different conclusions.

biggest disparity of mean outcome between the two groups. This is in line with what we showed in the theoretical analysis above: lower thresholds would lead to an increase of within-group disparity.

When using gini-coefficient as a metric, `DemoPar` leads to highest between-group disparity. In general, as cost ratio increases, gini-coefficient tends to increase for both within-group and between-group disparity, which implies that unevenness within each group increases as the cost ratio increases.

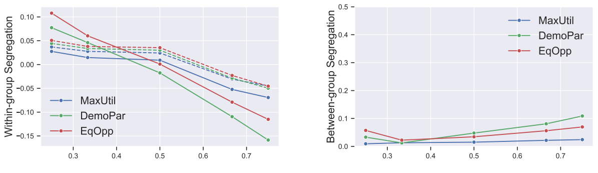


Figure 5: Using mean to measure disparity as a function of cost ratio q.

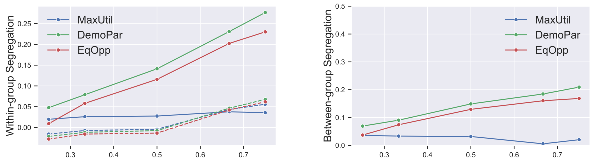


Figure 6: Using gini-coefficient to measure disparity as a function of cost ratio q.

Synthetic Gaussian (2d) The initial features for the two groups is sampled from $\mathcal{N}(\mu_0, I)$ and $\mathcal{N}(\mu_1, I)$ respectively, where $\mu_0 = [0, 0]^T$ and $\mu_1 = [1, 1]^T$. We plot the initial feature distributions in figure 2. The feature update is $X_+ = [0.02, 0.01]$ and $X_- = [0.01, 0.02]$). We simulate with two feature contribution matrix $M_1 = [1, 1]$ and $M_2 = [1, -1]$, where the first/second feature is a "bad" feature (negatively impacts the target variable) respectively. In figure 3, we plot the final distribution of the features under different structural equation f_Y. In the top row, both features positively contribute to the target variable ($f_Y(X) > 0$). In the bottom row, the second feature negatively contribute to the target variable. This shapes the feature spaces differently despite the feature update equation in the same.

4.3. Key Takeaways

In this work, we study if enforcing fairness constraints could lead to backfire effects. We show that repeatedly enforcing fairness constraints (`DemoPar,EqOpp`) could lead to within-group disparity and between-group disparity than `MaxUtil`. In addition, the evaluation of long-term impact of fairness constraints is very sensitive to the metrics used for measuring disparity and disparity, and different metrics could lead to different conclusions on which fairness policy leads to the biggest backfire effect.

5. Related Work

Several works have studied the dynamics between algorithmic decisions and the long-term population qualifications. One of the first works that touches on this topic is [3], which considers the one-step feedback model and shows that enforcing common static fairness metrics in constrained optimization does not in general promote average group scores. Later, [1] extends previous one-step analysis to multiple-step using simulation, and argues that long-term dynamics may lead to different conclusions from one-shot analysis. [4] studies whether enforcing demographic parity could lead to equality of qualifications. Most related to our work, [6] studies the problem under a partially observed Markov decision problem setting, and characterizes the impacts of fairness constraints can have on the equilibrium of group qualification rates. One thing that has been missing from previous work is that the analysis only focus on group mean or the features or the qualification, yet an algorithm or policy could have more profound impact on the shape of the population beyond group mean.

Acknowledgments

Dr. Cuesta-Infante is funded by the Spanish Government research fundings RTI2018-098743-B-I00 (MICINN/FEDER), PID2021-128362OB-I00 (MICINN/FEDER) and Y2018/EMT-5062 (Comunidad de Madrid).

References

[1] D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., and Halpern, Y. Fairness is not static: Deeper understanding of long term fairness via simulation studies. In *Proc. of the Conference on Fairness, Accountability, and Transparency*, pp. 525–534, 2020. doi: 10.1145/3351095.3372878.

[2] Dressel, J. and Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4, 2018.

[3] Liu, L., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. *ArXiv*, abs/1803.04383, 2018.

[4] Mouzannar, H., Ohannessian, M. I., and Srebro, N. From fair decision making to social equality. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

[5] Obermeyer, Z., Powers, B. W., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366:447 – 453, 2019.

[6] Zhang, X., Tu, R., Liu, Y., Liu, M., Kjellstrom, H., Zhang, K., and Zhang, C. How do fair decisions fare in long-term qualification? *ArXiv*, abs/2010.11300, 2020.

