PROCEDURAL FAIRNESS THROUGH ADDRESSING SOCIAL DETERMINANTS OF OPPORTUNITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Social determinants of opportunity are variables that, while not directly pertaining to any specific individual, capture key aspects of contexts and environments that have direct causal influences on certain attributes of an individual, e.g., environmental pollution in an area affects individual's health condition, and educational resources in an neighborhood influence individual's academic preparedness. Previous algorithmic fairness literature often overlooks *social determinants of opportunity*, leading to implications for procedural fairness and structural justice that are incomplete and potentially even inaccurate. We propose a modeling framework that explicitly incorporates *social determinants of opportunity* and their causal influences on individual-level attributes of interest. To demonstrate theoretical perspectives and practical applicability of our framework, we consider college admissions as a running example. Specifically, for three mainstream admission procedures that have historically been implemented or are still in use today, we distinguish and draw connections between the outcome of admission decision-making and the underlying distribution of academic preparedness in the applicant population. Our findings suggest that mitigation strategies centering solely around protected features may introduce new procedural unfairness when addressing existing discrimination. Considering both individual-level attributes and *social determinants of opportunity* facilitates a more comprehensive explication of benefits and burdens experienced by individuals from diverse demographic backgrounds as well as contextual environments, which is essential for understanding and achieving procedural fairness effectively and transparently.

Thank All Reviewers!

We are extremely grateful to all the reviewers for the comments and the time devoted. In the revised manuscript, we have carefully considered and incorporated the review comments. We provide color-coded side notes that correspond to comments/questions by each reviewer (Reviewer g5W4, Reviewer 8F2j, Reviewer h9Xa). We summarize the list of responses on **Page 16** at the beginning of the Appendix for the convenience of navigation.

1 INTRODUCTION

Structural injustice refers to circumstances in which social practices, social structures, or the environment reinforce and compound prior histories of injustice (Carmichael et al., 1967; Sowell, 1972; Tilly, 1998; Rothstein, 2017; Alexander, 2020). We use the term "social determinants of opportunity" to refer to the specific aspects of social practices, social structures, or the environment that have a profound impact on the opportunities of individuals. When members of specific demographic groups have been the subject of histories of unjust treatment, their demographic membership often correlates with circumstances in which they face significant social impediments to opportunity (Gee & Ford, 2011; Yearby, 2018; Robinson et al., 2020; Yearby et al., 2022; Chetty et al., 2024). Because the *social determinants of opportunity* are features of places, institutions, policies, or practices, they persist even if attitudes that cause unjust treatment have been subject to significant reform. Their effects may not be tied directly to demographic group membership but to broader traits (such as income level or job type) or to geographic areas. As a result, individuals within the same demographic group, depending on their unique circumstances, may experience different impediments to opportunity due to intersecting social determinants, e.g., various environmental impacts on health in different geographic locations (Comber et al., 2011; Yeum et al., 2016; Tan et al., 2020). Conversely, individuals from different demographic groups in the same geographic neighborhood may encounter similar impediments to opportunity, e.g., poverty and pollution in the neighborhood, lack of educational resource in the community (Connell, 1994; Tilak, 2002; Rose & Dyer, 2008).

Previous research on algorithmic fairness has focused on protected features, e.g., race, sex, gender, and age (Romei & Ruggieri, 2014; Loftus et al., 2018; Corbett-Davies & Goel, 2018; Mitchell et al., 2018; Narayanan, 2018; Verma & Rubin, 2018; Caton & Haas, 2020; Chouldechova & Roth, 2020; Makhlouf et al., 2020; Mehrabi et al., 2021; Zhang & Liu, 2021; Pessach & Shmueli, 2022; Tang et al., 2023b). Various fairness metrics that are directly defined upon protected features are proposed to estimate or bound empirical violations of fairness, based on observational statistics (Calders et al., 2009; Hardt et al., 2016; Zafar et al., 2017), causal properties and/or quantities (Kilbertus et al., 2017; Kusner et al., 2017; Nabi & Shpitser, 2018; Chiappa, 2019; Coston et al., 2020), and dynamic modelings (Liu et al., 2018; Zhang et al., 2020; Tang et al., 2023a). In terms of *auditing* potential fairness violations, the focus on protected features is natural since these are the features in virtue of which individuals might be subject directly to unfair treatment or might experience disproportionate burdens. However, the goal of *mitigation* goes beyond *auditing* fairness violations by seeking to intervene in ways that will reduce burdens and promote fairer outcomes in the future. Both protected features and *social determinants of opportunity* play important roles in the underlying causal mechanism, and therefore, need to be explicitly addressed when designing and evaluting mitigation strategies.

We consider *procedural fairness* that pertains directly to the data generating process itself (Rawls, 1971; 2001; Sen, 2011), and propose a framework that incorporates *social determinants of opportunity* for understanding and achieving procedural fairness. Our contributions can be summarized as follows:

- We advocate explicitly considering *social determinants of opportunity* because they capture key aspects of contexts and environments that have direct causal influences on specific attributes of individuals, overlooking which may result in incomplete or inaccurate claims for procedural fairness.
- We propose a modeling framework that incorporates *social determinants of opportunity*, and demonstrate how our approach facilitates nuanced analyses of benefits and burdens experienced by individuals with different demographic backgrounds as well as contexts and environments.
- Our findings suggest the importance of recognizing various factors in the underlying data generating process that have procedural fairness implications, some are individual-level protected features, others correlate with protected features but do not pertain to any specific individual.

2 Preliminaries

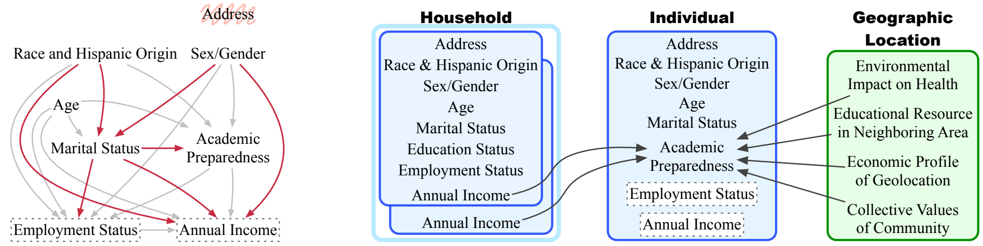
We first provide a brief introduction to causality and causal modeling (Section 2.1), and then, we present an overview of different procedures in college admissions as a running example (Section 2.2).

2.1 A Brief Introduction to Causal Modeling

We use uppercase letters to denote random variables, lowercase letters to denote values taken by variables, and calligraphic letters to denote corresponding domains of values. For instance, for a random variable Z, it can take a value z from its domain of values $\mathcal{Z}$. For two random variables W and V, we say that W is a direct cause of V if there is a change in distribution of V when we apply an intervention on W while holding all other variables fixed (Spirtes et al., 1993; Pearl, 2009). We can represent causal relations among variables via a directed acyclic graph (DAG), where nodes correspond to variables, and edges denote causal relations between variables and their direct causes.

2.2 Overview of Different Admission Procedures

There are three mainstream categories of admission procedures that were implemented in the history or are still in use today: quota-based admissions, holistic review with plus factors, and top-percentage plans. Variant forms of these procedures have been evaluated in law cases related to affirmative actions. At the core of these cases is the "compelling interest," a legally necessary and highly justified purpose that must be demonstrated to validate government measures that differentiate individuals based on race (Supreme Court, 1978; 2003a;b; 2013a;b; 2023a;b). Beyond jurisprudence, it is also established that the way and the extent to which race and ethnicity is utilized in the admission procedure should be under strict scrutiny, for instance, the studies on impacts of affirmative actions in education from economic literature (Sowell, 1972; 2004; Zimmerman, 2014; Bleemer, 2022) and from algorithmic fairness literature (Kusner et al., 2017; Nabi & Shpitser, 2018; Kannan et al., 2019; Chiappa, 2019). In this paper, we consider college admissions as a running example and present nuances in implications on procedural fairness, when employing different admission procedures.



(a) Previous causal modeling (b) Our causal model incorporating *social determinants of opportunity*

Figure 1: Comparison between different modeling choices when constructing a causal graph for the underlying data generating process. Panel (a) presents the modeling in previous causal fairness literature, where the color red denotes problematic edges (Kilbertus et al., 2017; Kusner et al., 2017; Nabi & Shpitser, 2018; Chiappa, 2019; Wu et al., 2019; Nabi et al., 2022). The variable `Address` is usually omitted in previous literature, as indicated by squiggles. Employment status and annual income of the applicant are enclosed in dashed boxes as they are typically not factors considered in college admissions. Panel (b) presents our modeling of influences from *social determinants of opportunity* on academic preparedness, corresponding to contexts and environments.

3 Social Determinants of Opportunity in Causal Mechanisms

We discuss in Section 3.1 issues of previous causal modeling for procedural fairness. In Section 3.2, we present our modeling framework incorporating *social determinants of opportunity*. [1]

Definition 3.1 (Social Determinants of Opportunity)**.** A *social determinant of opportunity* is a variable representing an aspect of the data generating process, in which one or more characteristics of the environment where individuals live or operate, that are not an attribute of any specific individual, have direct causal influence on the attributes of an individual.

3.1 Causal Modeling in Previous Works

Previous works in causal fairness represent discriminations in terms of problematic edges or paths in the causal graph. As shown in Figure 1(a), one can represent causal relations among relevant variables with a DAG, and further indicate objectionable aspects of data generating process with red edges or paths originating from protected features (Kilbertus et al., 2017; Kusner et al., 2017; Nabi & Shpitser, 2018; Chiappa, 2019; Wu et al., 2019; Nilforoshan et al., 2022; Nabi et al., 2022). Three potential issues may arise from this modeling choice.

The Recapitulation of Inappropriate Stereotypes In Figure 1(a), there are causal edges or paths originating from protected features to certain other variables, for instance, from sex or gender to annual income, and from race to educational status. While it seems intuitive to use the edge `Race` $\rightarrow$ `Education Status` to capture the potential racial discrimination in education (Kusner et al., 2017; Nabi & Shpitser, 2018; Chiappa, 2019; Nabi et al., 2022), the interpretation of the edge according to the definition of causal intervention reveals its potential controversies. Specifically, by definition of causality, this edge asserts that there is a difference in the distribution of education status, when we "intervene" on individual's race while keeping all other things unchanged.[2] Such modeling choice, although seemingly neutral from a technical perspective, unintentionally aligns with the controversial ideology of racial essentialism (racial groups possess underlying intrinsic essences, e.g., intellectual and biological, that make them different), which has been widely criticized due to the lack of scientific evidence supporting its claims (Roberts, 2011; Smedley, 2018; Delgado & Stefancic, 2023). If a certain edge or path in the causal model does not reflect an actual real-world causal process, subsequent causal fairness analyses based on causal effects may not provide informative conclusions.

The Limited Scope of Only Modeling Individual-Level Variables Compared to fairness notions based on observational statistics, causal fairness notions incorporate causal relations among features of individuals, such that interventional and counterfactual analyses can be conducted to reason about

[1] Due to space limit, we provide further discussions on related works in Appendix A.

[2] Here, we use "intervene" in quotes to signify the need of extra caution when discussing the manipulation of individual's race, due to both ethical and practical considerations.

different potential outcomes for specific individuals (Kilbertus et al., 2017; Kusner et al., 2017; Nabi & Shpitser, 2018; Chiappa, 2019; Wu et al., 2019; Nabi et al., 2022). Recent works have also utilized (sub)group-level statistics based on causal effects to capture group-level fairness (Zhang & Bareinboim, 2018b;a; Coston et al., 2020; Imai & Jiang, 2020; Mishler et al., 2021; Nilforoshan et al., 2022). However, the scope of considered variables is largely limited to attributes directly pertaining to an individual, e.g., demographic information, education status, and annual income in Figure 1(a). Other than individual-level variables, contextual environments actually have significant influences over the individual, for instance, the improvement in physical health observed in a randomized housing mobility social experiment (Ludwig et al., 2011), and the social determinants of health (Marmot & Wilkinson, 2005; Braveman & Gottlieb, 2014; Robinson et al., 2020; Yearby et al., 2022).

Omitting Relevant Variables In previous literature, it is a common practice to omit the variable `Address` in theoretical and empirical analyses. In particular, previous causal models do not include `Address` as a relevant variable (Kilbertus et al., 2017; Kusner et al., 2017; Nabi & Shpitser, 2018; Chiappa, 2019; Nabi et al., 2022). Furthermore, the empirical analyses tend to drop the `Address` information (or its alternatives) during data collection and/or preprocessing. For instance, there is no address information included in the Adult data set (Becker & Kohavi, 1996), the address information is dropped (Mary et al., 2019) when processing the Communities and Crimes data set (Redmond, 2009), the address-related information is dropped by the Folktables package when retrieving public-use data products from US Census Bureau and forming Adult-like data sets (Ding et al., 2021). This is problematic because through `Address` an individual can be related to a household via the family relation, and to a geographical location via the residence relation. This seemingly irrelevant variable (for college admission) actually contains information relevant to an individual's opportunity, since factors from contextual environments (e.g., environmental impact on health, educational resource available in neighboring area) affect academic preparedness of the applicant.

