Algorithmic greenlining: An approach to increase diversity*

Christian Borgs,[1] Jennifer Chayes,[1] Nika Haghtalab,[1] Adam Tauman Kalai,[1] and **Ellen Vitercik[2]**
[1]Microsoft Research, Cambridge, Massachusetts, USA
{christian.borgs, jchayes, adam.kalai, nika.haghtalab}@microsoft.com
[2]Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
vitercik@cs.cmu.edu

Abstract

In contexts such as college admissions, hiring, and image search, decision-makers often aspire to formulate selection criteria that yield both high-quality and diverse results. However, simultaneously optimizing for quality and diversity can be challenging, especially when the decision-maker does not know the true quality of any criterion and instead must rely on heuristics and intuition. We introduce an algorithmic framework that takes as input a user's selection criterion, which may yield high-quality but homogeneous results. Using an application-specific notion of substitutability, our algorithms suggest similar criteria with more diverse results, in the spirit of statistical or demographic parity. For instance, given the image search query "chairman", it suggests alternative queries which are similar but more gender-diverse, such as "chairperson". In the context of college admissions, we apply our algorithm to a dataset of students' applications and rediscover Texas's "top 10% rule": the input criterion is an ACT score cutoff, and the output is a class rank cutoff, automatically accepting the students in the top decile of their graduating class. Historically, this policy has been effective in admitting students who perform well in college and come from diverse backgrounds. We complement our empirical analysis with learning-theoretic guarantees for estimating the true diversity of any criterion based on historical data.

Introduction

In many application domains such as college admissions, hiring, and image search, domain experts aim to develop selection criteria that yield high-quality and diverse results. However, they often operate under an unavoidable lack of information: they do not know the true quality of any given criterion, and instead must rely upon heuristics and intuition. This makes it difficult to simultaneously optimize quality and diversity. For example, consider choosing university admissions criteria, a scenario where there is extremely limited information about the true quality of any candidate. An admissions officer's intuition might suggest that admitting students with an SAT score above 1400 would provide high-quality candidates, but the criterion might admit few students from minority groups. Nonetheless, there may be similar criteria, such as "SAT score above 1200 and class rank in the top 10%" which yield high-quality and diverse student bodies. Of course, a human decision-maker cannot search over all related criteria by hand in order to discover criteria that return diverse results. In this work, we present a framework for automating this search procedure.

Our algorithms take as input a user-specified criterion t, a pre-defined similarity measure among criteria, and a set of examples. They suggest alternate criteria $t'_1, \ldots, t'_k$ with greater estimated diversity on the set of examples but which are similar to t. For example, t might be the image search query "haircut" and $t'_1, \ldots, t'_k$ might include related searches such as "hairstyles" and "popular haircuts". We model a criterion t as a function $t : X \to R$, where X is a set of observables, such as job applications, and R is a set of results, perhaps indicating whether an applicant is hired. Our algorithms rely on application-specific similarity functions which measure how substitutable any two criteria are. They also depend on functions that measure the diversity of any criterion's results. Our algorithms optimize the similarity function while meeting a desired diversity constraint.

We apply our framework in three key areas: college admissions, job search, and image search. In each application, we suggest straight-forward, natural substitutability and diversity measures, and then exemplify how our techniques can be used to find similar criteria that yield more diverse results. For example, in the case of college admissions, given two criteria t and t' and the sets E_t and $E_{t'}$ of students they admit, we measure the similarity of t and t' as a simple function of the symmetric difference between E_t and $E_{t'}$. As another example which illustrates the generality of our approach, we say that two jobs are similar if the requisite skills, education, and training are transferable, as formalized by the Department of Labor. We emphasize that in all of our applications, a domain expert could alter these two measures as they see fit. Our goal is not to nail down a universally optimal definition of substitutablity and diversity, but to introduce this technique to domain experts who can determine appropriate metrics for their specific applications. We also

*This research uses public data from the Texas Higher Education Opportunity Project (THEOP) and acknowledges the following agencies that made THEOP data available through grants and support: Ford Foundation, The Andrew W. Mellon Foundation, The William and Flora Hewlett Foundation, The Spencer Foundation, National Science Foundation (NSF Grant SES-0350990), The National Institute of Child Health Human Development (NICHD Grant R24 H0047879) and The Office of Population Research at Princeton University.

provide provable guarantees for estimating the diversity of any criterion given historical data. Thus, our algorithms can use diversity estimates instead of true diversity scores.

We call our approach *algorithmic greenlining* as it is the antithesis of redlining, the historic and systematic denial of services to residents of specific communities, often due to demographics. Greenlining is the conscious effort to promote minority interests and representation by using unbiased criteria in day-to-day settings. In this work, we provide individuals with computational tools to help them reduce the intentional or unintentional bias encoded by their criteria.

Related work

Algorithmic fairness using similarity metrics. Our notion of criterion similarity is reminiscent of the similarity metric used over people for *individually fair classification* (Dwork et al., 2012). As defined by Dwork et al. (2012), under an individually fair classification system, similar individuals should be treated (i.e., classified) similarly. Thus, it relies on a similarity metric between individuals, often based on high-dimensional attribute vectors. In contrast, we rely on a notion of similarity between criteria.