A. Appendix

A.1. Additional Experiment Results

A.1.1. Thresholds for each policy as cost ratio increases

In figure 7, we plot the average thresholds of each policy as a function of the cost ratio.

Regardless of the group, `MaxUtil`'s threshold only depends on the parameter for the utility function ($\frac{u_-}{u_- + u_+}$), which is set as 0.5 in the experiment.

For `DemoPar` policy, it consistently overcompensate for the disadvantaged group by assigning a lower threshold for the disadvantaged group.

The case with `EqOpp` is a little bit more complicated. As cost ratio increases, `EqOpp` switches from lower threshold for disadvantaged group to lower threshold for advantaged group.

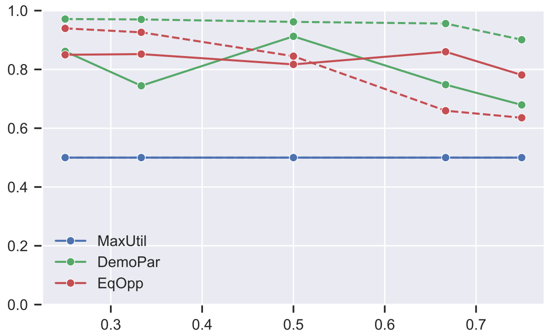


Figure 7: Threshold as a function of cost ratio. The dashline indicates the threshold for advantaged group, and the solid line indicates the threshold for disadvantaged group.

A.1.2. Thresholds of different policies as the cost ratio increases.

As shown in figure 8, simulation on individuals coming from different quantile of the initial target variable different could lead to different narratives. The `EqOpp` leads to the smallest impact for individuals starting at the 25th quantile, yet it leads to the biggest backfire effect for individuals starting at the 75th quantile.

A.2. Proof of Lemma 1 and Theorem 1

A.2.1. Individual Level Equilibrium

We first derive equilibrium conditions for target variable Y on an individual level, where $\mathbb{E}[Y^{t+1}] = Y^t$.

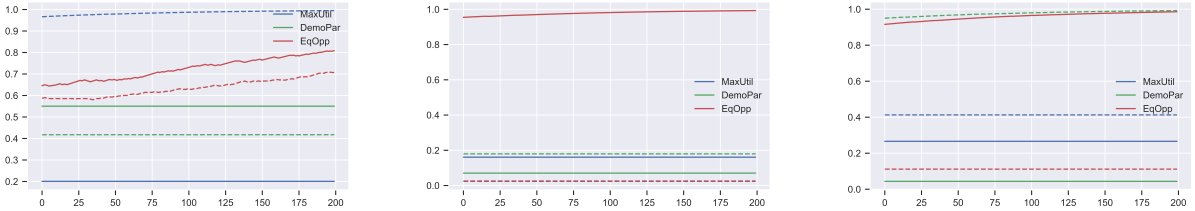


Figure 8: Evolution of target variable for individuals starting at 25th, 50th, and 75th quantile. Top: Synthetic Gaussian. Bottom: Loan application example

Let $\pi : \mathcal{S} \to [0,1]$ be the policy function/predictor that maps from states S^t to the probability of probability of a positive action, i.e., $\mathbb{P}(A^t = 1) = \pi(S^t)$.

With threshold policies, the action $A^t = 1$ if and only if $Y^t > \tau^t$, i.e. the feature values are above the threshold. For threshold policies, we have

$$\mathbb{P}(A^t = 1) = \pi^t(Y^t) = \mathbb{P}(Y^t \geq \tau^t) = 1 - \mathbb{P}(Y^t \leq \tau^t)$$

By construction, the target variable indicates the probability of getting a positive outcome , i.e., $\mathbb{P}(O^t = 1) = Y^t$.

Let X_+^t and X_-^t be the magnitude of change of feature value for positive outcome and negative outcome respectively.

$$\begin{aligned} X_+^t &= f_X(X^t, A^t = 1, O^t = 1) - X^t \\ X_-^t &= X^t - f_X(X^t, A^t = 1, O^t = 0) \end{aligned}$$

We can write $\mathbb{E}[Y^{t+1} | A^t = 1, O^t = 1]$ by expanding on the structural equation, i.e.,

$$\begin{aligned} \mathbb{E}[Y^{t+1} | A^t = 1, O^t = 1] &= f_Y(f_X(X^t, A^t = 1, O^t = 1)) = f_Y(X^t + X_+^t) \\ \mathbb{E}[Y^{t+1} | A^t = 1, O^t = 0] &= f_Y(f_X(X^t, A^t = 1, O^t = 0)) = f_Y(X^t - X_-^t) \end{aligned}$$