Reviewer h9Xa: **Q5.1**
Previous approaches tend to omit relevant variables

3.2 Our Modeling Approach

Presented in Figure 1(b), we unpack the semantics of edges or paths originating from protected features in Figure 1(a), and replace them with causal edges from potential *social determinants of opportunity* (Definition 3.1) in contextual environments to the individual. For instance, in our running example of college admissions, instead of using `Race` $\rightarrow$ `Education Status` to model potential racial discrimination in college admission, we consider the household and geographic location related to the individual (through applicant's address) and model causal influences from contextual environments. Historical or current racial discrimination is not instantiated through inherent biological or intellectual differences across demographic groups. Instead, it manifests through patterns where individuals from certain racial backgrounds are more likely to reside in areas with weaker socio-economic profile and fewer educational resources (Sowell, 2004; Rothstein, 2017; Alexander, 2020).

More generally, disadvantage can perpetuate through many means, which are not necessarily limited to particular properties of any specific individual. Even if we completely eliminate discrimination in terms of directly rejecting admission on the basis of group membership, disadvantage can get perpetuated through *social determinants of opportunity* (Carmichael et al., 1967; Sowell, 1972; Tilly, 1998; Rothstein, 2017; Young, 2008; Powers & Faden, 2019; Alexander, 2020). As we will see in Section 4, our approach captures nuances in data generating processes and enables more fine-grained and to-the-point analyses for procedural fairness.

4 Theoretical Analyses on College Admission Procedures

In this section, we consider the practical scenario of college admissions and demonstrate the nuanced analyses our modeling framework can provide. In Section 4.1, we present a summary of the assumptions we use to facilitate closed-formula theoretical analyses. In Sections 4.2 – 4.4, we consider three mainstream college admission procedures. We provide discussions in Section 4.5.

4.1 Assumptions in Our Analyses

Beyond the emphasis on the protected feature, race, the influence from *social determinants of opportunity* is seldom discussed and somewhat overlooked in current algorithmic fairness literature.

Therefore, in our analyses, we take `Address Region` as a surrogate for *social determinants of opportunity*, and utilize our modeling approach to explicate their influence on individuals' academic preparedness.[3] For clear illustration through closed-formula theoretical derivation, we incorporate certain quantitative assumptions in our theoretical analyses of different admission procedures.

Assumption 4.1 (Region-Specific Demographic Makeup)**.** Let us denote the protected feature as A, where $a \in \mathcal{A}$ denotes under-represented minority (URM) applicant group, and $a' \in \mathcal{A}$ denotes non-URM applicant group. There are two regions where applicants reside in, rich and poor regions, with different demographic compositions from URM/non-URM groups,

$$\begin{array}{rcc} & \text{poor region} & \text{rich region} \\ \text{URM applicants} & n_a^{(\text{poor})} & n_a^{(\text{rich})} \\ \text{Non-URM applicants} & n_{a'}^{(\text{poor})} & n_{a'}^{(\text{rich})} \end{array}, \tag{1}$$

where the following inequalities hold true:

(1) disproportionate geographic distribution due to historical injustice, i.e., $n_a^{(\text{poor})} > n_{a'}^{(\text{poor})}$,

(2) the definition of "underrepresented minority", i.e., $n_a^{(\text{poor})} + n_a^{(\text{rich})} < n_{a'}^{(\text{poor})} + n_{a'}^{(\text{rich})}$.

Condition (1) specifies that URM applicants are relatively more concentrated in the less well-off region due to historical injustice (Sowell, 2004; Rothstein, 2017; Alexander, 2020). Condition (2) holds by definition, i.e., the total number of URM applicants is smaller than that for non-URM applicants.

Assumption 4.2 (Determinant of Academic Preparedness)**.** Conditioning on the affluence of the region where the applicant resides in, the academic preparedness is conditionally independent from the protected feature race. In other words, we have the following relation ($\perp\!\!\!\perp$ denotes independence):

$$\texttt{Academic Preparedness} \perp\!\!\!\perp \texttt{Race} \mid \texttt{Address Region}. \tag{2}$$

While there can be dependence between `Race` and `Academic Preparedness` (without conditioning on `Address Region`) due to historical injustice (Sowell, 2004; Rothstein, 2017; Alexander, 2020), such dependence does *not* indicate that `Race` is a determinant of applicant's `Academic Preparedness`. Assumption 4.2 specifies that after conditioning on applicant's address region, applicant's academic preparedness is irrelevant to the demographic group. In other words, `Address Region` encloses region-specific *social determinants of opportunity* related to academic preparedness, for instance, the availability of educational resources in the area and the environmental impacts on applicant's health, but `Race` is *not* an inherent determinant of applicant's academic preparedness

Assumption 4.3 (Gamma Parameterization of Academic Preparedness Distribution)**.** Let S denote the non-negative overall academic index score of an applicant's academic preparedness. Further let S_{MAX} and S_{MIN} denote the highest and lowest possible values of the score. Within any specific region $r \in \{\text{poor}, \text{rich}\}$, the log-converted relative score Q is Gamma distributed with region-specific shape and scale parameters, $k^{(r)}$ and $\theta^{(r)}$, respectively. Furthermore, the rich region's cumulative distribution function (CDF) of log-converted relative score Q dominates that of the poor region:

$$\begin{aligned} &Q \sim \Gamma\big(k^{(r)}, \theta^{(r)}\big), \text{ where } Q := -\log\left(\frac{S - S_{\text{MIN}}}{S_{\text{MAX}} - S_{\text{MIN}}}\right), r \in \{\text{poor}, \text{rich}\}, \\ &\forall q \in [0, \infty),\ F^{(\text{rich})}(q) \geq F^{(\text{poor})}(q), \text{ where } F^{(r)}(q) \text{ is the CDF of } \Gamma\big(k^{(r)}, \theta^{(r)}\big). \end{aligned} \tag{3}$$

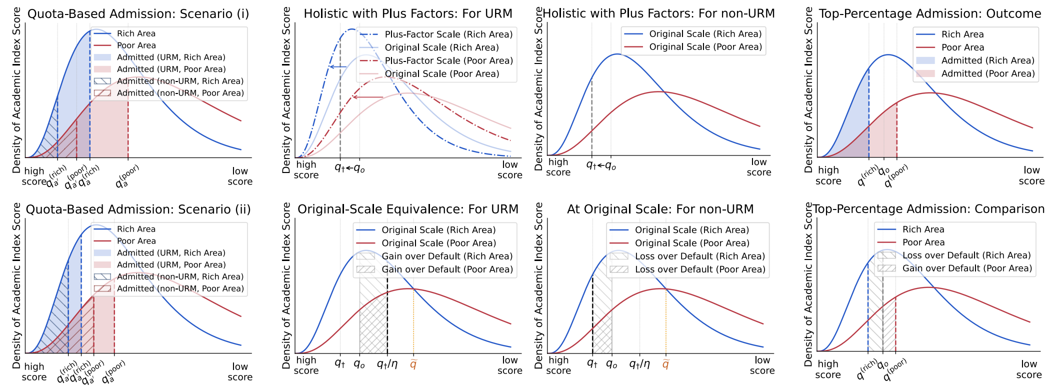
In Assumption 4.3, the conversion of the score maps the domain of values $[S_{\text{MIN}}, S_{\text{MAX}}]$ (higher score S is more competitive) to $[0, \infty)$, where the closer to 0 the converted score Q, the more competitive. The flexibility of Gamma distributions allows us to use combinations of shape and scale parameters to capture properties of the region-specific academic preparedness distribution.

Assumption 4.4 (Selective Admission and Open Enrollment)**.** The selective college employs thresholds on applicants' academic preparedness scores and has a limited availability of admissions g:

$$g < n, \text{ where } n = n_a^{(\text{poor})} + n_a^{(\text{rich})} + n_{a'}^{(\text{poor})} + n_{a'}^{(\text{rich})}, \tag{4}$$

and all applicants can get admitted to the open-enrollment college.

[3]For the purpose of this paper, we aim to demonstrate how our modeling framework dovetails ethical insights, and we do not intend to make any legal claim.



(a) Quota-based admission (b) Holistic review with plus factors (c) Top-percentage plan

Figure 2: Procedural fairness implications of different admission strategies. Panel (a): quota-based admission can introduce additional unfairness against non-URM applicants from the poor region. Panel (b): holistic review with plus factors tends to benefit URM applicants in the rich region more than these in the poor region. Panel (c): top-percentage plan transfer admission opportunity from the rich region to the poor region, and the redistribution is proportional to the natural region-specific demographic compositions.

Assumption 4.4 states that while the open-enrollment college can admit all applicants, the selective college uses score thresholds to distribute the limited availability of admissions. As we shall see in Sections 4.2 – 4.4, the exact values of thresholds depend on the admission strategy, and have procedural fairness implications in terms of the benefits and burdens experienced by individuals from different demographic groups, as well as regions with varying levels of affluence.[4]

4.2 Quota-Based Admissions

The quota-based admission is a type of affirmative-action admission strategy that sets specific limits on the number of admissions for applicants from different demographic backgrounds. This admission strategy was originally designed to rectify historical injustice by directly setting aside admission quotas to increase the representation of URM students. However, due to the rigid nature of the quota-based mechanism, this admission strategy has been controversial and addressed by the U.S. Supreme Court in the landmark case *University of California Regents v. Bakke (1978)* (Supreme Court, 1978). It was held that the use of strict racial quotas in college admission was unconstitutional, and was reaffirmed in another landmark case *Grutter v. Bollinger (2003)* (Supreme Court, 2003b).

Aside from the fact that the quota-based admission procedure is rigid and mechanical, it fails to account for the role of *social determinants of opportunity* which vary across regions and influence applicants' academic preparedness in different ways. As a result, employing quota-based admission can further disadvantage non-URM applicants from less well-off areas, effectively introducing additional unfairness during the attempt to rectify historical racial injustice:

Theorem 4.5 (Quota-Based Admission Incurs Unfairness w.r.t. Non-URM in Poor Region)**.** *Under Assumptions 4.1–4.4, let us denote with $\eta_{\text{quota}} \in \big[1, \frac{n}{n_a^{(\text{poor})}+n_a^{(\text{rich})}}\big]$ the weighting coefficient over the natural proportion of URM applicants in population, such that the quota for URM admissions in the selective college is $\eta_{\text{quota}} \cdot (\frac{n_a^{(\text{poor})}+n_a^{(\text{rich})}}{n} g)$. Then, the quota-based admission strategy imposes a more competitive requirements (in terms of score threshold) for non-URM applicants from the poor region, than that for URM applicants from the rich region, unless the following condition on region-specific academic preparedness CDF's is satisfied:*

$$\max_{q \in [0,\infty)} \frac{F^{(\text{rich})}(q)}{F^{(\text{poor})}(q)} \geq \frac{(n_{a'}^{(\text{poor})} + n_{a'}^{(\text{rich})})\eta_{\text{quota}}}{(n_a^{(\text{poor})} + n_a^{(\text{rich})})(1-\eta_{\text{quota}}) + (n_{a'}^{(\text{poor})} + n_{a'}^{(\text{rich})})} \, . \tag{5}$$

Theorem 4.5 states that, for quota-based admission to rectify historical racial injustice without introducing additional unfairness against non-URM individuals from less affluent areas, a rather

[4]Due to the space limit, we present proofs for our theoretical results in Appendix B.

strong condition must be met involving demographic composition, the academic preparedness across regions, and the imposed quota-based weighting coefficient η_{quota}, as summarized in Equation (5). In particular, the larger the quota (larger η_{quota}), the more spots are reserved for URM applicants (from both poor and rich regions), the more challenging for non-URM applicants in the poor region to be able to attend the selective college. In other words, non-URM applicants in the poor region, who face the same obstacles and disadvantages in contextual environments as their URM counterparts, are not reserved additional spots; on top of that, they have to compete with more advantaged peers (non-URM applicants from the rich region) over the spots that are already more limited. As we illustrate in Figure 2(a) Scenario (i), quota-based admission may result in a higher score threshold for non-URM in poor region than that for URM in rich region (since $q_{a'}^{(\text{poor})} < q_a^{(\text{rich})}$).

4.3 Holistic Review with Plus Factors

Holistic review with plus factors is another type of affirmative-action admission strategy, involving consideration of multiple factors that together define each individual applicant. The key element of this process is the use of plus factors, where certain characteristics, for instance, race and ethnic group, are given additional weight to promote diversity in the student body and rectify historical disadvantages. This approach was upheld by the U.S. Supreme Court in *Grutter v. Bollinger (2003)* (Supreme Court, 2003b), but was overruled in recent decisions for *Students for Fair Admissions (SFFA) v. Harvard & UNC (2023)* (Supreme Court, 2023a;b), effectively banning race-conscious admissions.

Putting aside the evolving jurisprudence, we aim to precisely characterize holistic review in terms of its implications on the distribution of benefits and burdens among individuals, when allocating the limited spots in selective college admissions. When taking into account of *social determinants of opportunity* signified by `Address Region`, we show that holistic review with plus factors may benefit applicants from better-off areas more than those from less well-off areas:

Theorem 4.6 (Holistic Review with Plus Factors Benefits URM in Rich Region More)**.** *Under Assumptions 4.1–4.4, let us denote with $\eta_\dagger < 1$ the multiplicative coefficient on the scale parameter of Gamma distributions for URM applicants' academic index scores, such that the perceived scores of URM applicants shift more probability density towards the high-score end. Let us denote with q_o the default threshold for selective admission, and with $q_\dagger$ the threshold if the admission procedure is a holistic review with plus factors. Further assume that region-specific shape parameters satisfy $k^{(\text{poor})} = k^{(\text{rich})} = k_o$. Then, the increase in the probability of selective admission for URM applicants from the rich region, is larger than that for URM applicants from the poor region:*

$$\textit{if the selective admission is limited in availability such that } q_o < \frac{k_o \ln(\theta^{(\text{poor})}/\theta^{(\text{rich})})}{1/\theta^{(\text{rich})} - 1/\theta^{(\text{poor})}}, \textit{ then}$$

$$\forall \eta_\dagger \in \left[\frac{q_o(1/\theta^{(\text{rich})} - 1/\theta^{(\text{poor})})}{k_o \ln(\theta^{(\text{poor})}/\theta^{(\text{rich})})}, 1\right), F^{(\text{rich})}(\frac{q_\dagger}{\eta_\dagger}) - F^{(\text{rich})}(q_o) > F^{(\text{poor})}(\frac{q_\dagger}{\eta_\dagger}) - F^{(\text{poor})}(q_o).$$

Theorem 4.6 characterizes different levels of benefits for URM applicants from different regions. Specifically, in terms of the increase in admission probability to the selective college, URM applicants from the rich region benefit more from the admission procedure that utilizes holistic review with plus factors, compared to URM applicants from the poor region. To better demonstrate our theoretical result, we provide illustrations in Figure 2(b).

As presented in top-row subfigures in Figure 2(b), at the original scale, the region-specific distributions of academic preparedness are the same for URM and non-URM applicants (Assumption 4.2). Holistic review with plus factors grants preference to URM applicants by perceiving their scores, at the distribution level, as if they were sampled from a distribution that is more concentrated at the high-score end (the plus-factor scale).[5] Because of the limited availability in selective admissions, the threshold $q_\dagger$ for admission under holistic review with plus factors is more competitive than the default q_o, i.e., $q_\dagger < q_o$, for both URM and non-URM applicants. While non-URM applicants are assessed on the original scale, URM applicants are evaluated on a plus-factor scale. Under the

[5] For holistic review with plus factors, we model its affirmative-action emphasis on the URM group through a distribution shift, i.e., from the original scale to the plus-factor scale, instead of an automatic awarding of points for each URM applicant. Our modeling choice is for the purpose of avoiding the introduction of rigid and mechanical characteristics to the process, as was addressed in *Gratz v. Bollinger (2003)* (Supreme Court, 2003a).

Gamma parameterization (Assumption 4.3), this is equivalent to employing a more competitive threshold $q_\dagger$ for non-URM applicants but a less competitive one $q_\dagger/\eta_\dagger$ for URM applicants, where $q_\dagger < q_o < q_\dagger/\eta_\dagger$. Although the mathematical form of $q_o < {k_o \ln(\theta^{(\mathrm{poor})}/\theta^{(\mathrm{rich})})}/{(1/\theta^{(\mathrm{rich})} - 1/\theta^{(\mathrm{poor})})}$ appears convoluted, the condition itself is relatively mild. Graphically speaking, the spots at the selective college are limited such that the threshold q_o does not reach the point where region-specific Gamma density curves (in the original scale) intersect, as depicted by $\widetilde{q}$ in Figure 2(b).

From the shaded areas in bottom-row subfigures in Figure 2(b), we can see that the increased admission probability for URM groups comes with a corresponding reduction in that for non-URM groups. However, such redistribution benefits URM applicants in the rich region more than those in the poor region, essentially disadvantaging URM applicants in less well-off areas.

4.4 Top-Percentage Plans

The top-percentage plans are college admission policies that guarantee admission to students who graduate in a certain top percentage of their high school classes. The top-percentage plans are generally not considered traditional affirmative-action admission strategies. Instead, these policies are race-neutral alternatives aiming to promote diversity by drawing students from a wide range of schools with different socioeconomic and geographic backgrounds, without explicitly considering race. A prominent example is the University of Texas's Top 10% Rule, which guarantees admission to students in the top 10% of their class. Another is the Eligibility in the Local Context (ELC) program of University of California, which was introduced after the 1996 California Proposition 209 banned the use of race, ethnicity, and gender in public university admissions in California.

Taking into account the demographic composition of applicants and the number of available spots at the selective college, we characterize the difference between top-percentage plans compared to the default selective admission. When explicitly considering the role of `Address Region` in applicants' academic preparedness, we show that the redistribution of limited selective admissions, as implied by top-percentage plans, is carried out by reallocating availability from the rich region to the poor region, regardless of the demographic group of applicants:

Theorem 4.7 (Top-Percentage Plans Reallocate Spots from Rich Region to Poor Region)**.** *Under Assumptions 4.1–4.4, let us denote with q_o the default threshold for selective admission, and with $q^{(\mathrm{poor})}$ and $q^{(\mathrm{rich})}$ the thresholds for poor and rich regions, respectively, if top-percentage plans are employed. Then, the increase in selective admissions (in terms of counts) for applicants from the poor region, comes from spots reallocated out of the rich region. This redistribution is a result of the top-percentage plans, and is not relevant to applicants' demographic group:*

$$\big(n_a^{(\mathrm{poor})} + n_{a'}^{(\mathrm{poor})}\big)\big[F^{(\mathrm{poor})}(q^{(\mathrm{poor})}) - F^{(\mathrm{poor})}(q^{(o)})\big] = \big(n_a^{(\mathrm{rich})} + n_{a'}^{(\mathrm{rich})}\big)\big[F^{(\mathrm{rich})}(q^{(o)}) - F^{(\mathrm{rich})}(q^{(\mathrm{rich})})\big].$$

Furthermore, if region-specific shape parameters satisfy $k^{(\mathrm{poor})} = k^{(\mathrm{rich})}$, we additionally have:

$$q^{(\mathrm{poor})}/q^{(\mathrm{rich})} = \theta^{(\mathrm{poor})}/\theta^{(\mathrm{rich})}.$$

Theorem 4.7 characterizes the reallocation of the selective admission spots performed by top-percentage plans. In Figure 2(c), we use shaded areas to illustrate the transfer of admission opportunity (in terms of the region-wise probability of selective admission) from the rich region to the poor region. The additional selective admissions gained by the poor region, compared to the default setting, are distributed proportionally to the natural demographic composition of each group.

4.5 Remark on Procedural Fairness Implications of Different Procedures

Although all three types of admission procedures share the goal of promoting fairness and diversity within the student body, the limited availability of selective admissions leads to varying redistributions of benefits and burdens among applicants. Quota-based admissions, while being rigid and mechanical, are more direct in reserving spots for URM applicants. However, as an unintended consequence, non-URM applicants from less well-off areas can be further disadvantaged when quota-based admissions are employed (Theorem 4.5). Holistic review with plus factors, in comparison, takes a more flexible approach when granting preferences to URM applicants. However, the increase in selective admission probability for URM applicants, which is reallocated from non-URM applicants, rewards the rich region more than the poor region (Theorem 4.6). Top-percentage plans, which provide race-neutral alternatives to the previous two affirmative-action strategies, transfer opportunities from rich region to poor region, operating in proportion to natural region-specific demographic compositions (Theorem 4.7).

The benefits and burdens experienced by applicants from different backgrounds in college admissions extend beyond whether or not and how the protected feature race is explicitly used in decision-making. Our theoretical results demonstrate the crucial role played by *social determinants of opportunity* enclosed in `Address Region` for procedural fairness analysis. Without them, it is impossible to identify the newly introduced unfairness, since the address variable is absent from the causal graph in previous literature (Kusner et al., 2017; Nabi & Shpitser, 2018; Chiappa, 2019; Wu et al., 2019).

5 Experiments

Commonly used data sets and benchmarks in algorithmic fairness literature tend to omit variables related to *social determinants of opportunity* (as we discussed in Section 3.1). However, the relative absence of comprehensive measurements does not render our framework unnecessary or ineffective. In this section, we demonstrate how to apply our analytical framework using the information available. We consider the publicly-available statistics for freshmen admissions to University of California, and reason about underlying academic preparedness from potential regions.[6]

5.1 Formulation of the Optimization Problem

Due to legal and ethical considerations (e.g., privacy protection and data confidentiality), the released data only contains summary statistics, and the detailed application or admission data is not publicly available. Nevertheless, we aim to utilize the information available and formulate a constrained optimization problem to estimate region-specific academic preparedness.

We do not regard race as a determinant of academic preparedness (Assumption 4.2), and incorporate the Gamma parameterization for region-specific distribution of academic preparedness among applicants (Assumption 4.3). Both the number of regions and demographic groups can take on values beyond the binary case. After specifying the number of regions, we formulate a constrained optimization problem to solve for region-specific shape and scale parameters, as well as demographic compositions across regions, where $\mathcal{L}(\cdot)$ denotes the loss function:

$$
\begin{aligned}
\min \quad & \mathcal{L}\Big(\substack{\text{demographic composition} \\ \text{(application \& admission)}}, \substack{\text{quantile statistics} \\ \text{(application \& admission)}}; k^{(R)}, \theta^{(R)}, q^{(R)}, n_A^{(R)}\Big) \\
s.t. \quad & \forall\, \text{race } a \in \mathcal{A},\ \sum\nolimits_r n_a^{(r)} \text{ matches demographic composition of applicants,} \\
& \forall\, \text{race } a \in \mathcal{A},\ \sum\nolimits_r n_a^{(r)} \cdot F^{(r)}(q^{(r)}) \text{ matches demographic composition of admissions,} \\
& \forall\, \text{specified } q^*, \sum\nolimits_r \Big[F^{(r)}(q^*) \cdot \sum\nolimits_a n_a^{(r)}\Big] \text{ matches application statistics,} \\
& \forall\, \text{specified } q^*, \sum\nolimits_r \Big[F^{(r)}\big(\min(q^*, q^{(r)})\big) \cdot \sum\nolimits_a n_a^{(r)}\Big] \text{ matches admission statistics,} \\
& \forall\, \text{region } r \in \mathcal{R},\ \text{the CDF (irrelevant to race) } F^{(r)}(q^{(r)}) := \int_0^{q^{(r)}} \Gamma(\xi; k^{(r)}, \theta^{(r)}) d\xi.
\end{aligned}
\tag{6}
$$

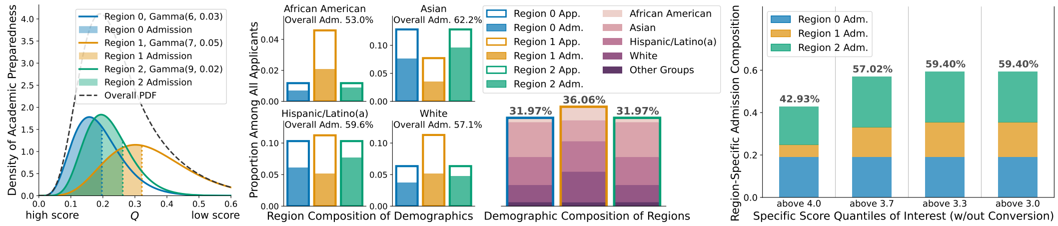
Here, q^*'s are certain quantiles specified in the publicly-available statistics provided by University of California undergrad admissions summary, that (before the relative log conversion) correspond to capped and weighted high-school GPA scores $\{4.0, 3.7, 3.3, 3.0\}$. We consider $\min(q^*, q^{(r)})$ when calculating estimated cumulative probabilities for admissions, $F^{(r)}\big(\min(q^*, q^{(r)})\big)$, because threshold values may differ across regions as a result of the employed admission procedure.[7]

5.2 Experimental Results

Despite the various constraints listed in Equation (6), the optimization problem can potentially remain under-constrained due to the limited information available provided by summary statistics. In practice, we solve the constrained optimization problem to match the estimation with

[6]The data is obtained from University of California undergrad admissions summary and freshmen fall admissions summary. We provide data descriptions as well as additional results and analyses in Appendix C.

[7]The 1996 California Proposition 209 banned the use of race, ethnicity, and gender in public university admissions. Therefore, thresholds are (potentially) region-specific but race-irrelevant, i.e., $q^{(r)}$ instead of $q_a^{(r)}$.



(a) Score densities (b) Race- and region- specific composition (c) Quantile-specific composition

Figure 3: Visualization of constrained optimization results fitted on University of California application and admission summary statistics. Panel (a): region-specific and overall densities of academic preparedness. Panel (b): for each group, the region-specific compositions of application and admission proportions (left four subplots); for each region, the demographic composition of applicants (right subplot). Panel (c): for specific quantiles of interest, the region composition of admitted students (in terms of the proportion among all applicants).

the university-wide statistics of capped and weighted high-school GPA scores (from year 2023).[8] We consider demographic groups recorded in the data, and limit the number of potential regions to three to avoid overfitting of summary statistics. In Figure 3, we present visualizations of the result of the constrained optimization, including the estimated region-wise and demographic composition of applicants, $n_a^{(r)}$, the parameters in region-specific Gamma distributions, $k^{(r)}$ and $\theta^{(r)}$, and the corresponding score thresholds $q^{(r)}$, where region $r \in \{0, 1, 2\}$ and race $a \in \{\text{African American}, \text{Asian}, \text{Hispanic/Latino(a)}, \text{White}, \text{Other Groups}\}$.