We can compute the expected value of X^{t+1} as a function of X^t using law of total expectation:

$$\begin{aligned} \mathbb{E}[X^{t+1}] &= \mathbb{E}[X^{t+1} | A^t = 1, O^t = 1]\mathbb{P}[A^t = 1, O^t = 1] \;\; \text{positive action and positive outcome} \\ &+ \mathbb{E}[X^{t+1} | A^t = 1, O^t = 0]\mathbb{P}[A^t = 1, O^t = 0] \;\; \text{positive action and negative outcome} \\ &+ \mathbb{E}[X^{t+1} | A^t = 0]\mathbb{P}[A^t = 0] \;\; \text{negative action} \\ &= (X^t + X_+^t)\pi(Y^t)Y^t + (X^t - X_-^t)\pi(Y^t)(1 - Y^t) + X^t(1 - \pi(Y^t)) \\ &= X^t + \pi(Y^t)[(X_+ + X_-)f_Y(X^t) - X_-] \\ \mathbb{E}[Y^{t+1}] &= \mathbb{E}[f_Y(X^{t+1})] \\ &= f_Y(X^t + X_+^t)\pi(Y^t)Y^t + f_Y(X^t - X_-^t)\pi(Y^t)(1 - Y^t) + f_Y(X^t)(1 - \pi(Y^t)) \\ &\approx f_Y(X^t) + \nabla f_Y(X^t)X_+^t\pi(Y^t)Y^t + f_Y(X^t) - \nabla f_Y(X^t)X_-^t\pi(Y^t)(1 - Y^t) + f_Y(X^t)(1 - \pi(Y^t)) \\ &= Y^t + \pi(Y^t)\nabla f_Y(X^t)[(X_+^t + X_-^t)Y^t - X_-^t] \end{aligned}$$

where we use first order linear approximation to approximate $f_Y(X^t + X_+^t)$, where $f_Y(X^t + \epsilon) \approx f_Y(X^t) + \nabla f_Y(X^t)\epsilon$.

Fixed point theorem The iteration achieves a fixed point x^* when $x^* = \Phi(x^*)$. According to Banach Fixed point, a unique fixed point x^* exists if Φ is a contraction mapping, i.e., $d(\Phi(x_1), \phi(x_2)) \leq Ld(x_1, x_2)$ for some distance function d and Lipschitz constant $L < 1$.

The equilibrium happens when $\pi^t \nabla f_Y(X^t)[Y^t(X_+^t + X_-^t) - X_-^t] = 0$, where

$$Y^t = \frac{\nabla f_Y(X^t)X_-^t}{\nabla f_Y(X^t)(X_+^t + X_-^t)}$$

Counter-intuitively, the change direction of the target variable only depends on the feature contribution direction $\nabla f_Y(X^t)$ and cost ratio $q^t = \frac{X_-^t}{X_-^t + X_+^t}$, not on the threshold value τ. On the other hand, the threshold value τ would impact the rate of changing. Specifically, when τ is lower, $\pi(Y^t) = \mathbb{P}(Y^t \geq \tau)$ would be greater, and the changing rate of the target variable would also be greater.

Using the recursion we have, we can find the cumulative change in Y:

$$\mathbb{E}[Y^T] - Y^0 = \sum_{t=0}^{T} \pi(Y^t)\nabla f_Y(X^t)(Y^t(X_+^t - X_-^t) - X_-^t)$$

A.2.2. Group Level Equilibrium

$$\begin{aligned}
\delta_z^T &= g(Y_z^T) - g(Y_z^0) \\
&= \sum_{t=0}^{T} g(\mathbb{P}(Y_z^t \geq \tau^t)\nabla f_Y(X^t)[Y_z^t(X_+^t + X_-^t) - X_-^t]) \\
&= \sum_{t=0}^{T} g((1 - F_{Y_z^t}(\tau^t))\nabla f_Y(X^t)[Y_z^t(X_+^t + X_-^t) - X_-^t])
\end{aligned}$$

where $F_{Y_z^t}$ is the CDF distribution of the random variable Y from group z at time t.

$$\begin{aligned}
\Delta^t &= |\delta_{z=0}^t - \delta_{z=1}^t| \\
&= |h(\tau_1^t) - h(\tau_2^t)|
\end{aligned}$$

where $h(\tau^t) = \sum_{t=0}^{T} g((1 - F_{Y_z^t}(\tau^t))\nabla f_Y(X^t)[Y^t(X_+^t + X_-^t) - X_-^t])$, and τ_1 ,τ_2 are thresholds for group z_1 and z_2 respectively.