In Figure 3(a), we present region-specific densities of academic preparedness, as well as the overall density if we consider all applicants. The distinct shapes of region-specific densities reflect the varying influences on applicants' academic preparedness across different regions. For instance, the densities of Region 0 (blue) and Region 2 (green) concentrate more at the high-score end, compared to Region 1 (orange), indicating the more positive influence on applicant's academic preparedness. In Figure 3(b), in the left four subplots, for different demographic groups we present region-specific compositions of application and admission proportions; in the right-hand-side subplot, we present the demographic composition of applicants within each region. In Figure 3(c), we present the proportion (among all applicants) of admitted students whose scores are above specific quantiles. As we can see from Figure 3, there is a correlation between race and *social determinants of opportunity*, as indicated by different academic preparedness across regions, and also by the disproportionate demographic compositions of admission even if the procedure does not utilize race (as per 1996 California Proposition 209).

6 Concluding Remarks

In this paper, we advocate the explicit consideration of *social determinants of opportunity* in causal mechanisms for the purpose of understanding and achieving procedural fairness. We propose a modeling framework that encompasses variables characterizing influences from contexts and environments to individuals, namely, *social determinants of opportunity*. In the running example of college admissions, we demonstrate nuanced analyses that our framework facilitates, and explicate procedural fairness implications when different decision-making procedures are employed.

Because social determinants correlate with protected features, explicitly considering social determinants through which structural injustice potentially perpetuates can help us better understand the underlying data generating process. This, in turn, facilitates more precise and comprehensive characterizations of procedural fairness implications, and makes it more transparent to see benefits and burdens experienced by individuals with different demographic backgrounds as well as contexts and environments, when they are subject to different algorithmic decision-making procedures. Future works naturally include designing and utilizing appropriate measurements of *social determinants of opportunity* to develop fairness auditing and mitigation strategies, so that we can achieve procedural fairness in an effective, principled, and transparent way.

[8]Our implementation can be found at the anonymous Github repository `https://anonymous.4open.science/r/ProceduralFairnessSocialDeterminants`.

REFERENCES

Michelle Alexander. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press, 2020.

Noah Arthurs, Ben Stenhaug, Sergey Karayev, and Chris Piech. Grades are not normal: Improving exam score models using the logit-normal distribution. *International Educational Data Mining Society*, 2019.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Zachary Bleemer. Affirmative action, mismatch, and economic mobility after california's proposition 209. *The Quarterly Journal of Economics*, 137(1):115–160, 2022.

Paula Braveman and Laura Gottlieb. The social determinants of health: It's time to consider the causes of the causes. *Public Health Reports*, 129:19–31, 2014.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18. IEEE, 2009.

Wenbin Cao, Hui Wang, and Huihui Ying. The effect of environmental regulation on employment in resource-based areas of china—an empirical research based on the mediating effect model. *International Journal of Environmental Research and Public Health*, 14(12):1598, 2017.

Stokely Carmichael, Charles V Hamilton, and Stokely Carmichael. *Black Power*, volume 48. Random House New York, 1967.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

US Census Bureau. *American Community Survey Design and Methodology*. American Community Survey (ACS), 1.0 edition, 2009.

US Census Bureau. *American Community Survey Design and Methodology*. American Community Survey (ACS), 2.0 edition, 2014.

US Census Bureau. *2021 ACS 1-Year PUMS Data Dictionary*. American Community Survey (ACS), 2021.

US Census Bureau. *American Community Survey and Puerto Rico Community Survey Design and Methodology*. American Community Survey (ACS), 3.0 edition, 2022.

Raj Chetty, Will S Dobbie, Benjamin Goldman, Sonya Porter, and Crystal Yang. Changing opportunity: Sociological mechanisms underlying growing class gaps and shrinking race gaps in economic mobility. Technical report, National Bureau of Economic Research, 2024.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7801–7808, 2019.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

James S Coleman. Equality of educational opportunity. *Integrated Education*, 6(5):19–28, 1968.

James S Coleman. Social capital in the creation of human capital. *American Journal of Sociology*, 94:S95–S120, 1988.

Alexis J Comber, Chris Brunsdon, and Robert Radburn. A spatial analysis of variations in health access: Linking geography, socio-economic status and access perceptions. *International Journal of Health Geographics*, 10: 1–11, 2011.

Raewyn Connell. Poverty and education. *Harvard Educational Review*, 64(2):125–150, 1994.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 582–593, 2020.

Richard Delgado and Jean Stefancic. *Critical Race Theory: An Introduction*, volume 87. NYU Press, 2023.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6478–6490, 2021.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.

James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 1918–1921. IEEE, 2020.

Gilbert C Gee and Chandra L Ford. Structural racism and health inequities: Old issues, new directions. *Du Bois Review: Social Science Research on Race*, 8(1):115–132, 2011.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.

Miguel A Hernán and James M Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

David R Howell and Arne L Kalleberg. Declining job quality in the united states: Explanations and evidence. *The Russell Sage Foundation Journal of the Social Sciences*, 5(4):1–53, 2019.

Kosuke Imai and Zhichao Jiang. Principal fairness for human and algorithmic decision-making. *arXiv preprint arXiv:2005.10400*, 2020.

Christopher Jencks. Inequality: A reassessment of the effect of family and schooling in america, 1972.

Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pp. 240–248, 2019.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pp. 2564–2572. PMLR, 2018.

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 656–666, 2017.

Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR, 2018.

Joshua Loftus, Chris Russell, Matt Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

Jens Ludwig, Lisa Sanbonmatsu, Lisa Gennetian, Emma Adam, Greg J Duncan, Lawrence F Katz, Ronald C Kessler, Jeffrey R Kling, Stacy Tessler Lindau, Robert C Whitaker, et al. Neighborhoods, obesity, and diabetes–a randomized social experiment. *New England Journal Of Medicine*, 365(16):1509–1519, 2011.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*, 2020.

Ioannis Manisalidis, Elisavet Stavropoulou, Agathangelos Stavropoulos, and Eugenia Bezirtzoglou. Environmental and health impacts of air pollution: A review. *Frontiers in Public Health*, 8:14, 2020.

Michael Marmot and Richard Wilkinson. *Social Determinants of Health*. OUP Oxford, 2005.

Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pp. 4382–4391, 2019.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021.

Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 386–400, 2021.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 1931–1940, 2018.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. In *International Conference on Machine Learning*, pp. 4674–4682. PMLR, 2019.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Optimal training of fair predictive models. In *Conference on Causal Learning and Reasoning*, pp. 594–617. PMLR, 2022.

Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, volume 1170, pp. 3, 2018.

Hamed Nilforoshan, Johann D Gaebler, Ravi Shroff, and Sharad Goel. Causal conceptions of fairness and their consequences. In *International Conference on Machine Learning*, pp. 16848–16887. PMLR, 2022.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

Madison Powers and Ruth Faden. *Structural Injustice: Power, Advantage, and Human Rights*. Oxford University Press, 2019.

Bingtao Qin, Lei Liu, Le Yang, and Liming Ge. Environmental regulation and employment in resource-based cities in china: The threshold effect of industrial structure transformation. *Frontiers in Environmental Science*, 10, 2022.

John Rawls. *A Theory of Justice*. Cambridge: Harvard University Press, 1971.

John Rawls. *Justice as Fairness: A Restatement*. Harvard University Press, 2001.

Michael Redmond. Communities and Crime. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C53W3X.

Kyriaki Remoundou and Phoebe Koundouri. Environmental effects on public health: An economic perspective. *International Journal of Environmental Research and Public Health*, 6(8):2160–2178, 2009.

Dorothy Roberts. *Fatal invention: How science, politics, and big business re-create race in the twenty-first century*. New Press/ORIM, 2011.

Whitney R Robinson, Audrey Renson, and Ashley I Naimi. Teaching yourself about structural racism will improve your machine learning. *Biostatistics*, 21(2):339–344, 2020.

Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638, 2014.

Pauline M Rose and Caroline Dyer. Chronic poverty and education: A review of literature. *Chronic Poverty Research Centre Working Paper*, (131), 2008.

Richard Rothstein. *The Color of Law: A Forgotten History of How Our Government Segregated America*. Liveright Publishing, 2017.

Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pp. 793–810, 2019.

Amartya Sen. *The Idea of Justice*. Harvard University Press, 2011.

Audrey Smedley. *Race in North America: Origin and Evolution of a Worldview*. Routledge, 2018.

Thomas Sowell. *Black Education: Myths and Tragedies*. ERIC, 1972.

Thomas Sowell. *Affirmative Action Around the World: An Empirical Study*. Yale University Press, 2004.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer New York, 1993.

US Supreme Court. *University of California Regents v. Bakke*. Number 76-811. 438 U.S. 265, 1978.

US Supreme Court. *Gratz v. Bollinger*. Number 02-516. 539 U.S. 244, 2003a.

US Supreme Court. *Grutter v. Bollinger*. Number 02-241. 539 U.S. 306, 2003b.

US Supreme Court. *Fisher v. University of Texas at Austin*. Number 11-345. 570 U.S. 297, 2013a.

US Supreme Court. *Fisher v. University of Texas at Austin*. Number 14-981. 579 U.S. 365, 2013b.

US Supreme Court. *Students for Fair Admissions, Inc. v. President and Fellows of Harvard College*. Number 20-1199. 600 U.S. 181, 2023a.

US Supreme Court. *Students for Fair Admissions, Inc. v. University of North Carolina*. Number 21-707. 600 U.S. 181, 2023b.

Tina Q Tan, Ravina Kullar, Talia H Swartz, Trini A Mathew, Damani A Piggott, and Vladimir Berthaud. Location matters: Geographic disparities and impact of coronavirus disease 2019. *The Journal of Infectious Diseases*, 222(12):1951–1954, 2020.

Zeyu Tang, Yatong Chen, Yang Liu, and Kun Zhang. Tier Balancing: Towards dynamic fairness over underlying causal factors. In *International Conference on Learning Representations*, 2023a.

Zeyu Tang, Jiji Zhang, and Kun Zhang. What-is and how-to for fairness in machine learning: A survey, reflection, and perspective. *ACM Computing Surveys*, 55(13s):1–37, 2023b. ISSN 0360-0300.

Zeyu Tang, Jialu Wang, Yang Liu, Peter Spirtes, and Kun Zhang. Procedural fairness through decoupling objectionable data generating components. In *International Conference on Learning Representations*, 2024.

Jandhyala BG Tilak. Education and poverty. *Journal of Human Development*, 3(2):191–207, 2002.

Charles Tilly. *Durable Inequality*. University of California Press, 1998.

Edwin AJ van Hooft, John D Kammeyer-Mueller, Connie R Wanberg, Ruth Kanfer, and Gokce Basbug. Job search and employment success: A quantitative review and future research agenda. *Journal of Applied Psychology*, 106(5):674, 2021.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pp. 1–7. IEEE, 2018.

Rueben Warren, Bailus Walker Jr, and Vincent R Nathan. Environmental factors influencing public health and medicine: Policy implications. *Journal of the National Medical Association*, 94(4):185, 2002.

Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. Pc-fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3399–3409, 2019.

Ruqaiijah Yearby. Racial disparities in health status and access to healthcare: The continuation of inequality in the united states due to structural racism. *American Journal of Economics and Sociology*, 77(3-4):1113–1152, 2018.

Ruqaiijah Yearby, Brietta Clark, and José F Figueroa. Structural racism in historical and modern us health care policy: Study examines structural racism in historical and modern US health care policy. *Health Affairs*, 41 (2):187–194, 2022.

Kyung-Jin Yeum, Byeng Chun Song, and Nam-Seok Joo. Impact of geographic location on vitamin D status and bone mineral density. *International Journal of Environmental Research and Public Health*, 13(2):184, 2016.

Iris Marion Young. Structural injustice and the politics of difference. In *Intersectionality and Beyond*, pp. 289–314. Routledge-Cavendish, 2008.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 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.

Junzhe Zhang and Elias Bareinboim. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*, volume 31, 2018a.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making – the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b.

Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 3929–3935, 2017.

Xueru Zhang and Mingyan Liu. Fairness in learning-based sequential decision algorithms: A survey. In *Handbook of Reinforcement Learning and Control*, pp. 525–555. Springer, 2021.

Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? In *Advances in Neural Information Processing Systems*, volume 33, pp. 18457–18469, 2020.

Seth D Zimmerman. The returns to college admission for academically marginal students. *Journal of Labor Economics*, 32(4):711–754, 2014.

SUPPLEMENT TO "PROCEDURAL FAIRNESS THROUGH ADDRESSING SOCIAL DETERMINANTS OF OPPORTUNITY"

Anonymous authors
Paper under double-blind review

LIST OF RESPONSES TO REVIEWERS

Re: All Reviewers 1
Re: Q5.1 by Reviewer h9Xa 4
Re: Q5.2 by Reviewer h9Xa 19
Re: C2 by Reviewer g5W4 20
Re: C3 by Reviewer g5W4 20
Re: C1 by Reviewer h9Xa 24
Re: Q6 by Reviewer g5W4 25
Re: C3 by Reviewer h9Xa 25
Re: C2 by Reviewer h9Xa 26
Re: Q2 by Reviewer 8F2j 26

Table of Contents: Appendix

A Further Discussions on Related Works **18**

A.1 Fairness Notions Based on Observational Statistics and Causal Analysis 18

A.2 Detailed Comparison with Causal Fairness Approaches 19

A.2.1 Question of Interest 19

A.2.2 No Conflict in Principle with Causal Fairness 19

A.3 Different Uses of Term “Structure” in Related Disciplines 20

A.4 Common Presence of *Social Determinants of Opportunity* 20

B Proofs of Theoretical Results **21**

B.1 Proof of Theorem 4.5 in Section 4.2 21

B.2 Proof of Theorem 4.6 in Section 4.3 22

B.3 Proof of Theorem 4.7 in Section 4.4 23

C Additional Results and Discussions on Empirical Analyses **24**

C.1 Empirical Analyses on University of California Undergrad Admission 24

C.1.1 Description of the Data 24

C.1.2 Gamma Parameterization for Score Distribution 25

C.1.3 Further Discussions on Empirical Results 25

C.2 Additional Analyses on US Census Data 26

C.2.1 Age Structure of Population in PUMAs 26

C.2.2 Racial Composition in PUMAs 26

C.2.3 Occupation Distribution in PUMAs 27

C.2.4 Combination of Factors in PUMA 27

Table 1: Summary of comparisons between our approach and closely related previous works.

Fairness considerations	Avoiding (potential) recapitulation of inappropriate stereotypes in causal modeling	Precisely defined disadvantaged individuals	Emphasize on procedural fairness implications	Include *social determinants of opportunity*
Observational statistics (Dwork et al., 2012; Hardt et al., 2016; Zafar et al., 2017)	not applicable	✘	✘	✘
Path-specific (interventional) causal effect (Kilbertus et al., 2017; Zhang et al., 2017; Nabi & Shpitser, 2018; Nabi et al., 2019; 2022; Salimi et al., 2019)	✘	✘	✓	✘
(Path-specific) counterfactual causal effect (Kusner et al., 2017; Chiappa, 2019; Wu et al., 2019)	✘	✘	✓	✘
Intersectional definition of subgroups (Kearns et al., 2018; Foulds et al., 2020)	not applicable	✓	✘	✘
Our approach	✓	✓	✓	✓

A FURTHER DISCUSSIONS ON RELATED WORKS

In this section, we present further discussions on related works. In Section A.1, we consider types of information utilized when characterizing algorithmic fairness, and their relative emphases. In Section A.2, we provide a detailed comparison between our approach and previous works on causal fairness. In Section A.3, we present an additional remark on the use of term "structure" in related disciplines. In Section A.4, we discuss the common presence of *social determinants of opportunity* in various practical scenarios. We summarize the comparisons of our approach and the previous literature in Table 1.

A.1 FAIRNESS NOTIONS BASED ON OBSERVATIONAL STATISTICS AND CAUSAL ANALYSIS

Various notions have been proposed in the algorithmic fairness literature to characterize fairness with respect to the prediction or the prediction-based decision-making (Dwork et al., 2012; Hardt et al., 2016; Chouldechova, 2017; Zafar et al., 2017), and also notions that are based on causal modeling of the data generating process (Kusner et al., 2017; Kilbertus et al., 2017; Nabi & Shpitser, 2018; Chiappa, 2019; Wu et al., 2019; Coston et al., 2020). Recent survey papers have presented overviews on fairness notions in static settings (Loftus et al., 2018; Makhlouf et al., 2020; Mehrabi et al., 2021), dynamic settings (Zhang & Liu, 2021), and also the connection between algorithmic fairness and the literature from moral and political philosophy (Tang et al., 2023b).

The type of information utilized reflects different emphases of algorithmic fairness studies. Notions based on observational statistics analyze the fairness implications in terms of the *outcome* of predictions or decision-making (Dwork et al., 2012; Hardt et al., 2016; Chouldechova, 2017; Zafar et al., 2017; Kearns et al., 2018; Foulds et al., 2020). Approaches that capture causal influences from the protected feature to the target variable at the individual-level (Kusner et al., 2017; Kilbertus et al., 2017; Nabi & Shpitser, 2018; Chiappa, 2019; Wu et al., 2019) and the (sub-)group-level (Coston et al., 2020; Imai & Jiang, 2020; Mishler et al., 2021) put more emphases on the *procedural* aspect of algorithmic fairness inquiries, focusing on the data generating process of interest. Recent work has also proposed to address procedural fairness over all objectionable data generating components (Tang et al., 2024) according to John Rawls's advocacy for pure procedural justice (Rawls, 1971; 2001).

A.2 Detailed Comparison with Causal Fairness Approaches

Among previous algorithmic fairness approaches, causal fairness analyses are most closely related to our work since they also emphasize the role of data generating process (Section A.1). In this subsection, we provide a detailed comparison between our approach and previous works on causal fairness, in terms of the question of interest (Section A.2.1), and whether or not our framework are in tension with previous causal fairness approaches (Section A.2.2).

Reviewer h9Xa: **Q5.2**

We include a detailed compasion between our framework and previous approach.

A.2.1 Question of Interest

To avoid overloading the term "counterfactual" in the causal inference literature (Spirtes et al., 1993; Pearl, 2009; Peters et al., 2017), we use "counter-factual" (with a hyphen, as an opposite to "factual") to denote that something does not happen in the current reality. Previous causal fairness approaches have utilized interventional (Kilbertus et al., 2017; Nabi & Shpitser, 2018; Nabi et al., 2019; 2022) and/or counterfactual (Kusner et al., 2017; Chiappa, 2019; Wu et al., 2019) causal effects in the technical formulation, and aim to answer the following question:

Question A.1 (**Counter-Factual Analysis Starting from Protected Features**). Under certain conditions and assumptions, what would happen to the predicted outcome in the factual world and the counter-factual world, had **the protected feature(s)** taken different values?

Based on estimating or bounding certain causal effects among variables, including the protected feature, the (predicted) outcome, and certain variables that are closely related to but not the protected feature itself, e.g., proxy variables (Kilbertus et al., 2017), redlining attributes (Zhang et al., 2017), admissible variables (Salimi et al., 2019), and so on, the fairness violation is quantified in terms of causal effects between the protected feature and the (predicted) outcome. There is a reductive focus solely upon the protected feature when modeling the discrimination. For instance, it is a common practice for causal fairness notions to consider varying the value of protected feature (Kilbertus et al., 2017; Kusner et al., 2017; Nabi & Shpitser, 2018; Nabi et al., 2019; 2022; Chiappa, 2019; Wu et al., 2019) as the starting point. Recently, Tang et al. (2024) have also proposed to consider not only edges or paths originating from the protected feature, but also all objectionable components in the data generating process, to address procedural fairness.

However, the modeling choice of "summarizing" discrimination only through edges/paths originating from protected feature, or solely among individual-level variables, falls short of the need to capture procedural unfairness and structural injustice. The characteristics of the environment and the context that individuals operate in typically do not correspond to individual-level attributes, and are not considered in previous literature. Different from causal fairness approaches, our framework explicitly incorporates the influence of contextual environments, and aims to address the following question:

Question A.2 (**Factual Analysis Starting from Social Determinants of Opportunity**). Under certain conditions and assumptions, what are the aspects of the data generating process that characterize **the influence from contextual environments to the individual**?

As we discussed in Section 3, while *social determinants of opportunity* often correlate with individual-level attributes, they cannot be captured by features of any particular individual. Explicit consideration and modeling of *social determinants of opportunity* facilitate a more comprehensive understanding of the benefits and burdens experienced by individuals from diverse demographic backgrounds as well as contextual environments, which is essential for understanding and achieving procedural fairness effectively and transparently.

A.2.2 No Conflict in Principle with Causal Fairness

In principle, our framework is not in conflict with previous causal fairness approaches, and the two complement each other. Both our framework and previous causal fairness approaches aim to model the data generating process, and both emphasize the procedural fairness implications.

However, our framework extends the scope of consideration beyond individual-level variables, and explicitly incorporates the influence of contextual environments. For instance, when operationalizing our framework, we do not drop relevant variables, e.g., the `Address` of an individual, which is often omitted in previous literature (Kilbertus et al., 2017; Kusner et al., 2017; Nabi & Shpitser, 2018; Chiappa, 2019; Wu et al., 2019; Mary et al., 2019; Ding et al., 2021). Furthermore, the findings of our analyses suggest that we should utilize all information available, and furthermore, actively look for and develop better measurements for *social determinants of opportunity*, so that we can better understand and address procedural unfairness and structural injustice. Future works naturally include the development of causal effect estimands that incorporate both individual-level attributes and *social determinants of opportunity*, and our framework and previous causal fairness approaches can be used in conjunction to achieve the goal.

A.3 Different Uses of Term "Structure" in Related Disciplines

The term "structure" and "structural" are utilized in different ways by related disciplines. For the literature of causal learning and reasoning, the term "structure" and "structural" are often used to describe how causal structures look like among variables of interest (Spirtes et al., 1993; Pearl, 2009; Peters et al., 2017; Hernán & Robins, 2020), e.g., in terms of causal graphs and/or structural equation models (SEMs). For the literature of structural justice and social determinants of health, the term "structural" is used to denote the systemic ways in which society is organized, e.g., through policies, laws, and social norms, that perpetuate discrimination and animus towards certain groups (Carmichael et al., 1967; Sowell, 1972; Tilly, 1998; Yearby, 2018; Robinson et al., 2020; Alexander, 2020; Yearby et al., 2022). There are interests in the social determinants of health literature to use DAGs as a tool for illustrative purposes, abstracting key concepts or areas that are interrelated at a high level, and modeling the mechanism through which structural forms of discriminations get realized (racism, sexism, etc.) (Robinson et al., 2020; Yearby et al., 2022).

A.4 Common Presence of *Social Determinants of Opportunity*

To strike a balance between a broad discussion and a case study, we considered a concrete empirical setting of college admissions in the main paper, and demonstrate the nuanced analyses our framework facilitates. However, the implications of explicitly and carefully considering *social determinants of opportunity* are not limited to the college admissions setting. In this section, we discuss the common presence of *social determinants of opportunity* in various practical scenarios, where influence of contextual environments on individuals is often substantial.

Reviewer g5W4: **C2**
The presence of *social determinants of opportunity* is ubiquitous and not limited to college admissions.

***Social Determinants of Opportunity* – Health** In terms of the influence of environments on individual's health, previous literature has considered how environmental hazards disproportionately affect low-income populations and communities of color (Warren et al., 2002), how indoor air pollution affects women and children in low-income regions (Manisalidis et al., 2020), and the structural implications of social determinants of health on how society should be organized (Robinson et al., 2020; Yearby et al., 2022). More broadly, a review on economic research has also been conducted to show how environmental changes impact public health in both developed and developing countries (Remoundou & Koundouri, 2009).

***Social Determinants of Opportunity* – Education** In terms of the influence of environments on individual's educational attainments, previous literature has considered how the quality of schools and the availability of educational resources affect students' academic performance (Coleman, 1968; 1988), how the family and neighborhood environments influence education (Jencks, 1972), and implications of various affirmative-action policies (usually under different names) across countries with different histories and cultures (Sowell, 2004).

Reviewer g5W4: **C3**
Similar policies exist in different countries, and the implications of our analyses are not limited to US.

***Social Determinants of Opportunity* – Employment** In terms of the influence of environments on individual's employment opportunities, previous literature has considered the relation between the employment of residents and the rationalization and optimization level of region's industrial structures (Cao et al., 2017; Qin et al., 2022), the psychological perspective of (e.g., influence from collective values of community) job search behaviors (van Hooft et al., 2021), and how the employment rate of residents is influenced by job quality (Howell & Kalleberg, 2019).

B Proofs of Theoretical Results

In this section, we present proofs of our theoretical results.

B.1 Proof of Theorem 4.5 in Section 4.2

Theorem (Quota-Based Admission Incurs Unfairness w.r.t. Non-URM in Poor Region)**.** *Under Assumptions 4.1–4.4, let us denote with $\eta_{\text{quota}} \in \big[1, \frac{n}{n_a^{(\text{poor})}+n_a^{(\text{rich})}}\big]$ the weighting coefficient over the natural proportion of URM applicants in population, such that the quota for URM admissions in the selective college is $\eta_{\text{quota}} \cdot (\frac{n_a^{(\text{poor})}+n_a^{(\text{rich})}}{n} g)$. Then, the quota-based admission strategy imposes a more competitive requirements (in terms of score threshold) for non-URM applicants from the poor region, than that for URM applicants from the rich region, unless the following condition on region-specific academic preparedness CDF's is satisfied:*

$$\max_{q \in [0,\infty)} \frac{F^{(\text{rich})}(q)}{F^{(\text{poor})}(q)} \geq \frac{(n_{a'}^{(\text{poor})} + n_{a'}^{(\text{rich})})\eta_{\text{quota}}}{(n_a^{(\text{poor})} + n_a^{(\text{rich})})(1-\eta_{\text{quota}}) + (n_{a'}^{(\text{poor})} + n_{a'}^{(\text{rich})})} \,. \tag{B.1}$$

Proof. Quota-based admission reserves certain number of selective admission spots for the URM group, weighted by a coefficient $\eta_{\text{quota}} > 1$ over natural proportion of URM applicants, i.e., $\eta_{\text{quota}} \cdot (\frac{n_a^{(\text{poor})}+n_a^{(\text{rich})}}{n} g)$. Then, the available selective admission spots for the non-URM group is $g - \eta_{\text{quota}} \cdot (\frac{n_a^{(\text{poor})}+n_a^{(\text{rich})}}{n} g)$.

For the convenience of notation, let us denote η'_{quota} the weight coefficients for the non-URM group over the natural proportion of non-URM applicants in the population, such that:

$$\eta'_{\text{quota}} \cdot (\frac{n_{a'}^{(\text{poor})} + n_{a'}^{(\text{rich})}}{n} g) = g - \eta_{\text{quota}} \cdot (\frac{n_a^{(\text{poor})} + n_a^{(\text{rich})}}{n} g), \tag{B.2}$$

Notice that $\eta'_{\text{quota}} \in [0,1]$ since $\eta_{\text{quota}} \in \big[1, \frac{n}{n_a^{(\text{poor})}+n_a^{(\text{rich})}}\big]$. Additionally, η'_{quota} is not an additional parameter whose value can vary freely, and it is fully determined by the numeric relation specified in Equation (B.2).

Because of the limited availability of selective admissions g, when employing the quota-based admission strategy, the score thresholds for each group will change as a result of the introduced quota requirements specified by weighting factors η_{quota} and η'_{quota}. In particular, under Assumptions 4.1–4.4, the number of selective admissions for each group is calculated by the weighted sum (according to the probability of getting admitted to the selective college) of applicants from the group across regions, and the selective admission counts need to satisfy the quota requirements:

$$\begin{aligned} n_a^{(\text{poor})} \cdot F^{(\text{poor})}\big(q_a^{(poor)}\big) + n_a^{(\text{rich})} \cdot F^{(\text{rich})}\big(q_a^{(rich)}\big) &= \eta_{\text{quota}} \cdot (\frac{n_a^{(\text{poor})} + n_a^{(\text{rich})}}{n} g), \\ n_{a'}^{(\text{poor})} \cdot F^{(\text{poor})}\big(q_{a'}^{(poor)}\big) + n_{a'}^{(\text{rich})} \cdot F^{(\text{rich})}\big(q_{a'}^{(rich)}\big) &= \eta'_{\text{quota}} \cdot (\frac{n_{a'}^{(\text{poor})} + n_{a'}^{(\text{rich})}}{n} g). \end{aligned} \tag{B.3}$$

Since the quota-based admission strategy ensures Equation (B.3) is satisfied given the region-specific demographic makeup (Assumption 4.1), we have:

$$F^{(\text{poor})}\big(q_a^{(poor)}\big) = \frac{g \cdot \eta_{\text{quota}}}{n} = F^{(\text{rich})}\big(q_a^{(rich)}\big), \tag{B.4}$$

$$F^{(\text{poor})}\big(q_{a'}^{(poor)}\big) = \frac{g \cdot \eta'_{\text{quota}}}{n} = F^{(\text{rich})}\big(q_{a'}^{(rich)}\big). \tag{B.5}$$

Let us consider the left-hand-side (LHS) and right-hand-side (RHS) of each equation.

- LHS equals to RHS of Equation (B.4): since $F^{(\text{rich})}$ dominates $F^{(\text{poor})}$ (Assumption 4.3), we have $q_a^{(poor)} > q_a^{(rich)}$, i.e., among URM applicants, the threshold for the raw score in the poor region is lower than that for the rich region.

- LHS equals to RHS of Equation (B.5): for the same reason as above, we have $q_{a'}^{(poor)} > q_{a'}^{(rich)}$, i.e., among non-URM applicants, the threshold for the raw score in the poor region is lower than that for the rich region.
- LHS of Equation (B.4) and LHS of Equation (B.5): since $\eta'_{\text{quota}} < 1 < \eta_{\text{quota}}$, we have $q_a^{(poor)} > q_{a'}^{(poor)}$, i.e., for the poor region, the threshold for the raw score of URM applicants is lower than that for non-URM applicants.
- RHS of Equation (B.4) and RHS of Equation (B.5): for the same reason as above, we have $q_a^{(rich)} > q_{a'}^{(rich)}$, i.e., for the rich region, the threshold for the raw score of URM applicants is lower than that for non-URM applicants.

However, the relative magnitude relation between $q_{a'}^{(poor)}$ (for non-URM applicants residing in the poor region) and $q_a^{(rich)}$ (for URM applicants residing in the rich region) can go either way. Specifically, we can show that if $\max_{q\in[0,\infty)} \frac{F^{(\text{rich})}(q)}{F^{(\text{poor})}(q)} < \frac{\eta_{\text{quota}}}{\eta'_{\text{quota}}}$, then $q_{a'}^{(poor)} < q_a^{(rich)}$, i.e., the threshold at the raw score for non-URM applicants in the poor region is higher than that for URM applicants from the rich region:

$$\text{when } \max_{q\in[0,\infty)} \frac{F^{(\text{rich})}(q)}{F^{(\text{poor})}(q)} < \frac{\eta_{\text{quota}}}{\eta'_{\text{quota}}}, \text{ we have } \frac{\eta_{\text{quota}}}{\eta'_{\text{quota}}} \cdot F^{(\text{poor})}(q_{a'}^{(poor)}) > F^{(\text{rich})}(q_{a'}^{(poor)}), \tag{B.6}$$

and at the same time

$$\frac{\eta_{\text{quota}}}{\eta'_{\text{quota}}} \cdot F^{(\text{poor})}(q_{a'}^{(poor)}) \overset{(i)}{=} F^{(\text{poor})}(q_a^{(poor)}) \overset{(ii)}{=} F^{(\text{rich})}(q_a^{(rich)}), \tag{B.7}$$

where (i) results from Equations B.4 and B.5, and (ii) follows Equation (B.4).

Because $F^{(\text{rich})}(q_a^{(rich)}) > F^{(\text{rich})}(q_{a'}^{(poor)})$ and the CDF function $F^{(\text{rich})}(\cdot)$ is non-decreasing, we have $q_{a'}^{(poor)} < q_a^{(rich)}$. In other words, as a necessary condition to prevent this, we need

$$\max_{q\in[0,\infty)} \frac{F^{(\text{rich})}(q)}{F^{(\text{poor})}(q)} \geq \frac{\eta_{\text{quota}}}{\eta'_{\text{quota}}}, \tag{B.8}$$

after re-arranging, and incorporating Equation (B.2), gives us

$$\max_{q\in[0,\infty)} \frac{F^{(\text{rich})}(q)}{F^{(\text{poor})}(q)} \geq \frac{(n_{a'}^{(\text{poor})} + n_{a'}^{(\text{rich})})\eta_{\text{quota}}}{(n_a^{(\text{poor})} + n_a^{(\text{rich})})(1-\eta_{\text{quota}}) + (n_{a'}^{(\text{poor})} + n_{a'}^{(\text{rich})})} \ .$$

□

B.2 Proof of Theorem 4.6 in Section 4.3

Theorem (Holistic Review with Plus Factors Benefits URM in Rich Region More)**.** *Under Assumptions 4.1–4.4, let us denote with $\eta_\dagger < 1$ the multiplicative coefficient on the scale parameter of Gamma distributions for URM applicants' academic index scores, such that the perceived scores of URM applicants shift more probability density towards the high-score end. Let us denote with q_o the default threshold for selective admission, and with $q_\dagger$ the threshold if the admission procedure is a holistic review with plus factors. Further assume that region-specific shape parameters satisfy $k^{(\text{poor})} = k^{(\text{rich})} = k_o$. Then, the increase in the probability of selective admission for URM applicants from the rich region, is larger than that for URM applicants from the poor region:*

$$\textit{if the selective admission is limited in availability such that } q_o < \frac{k_o \ln(\theta^{(\text{poor})}/\theta^{(\text{rich})})}{1/\theta^{(\text{rich})} - 1/\theta^{(\text{poor})}}, \textit{ then}$$

$$\forall \eta_\dagger \in \left[\frac{q_o(1/\theta^{(\text{rich})} - 1/\theta^{(\text{poor})})}{k_o \ln(\theta^{(\text{poor})}/\theta^{(\text{rich})})}, 1\right), F^{(\text{rich})}(\frac{q_\dagger}{\eta_\dagger}) - F^{(\text{rich})}(q_o) > F^{(\text{poor})}(\frac{q_\dagger}{\eta_\dagger}) - F^{(\text{poor})}(q_o).$$

Proof. The holistic review with plus factors changes the scale parameter of the Gamma distribution corresponding to URM applicants' academic index scores, from the original scale, i.e., $\Gamma(k_o, \theta^{(r)})$, to the plus-factor scale, i.e., $\Gamma(k_o, \eta_\dagger \cdot \theta^{(r)})$, where $r \in \{\text{poor}, \text{rich}\}$. The admission procedure does not change how non-URM applicants' scores are perceived, i.e., it remains at the original scale, $\Gamma(k_o, \theta^{(r)})$.

Then, we can calculate the default threshold q_o and that when the admission strategy is employed, $q_\dagger$, as follows:

$$(n_a^{(\text{poor})} + n_{a'}^{(\text{poor})}) \cdot F^{(\text{poor})}(q_o) + (n_a^{(\text{rich})} + n_{a'}^{(\text{rich})}) \cdot F^{(\text{rich})}(q_o) = g, \tag{B.9}$$

$$n_a^{(\text{poor})} \cdot F_\dagger^{(\text{poor})}(q_\dagger) + n_a^{(\text{rich})} \cdot F_\dagger^{(\text{rich})}(q_\dagger) + n_{a'}^{(\text{poor})} \cdot F^{(\text{poor})}(q_\dagger) + n_{a'}^{(\text{rich})} \cdot F^{(\text{rich})}(q_\dagger) = g, \tag{B.10}$$

where $F^{(r)}(\cdot)$ is the CDF of $\Gamma(k_o, \theta^{(r)})$, and $F_\dagger^{(r)}(\cdot)$ is that of $\Gamma(k_o, \eta_\dagger \cdot \theta^{(r)})$.

Because of the numerical property of Gamma CDF's, we have:

$$\forall q \in [0, \infty), \quad F_\dagger^{(r)}(q) = \frac{1}{\Gamma(k)} \gamma\big(k_o, \frac{q}{\eta_\dagger \cdot \theta^{(r)}}\big) = \frac{1}{\Gamma(k)} \gamma\big(k_o, \frac{q/\eta_\dagger}{\theta^{(r)}}\big) = F^{(r)}\big(\frac{q}{\eta_\dagger}\big), \tag{B.11}$$

where $\gamma(\cdot, \cdot)$ is the incomplete gamma function. In other words, when employing holistic review with plus factors, having the same threshold $q_\dagger$ operating on $F_\dagger^{(r)}(\cdot)$ for URM applicants and $F^{(r)}(\cdot)$ for non-URM applicants, is equivalent to having a threshold $q_\dagger/\eta_\dagger$ for URM applicants and $q_\dagger$ for non-URM applicants but operating only on $F^{(r)}(\cdot)$, where $q_\dagger/\eta_\dagger > q_o > q_\dagger$.

Since $k^{(\text{poor})} = k^{(\text{rich})} = k_o$, the two PDF curves only have one intersecting point:

$$\begin{aligned} \frac{1}{\Gamma(k_o)(\theta^{(\text{poor})})^{k_o}} q^{k_o - 1} e^{-q/\theta^{(\text{poor})}} &= \frac{1}{\Gamma(k_o)(\theta^{(\text{rich})})^{k_o}} q^{k_o - 1} e^{-q/\theta^{(\text{rich})}} \\ \implies \quad q &= \frac{k_o \ln(\theta^{(\text{poor})}/\theta^{(\text{rich})})}{1/\theta^{(\text{rich})} - 1/\theta^{(\text{poor})}}. \end{aligned} \tag{B.12}$$

Then, when the selective admission availability is limited such that $q_o < \frac{k_o \ln(\theta^{(\text{poor})}/\theta^{(\text{rich})})}{1/\theta^{(\text{rich})} - 1/\theta^{(\text{poor})}}$, because of the CDF dominance of the rich region over the poor region (Assumption 4.3), and that we can equivalently compare thresholds $q_\dagger/\eta_\dagger > q_o > q_\dagger$ at the original-scale CDF $F^{(r)}(\cdot)$, we have:

$$\forall \eta_\dagger \in \left[\frac{q_o(1/\theta^{(\text{rich})} - 1/\theta^{(\text{poor})})}{k_o \ln(\theta^{(\text{poor})}/\theta^{(\text{rich})})}, 1\right), F^{(\text{rich})}\big(\frac{q_\dagger}{\eta_\dagger}\big) - F^{(\text{rich})}(q_o) > F^{(\text{poor})}\big(\frac{q_\dagger}{\eta_\dagger}\big) - F^{(\text{poor})}(q_o).$$

□

B.3 Proof of Theorem 4.7 in Section 4.4

Theorem (Top-Percentage Plans Reallocate Spots from Rich Region to Poor Region)**.** *Under Assumptions 4.1–4.4, let us denote with q_o the default threshold for selective admission, and with $q^{(\text{poor})}$ and $q^{(\text{rich})}$ the thresholds for poor and rich regions, respectively, if top-percentage plans are employed. Then, the increase in selective admissions (in terms of counts) for applicants from the poor region, comes from spots reallocated out of the rich region. This redistribution is a result of the top-percentage plans, and is not relevant to applicants' demographic group:*

$$\big(n_a^{(\text{poor})} + n_{a'}^{(\text{poor})}\big)\big[F^{(\text{poor})}(q^{(\text{poor})}) - F^{(\text{poor})}(q^{(o)})\big] = \big(n_a^{(\text{rich})} + n_{a'}^{(\text{rich})}\big)\big[F^{(\text{rich})}(q^{(o)}) - F^{(\text{rich})}(q^{(\text{rich})})\big].$$

Furthermore, if region-specific shape parameters satisfy $k^{(\text{poor})} = k^{(\text{rich})}$, we additionally have:

$$q^{(\text{poor})}/q^{(\text{rich})} = \theta^{(\text{poor})}/\theta^{(\text{rich})}.$$

Proof. Top-percentage plans distribute the limited availability of selective admissions in a way that guarantee admissions to top-percentage applicants in their regions, and the resulting thresholds are region-specific. Then, we can calculate the default threshold q_o and the region-specific thresholds when top-percentage plans are employed:

$$(n_a^{(\text{poor})} + n_{a'}^{(\text{poor})}) \cdot F^{(\text{poor})}(q_o) + (n_a^{(\text{rich})} + n_{a'}^{(\text{rich})}) \cdot F^{(\text{rich})}(q_o) = g, \tag{B.13}$$

$$
\begin{aligned}
&(n_a^{(\text{poor})} + n_{a'}^{(\text{poor})}) \cdot F^{(\text{poor})}(q^{(\text{poor})}) + (n_a^{(\text{rich})} + n_{a'}^{(\text{rich})}) \cdot F^{(\text{rich})}(q^{(\text{rich})}) = g, \\
&\text{where} \quad F^{(\text{poor})}(q^{(\text{poor})}) = F^{(\text{rich})}(q^{(\text{rich})}) = \frac{g}{n_a^{(\text{poor})} + n_a^{(\text{rich})} + n_{a'}^{(\text{poor})} + n_{a'}^{(\text{rich})}}.
\end{aligned} \tag{B.14}
$$

Compare Equations B.13 and B.14, we have:

$$
(n_a^{(\text{poor})} + n_{a'}^{(\text{poor})}) \big[F^{(\text{poor})}(q^{(\text{poor})}) - F^{(\text{poor})}(q^{(o)})\big] = (n_a^{(\text{rich})} + n_{a'}^{(\text{rich})}) \big[F^{(\text{rich})}(q^{(o)}) - F^{(\text{rich})}(q^{(\text{rich})})\big].
$$

Because of the numerical property of Gamma CDF's (as we have seen in the proof for Theorem 4.6), when region-specific shape parameters satisfy $k^{(\text{poor})} = k^{(\text{rich})} = k$, we have:

$$
\begin{aligned}
F^{(\text{poor})}(q^{(\text{poor})}) &= \frac{1}{\Gamma(k)} \gamma\big(k, \frac{q^{(\text{poor})}}{\theta^{(\text{poor})}}\big), \\
F^{(\text{rich})}(q^{(\text{rich})}) &= \frac{1}{\Gamma(k)} \gamma\big(k, \frac{q^{(\text{rich})}}{\theta^{(\text{rich})}}\big),
\end{aligned}
$$

together with Equation (B.14), and we have:

$$
F^{(\text{poor})}(q^{(\text{poor})}) = F^{(\text{rich})}(q^{(\text{rich})}) \implies \frac{q^{(\text{poor})}}{\theta^{(\text{poor})}} = \frac{q^{(\text{rich})}}{\theta^{(\text{rich})}}, \text{ i.e., } \frac{q^{(\text{poor})}}{q^{(\text{rich})}} = \frac{\theta^{(\text{poor})}}{\theta^{(\text{rich})}}.
$$

□

C Additional Results and Discussions on Empirical Analyses

In this section, we present additional results and discussions on empirical experiments. In Section C.1, we provide experimental details on University of California undergrad admission data, as well as further discussions of the empirical results. Then in Section C.2, we present additional empirical analyses based on the US Census data.

C.1 Empirical Analyses on University of California Undergrad Admission

We provide description of the data, clarification of the Gamma parameterization for score distribution, and further discussions on the empirical results presented in Section 5.

Reviewer h9Xa: **C1**

We include the data description for UC summary statistics, and further discussions of the empirical results.

C.1.1 Description of the Data

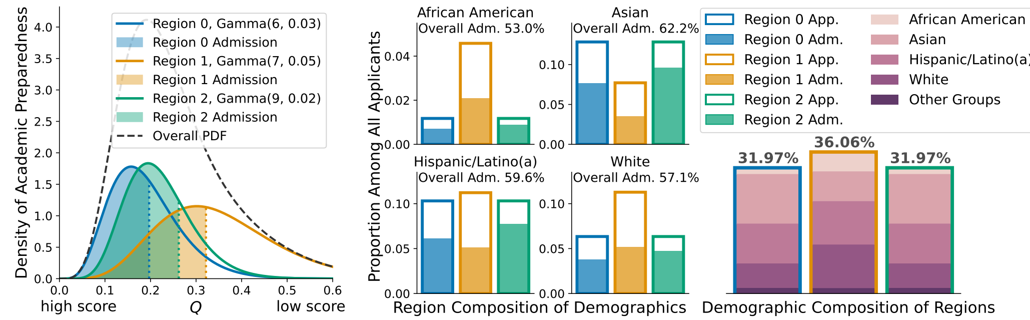
The University of California (UC) system is a public university system in the US. The UC Information Center provide summary statistics of undergrad admissions each year, including the undergraduate admissions summary, and the freshmen fall admissions summary. Because of legal and ethical considerations, the detailed data points at the individual level are not publicly available.

In the empirical analyses presented in Section 5, we utilize the university-wide (i.e., across the UC system) summary statistics of undergraduate admissions. Specifically, among the data for applicants (those who applied to at least one colleges in UC system), admissions (those who got offers from at least one college in UC), and enrollments (those who accepted the offers and enrolled in a specific college in UC), we utilize the application and admission statistics.

The undergraduate admissions summary provides the number of applicants and admitted students.[9] For a specific year and campus, the data takes a form of breakdown-counts across different demographic groups, including African American, American Indian, Asian, Hispanic/Latino(a), Pacific Islander, White, Unknown, International. The freshmen fall admissions summary provides the proportion of applicants and admitted students whose characteristics satisfy certain conditions.[10] For instance, the quantile statistics for high school weighted cumulative grade point average can be retrieved with the "HS weighted, capped GPA" option. All summary statistics are de-duplicated to avoid multiple-counting of students who applied to or admitted by multiple colleges at UC.

[9] https://www.universityofcalifornia.edu/about-us/information-center/admissions-residency-and-ethnicity

[10] https://www.universityofcalifornia.edu/about-us/information-center/freshman-admissions-summary



(a) Score densities

(b) Race- and region- specific composition

Figure 4: Recapitulation of Figures 3(a) and 3(b) in appendix, enlarged for better readability. Panel (a): region-specific and overall densities of academic preparedness. Panel (b): for each group, the region-specific compositions of application and admission proportions (left four subplots); for each region, the demographic composition of applicants (right subplot).

C.1.2 Gamma Parameterization for Score Distribution

Reviewer g5W4: **Q6**

The Gamma parameterization is not arbitrary.

Previous literature in educational research found that the distribution of student scores is roughly bell-shaped but is often not perfectly Gaussian (see, e.g., Arthurs et al. 2019). The distribution tends to skew towards the low-score end, and the support is often bounded (e.g., falls in $[S_{\mathrm{MIN}}, S_{\mathrm{MAX}}]$). Therefore, we use Gamma distributions to parameterize the score distribution, and utilize the shape and scale parameters to model the skewness and long-tail behaviors of the score distribution. This is consistent to Assumption 4.3 utilized in our theoretical analyses.

C.1.3 Further Discussions on Empirical Results

We provide further discussions on empirical results, especially Figures 3(a) and 3(b), enlarged and recapitulated in Figure 4 for better readability. Here, the regions may not correspond to real geographical locations due to the the under-constrained nature of the optimization problem (Section 5.1), and we focus on the interpretation of the results in terms of the relation among characteristics of regions, demographic groups, and academic preparedness. In Section C.2, we will present data analyses based on the US Census data, where more detailed geographical information is available.

Reviewer h9Xa: **C3**

We provide further discussions on Figure 3, enlarged and recapitulated in Figure 4 for better readability.

Figure 4(a) presents the region-specific densities of academic preparedness of applicants, as well as the overall density if we consider all applicants. We consider the pool of applicants, instead of that of admitted or enrolled students, since the application data is not yet "selected" by the university through the admission decision-making process, and therefore, more closely represents the underlying distribution of academic preparedness. Since the mean of a variable that follows Gamma distribution $\Gamma(k, \theta)$ is $k \cdot \theta$, the average score is 3.34 ($6 * 0.03 = 0.18$ converted back to the original scale) for Region 0 (blue), 2.82 for Region 1 (orange), and 3.34 for Region 2 (green). On average, the applicants in Region 0 and Region 2 have higher scores compared to those in Region 1, indicating the relative lack of educational resource in Region 1 (which results in overall insufficient academic preparedness). While the mean score is roughly the same for Region 0 and Region 2, the density of Region 0 is more concentrated at the high-score end compared to Region 2. From the resulting thresholds for the selective admissions, we can see that the threshold for Region 0 is more competitive than that for Region 2, which is further more competitive than that for Region 1.

In order to see the race-specific compositions of admissions indicated by the color-shaded areas under region-specific curves in Figure 4(a), we present Figure 4(b). We use the height of color-coded bars to denote the proportion of applicants that reside in specific regions, and the color-shaded part to indicate the proportion of admissions. For instance, for the African American group, the majority of applicants are from Region 1 (since the orange bar is highest in the upper-left subplot of Figure 4(b), corresponding to Region 1). Although more applications come from Region 1 (36.06% among all applicants), Region 1 appears to be the area where the educational resource is most scarce, and the relative concentration of African American applicants is more pronounced compared to other groups. The fact that the overall admission rate (53%) is lowest for the African American group also

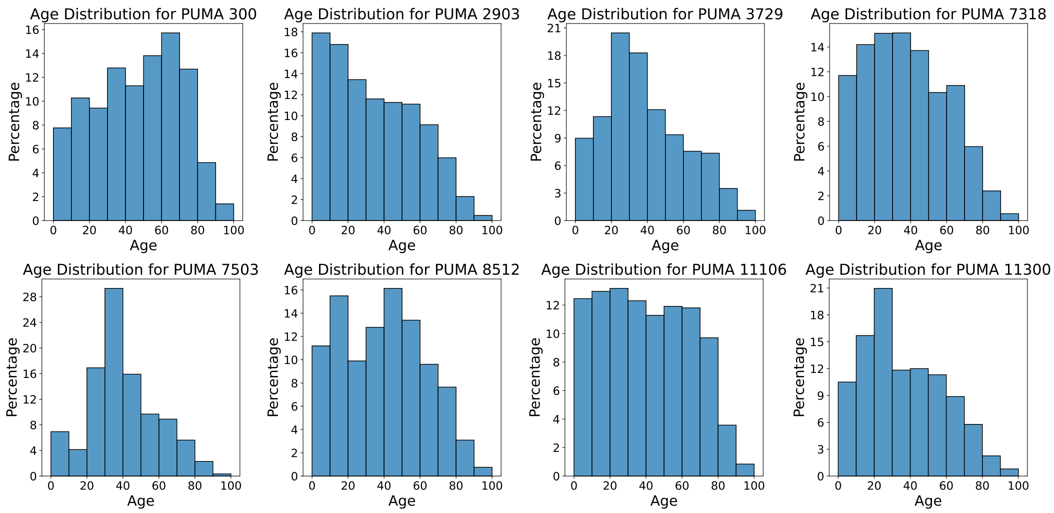


Figure 5: Age distribution in different PUMA regions in California based on US Census data.

corroborates with the previous observation. In other words, there is a correlation between region's ethnicity composition and the state-of-affairs of *social determinants of opportunity*, as indicated by the academic preparedness of applicants and the admission outcomes.

C.2 Additional Analyses on US Census Data

In this section, we present additional analyses on the US Census data (Census Bureau, 2009; 2014; 2022) to further illustrate the importance of considering the *social determinants of opportunity* in procedural fairness. We retrieve the public use microdata sample (PUMS) data from the US Census Bureau (Census Bureau, 2021), and provide visualizations of the age structure, racial composition, and occupation distribution in different Public Use Microdata Areas (PUMAs) in California based on the 2021 US Census PUMS data. PUMA is a geographical region smaller than counties, and the PUMA region is a strict subset of the corresponding state. Each PUMA contains at least 100,000 residents and provides reliable, detailed demographic, economic, and housing statistics at a sub-state level while also protecting the confidentiality of respondents (Census Bureau, 2021).

Reviewer h9Xa: **C2**
Following your suggestion, we have provided additional analyses on the US Census data.

Reviewer 8F2j: **Q2**
The region information is available in the US Census data.

C.2.1 Age Structure of Population in PUMAs

In Figure 5, we present age distributions in different PUMAs. For instance, PUMAs 3729, 7503, 11300 show noticeable concentrations of younger individuals, particularly in the 20–40 age range, suggesting a potentially more dynamic, working-age population which may affect local labor markets and educational demands. In contrast, PUMAs 7318 and 11106 exhibit a more balanced distribution across age groups, but with a slight skew towards middle-aged populations, which could indicate stable, established communities possibly with higher home ownership and lower school enrollment rates. For PUMA 8512, there are peaks in the 20s and again in the 50s, represent a mix of young adults possibly associated with entry-level professional work, and also senior adults in established careers or nearing retirement. The age distribution for PUMA 300 shows a peak around the age of 70s, reflecting a demographic profile with a substantial proportion of senior adults. Each area's age distribution can profoundly impact local policies, economic conditions, and community services tailored to the dominant age groups' needs. Therefore, the residents will be positioned differently in terms of *social determinants of opportunity* such as educational resources, employment opportunities, and healthcare providers.

C.2.2 Racial Composition in PUMAs

In Figure 6, we present racial compositions across PUMAs. In the context of US Census data, "Hispanic or Latino(a)" origin is considered an ethnicity, not a race. Individuals of Hispanic or Latino(a) origin can be of any race and are often asked to identify both their race and their ethnicity

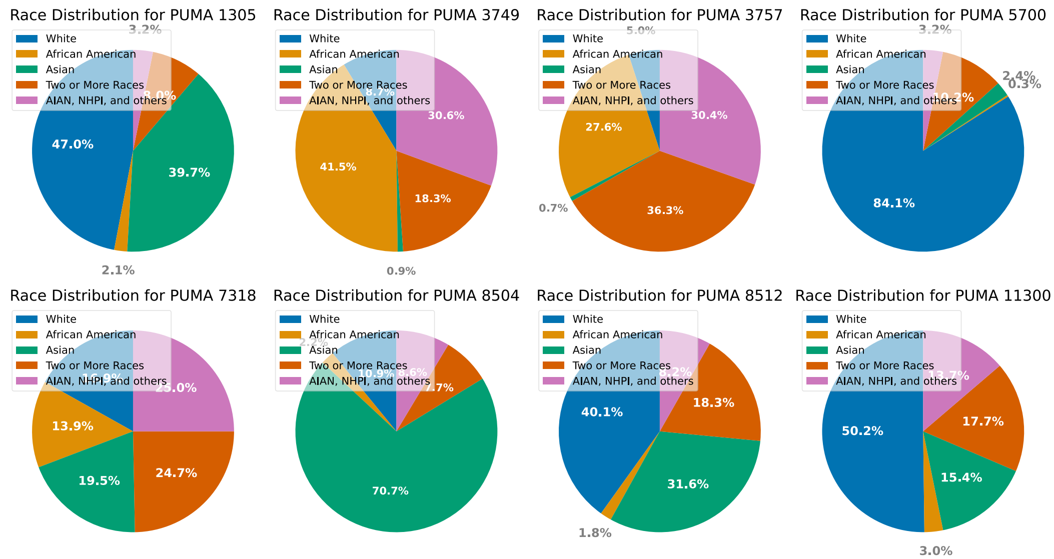


Figure 6: Racial composition in various PUMA regions in California based on US Census data.

during the data collection. Therefore, the racial composition does not contain a separate category for Hispanic or Latino(a) individuals.

As we can see, for historical and cultural reasons, the racial compositions vary quite a bit across different regions. For instance, PUMA 5700 predominantly consists of White individuals, making up 84.1% of its population, indicating a less racially diverse area compared to others. Similarly, PUMA 8504 displays a vast majority of Asian residents, accounting for 70.7% of the population. In contrast, PUMA 7318 offers a more balanced racial mix with no single group exceeding more than 30%, suggesting a more racially integrated community. These variations in racial composition can impact community needs, including educational services, cultural programs, and language services, and may influence local policy-making and resource allocation. Therefore, the association between *social determinants of opportunity* and racial composition of the population can differ significantly across regions.

C.2.3 Occupation Distribution in PUMAs

In Figure 7, we present distribution of occupations from certain categories in various PUMAs. The diverse workforce compositions reflect varying regional economic profiles and potential educational infrastructures. For instance, PUMAs 101 and 8503 display a strong presence of occupations related to science, engineering, education, and so on. In contrast, PUMA 6712 shows a more balanced distribution across different occupation categories (except for primary industries), suggesting a balanced mix of professional services and healthcare employment sectors. In terms of the category of farming, fishing, and forestry occupations, PUMAs 1901 and 8301 differ from other PUMAs (e.g., 101 and 8503). This category forms a significant part of the workforce (more than a third in both 1901 and 8301), reflecting an economy heavily reliant on primary industries. These patterns highlight how local natural and industrial resources, as well as economies, can significantly influence the occupational structures and, by extension, the training and education needed to support these sectors. Therefore, the *social determinants of opportunity* in different regions can be shaped differently.

C.2.4 Combination of Factors in PUMA

In Figure 8, we present how PUMAs can have very different profiles in terms of residents' age structure, race decomposition, and occupation distribution. In terms of the age structure, PUMAs 3749 and 8504 show more concentrations in the 20–40 age range, while PUMA 1700 has a high proportion of senior adults. In terms of the race decomposition, the majority of residents are white (75.9%) for PUMA 1700, African American (41.5%), and Asian (70.7%) for PUMA 8504. In terms of the occupation distribution, while the proportion of medical and healthcare practitioners is similar

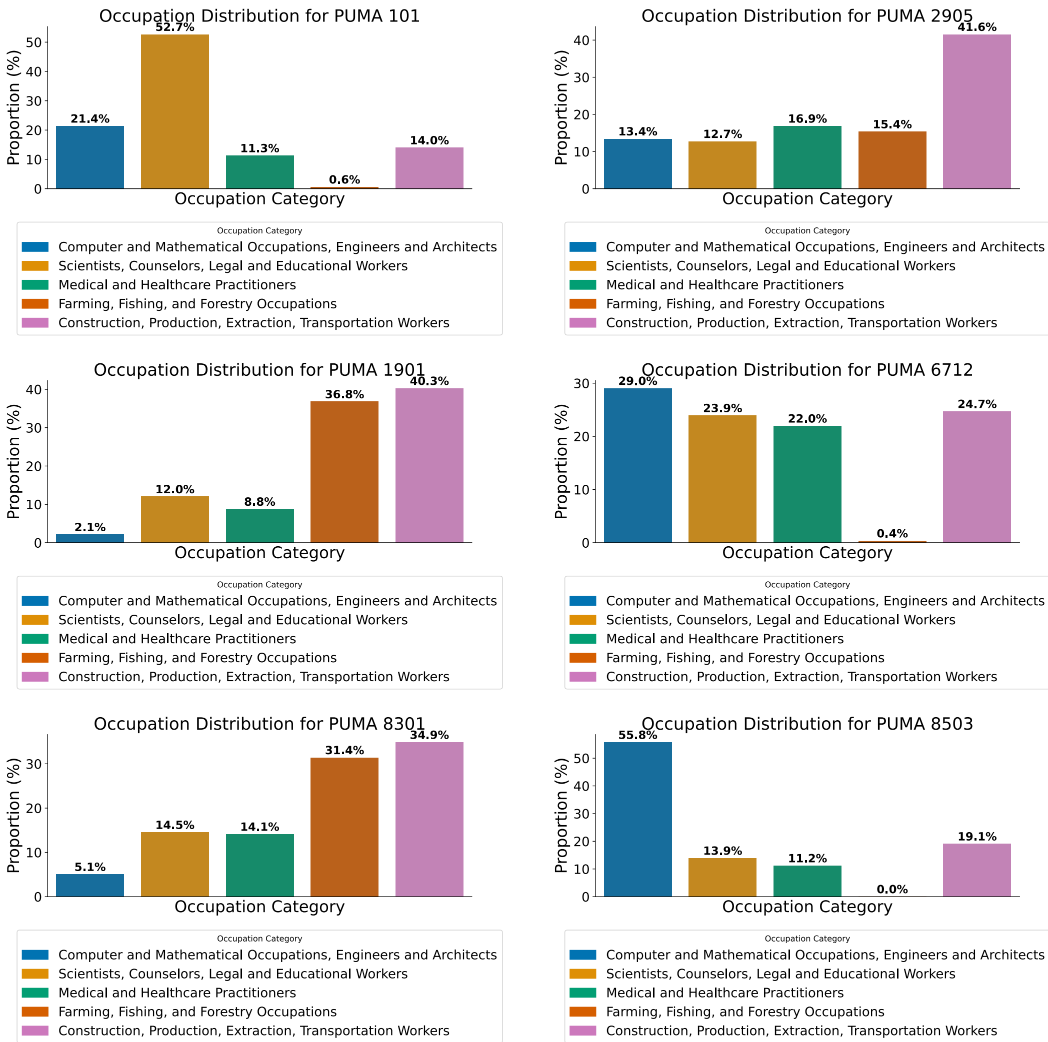
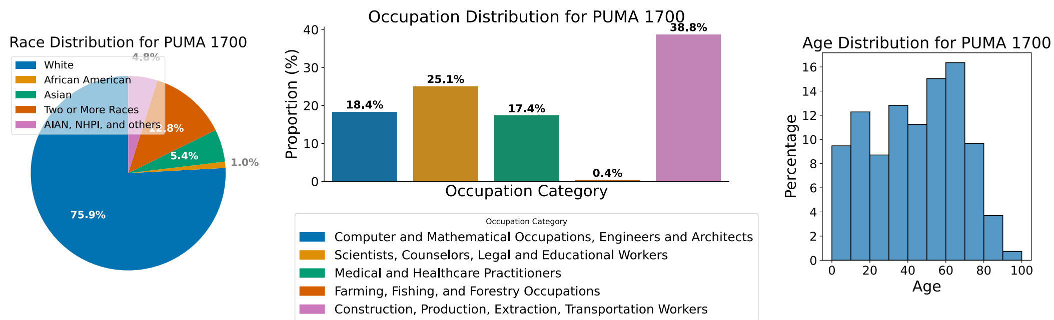
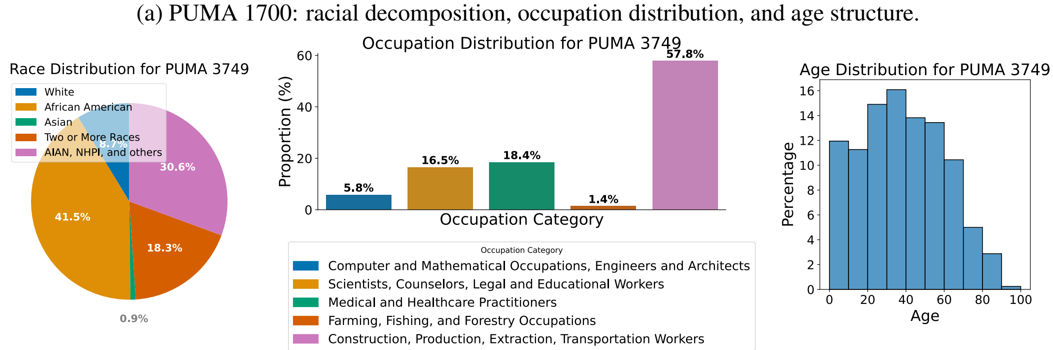


Figure 7: Occupational structure in various PUMA regions in California based on US Census data.

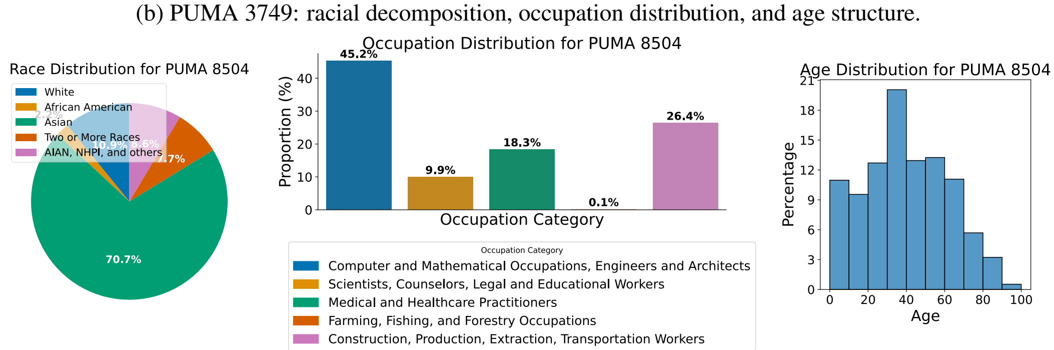
across the three regions, the occupational structures are very different. For instance, nearly one half of the working force in PUMA 8504 is within the category of computer and mathematical occupations, while the number is significantly lower in PUMAs 1700 and 3749, with a proportion of 18.4% and 5.8%, respectively. The comprehensive understanding of the *social determinants of opportunity* in different regions can help inform policy-making and resource allocation decisions, so that we can achieve procedural fairness in a more principled and transparent way.



(a) PUMA 1700: racial decomposition, occupation distribution, and age structure.



(b) PUMA 3749: racial decomposition, occupation distribution, and age structure.



(c) PUMA 8504: racial decomposition, occupation distribution, and age structure.

Figure 8: PUMAs with different profiles in terms of residents' age, race, and occupation.