Algorithmic fairness with and without quality scores. We do not assume the user knows the true quality of any given criterion or outcome. Thus, we cannot measure criterion fairness using notions such as equalized odds (Hardt, Price, and Srebro, 2016) which depend on ground truth. Rather, we can use any notion that only depends on the outcomes of the criteria (e.g., the fraction of students accepted under a given university admissions criterion from a minority group), such as statistical parity (Dwork et al., 2012).

Image search. Several works analyze the presence of bias in image results from major search engines (Kay, Matuszek, and Munson, 2015; Otterbacher, Bates, and Clough, 2017). Others study how to achieve diversity in image search outside the context of algorithmic fairness. Given a query, these works provide algorithms that return many different kinds of images matching that query (e.g., (Kennedy and Naaman, 2008; Wang et al., 2010)). For example, an image search for "apple" should not only return images of the fruit, but also of the computer. Some works also study how to ensure the image search results exhibit racial and gender diversity (Celis et al., 2018). In contrast, our goal is to provide search suggestions whose results are more diverse, rather than altering the images returned by a search.

Fair ranking. In a vein similar to image search, many researchers have studied how to rank a set of items fairly (Zehlike et al., 2017; Yang and Stoyanovich, 2017; Karako and Manggala, 2018; Biega, Gummadi, and Weikum, 2018; Singh and Joachims, 2018; Celis, Straszak, and Vishnoi, 2018) and how to measure the fairness of an existing ranking (Yang and Stoyanovich, 2017; Yang et al., 2018). Unlike our work, these works assume the algorithm has access to ground-truth quality scores.

Query suggestions. In addition to ranking, Information Retrieval systems often provide suggestions of alternative relevant searches (e.g., (Jones et al., 2006)). The goal has generally been to help the user better meet their information seeking needs. However, to the best of our knowledge, no prior work in this domain has considered fairness or diversity.

We include additional related work in the appendix.

The general model and notation

In our model, there is an underlying set $X \times Z$ of examples (x, z) where $x \in X$ represents the observable attributes of an instance and $z \in Z$ represents its protected attributes. There is a set T of criteria. Each criterion $t \in T$ is a function mapping the set X of observable attributes to a set R of results. For example, in the case of college admissions, $R = \{0, 1\}$ and $t(x) = 1$ if and only if the student with observable attributes x is admitted under criterion t.

We want to measure the diversity of a criterion given an arbitrary multi-set $E \subseteq X \times Z$ of examples. In the admissions example, E could be the set of students who apply to some university on a given year. When the results are binary ($R = \{0, 1\}$) given a multi-set of examples E, we use the notation $E_t = \{(x, z) \mid (x, z) \in E, t(x) = 1\}$ to denote the subset of examples $(x, z) \in E$ such that $t(x) = 1$.

For every multi-set $E \subseteq X \times Z$ of examples and every criterion $t \in T$, there is a diversity score $\texttt{div}(t, E) \in [0, 1]$. For example, t might be a threshold indicating that all students with an SAT score above 600 should be admitted. Given the set E of students, $\texttt{div}(t, E)$ might measure the fraction of admitted students who are Hispanic.

We assume we have a substitutability function σ where $\sigma(t, t', E) \in [0, 1]$ indicates how good a substitute t' is for t given the multi-set of examples E, not accounting for fairness or diversity. This captures how willing the agent would be to replace their criterion t with the criterion t'. For instance, someone searching for "doctor" might be happy to instead search for "physician". If each criterion is a range of admissible standardized test scores, a small change in thresholds might lead to a much more diverse set of admits. We assume that $\sigma(t, t, E) = 1$ for all $t \in T$ and all multi-sets $E \subseteq X \times Z$. We occasionally refer to the distance $\texttt{dist}(t, t', E)$ between two criteria (using the word "distance" loosely: we do not require that $\texttt{dist}$ be symmetric or that it satisfy the triangle inequality) and we define $\texttt{dist}(t, t', E) = 1 - \sigma(t, t', E)$.

This definition of substitutability is extremely versatile: it can measure both intrinsic and extrinsic similarity. As an example of intrinsic similarity, we say at a high level that two jobs are similar if the requisite skills, education, and training are transferable, as formalized by the Department of Labor. In this case, $\sigma(t, t', E)$ is independet of E. Meanwhile, as an example of extrinsic similarity, we say that two university admissions criteria are similar if moving from one criterion to the other does not drastically change the set of admitted students. Thus, similarity is not intrinsic to the criteria themselves, but is a function of how they interact with the dataset at hand, i.e., $\sigma(t, t', E)$ depends on E as well as t and t'. Again, we emphasize that the similarity metrics we employ are only examples. A domain expert could employ any definition they see fit.

Given an initial criterion t, our goal is to suggest a criterion t' that is largely substitutable ($\sigma(t, t', E)$ is close to 1) and is diverse ($\texttt{div}(t, E)$ is within a user-specified range).

Notation. We denote the i^{th} entry of a vector $\boldsymbol{v}$ as $v[i]$.

General model instantiations

We now instantiate our model in several settings: university admissions, image search, and job applicant search. In the appendix, we discuss the ethics of our data usage.

University admissions

We begin with a problem motivated by university admissions. We represent a set of applicants by feature vectors in $[0, H)^d$, for some $H \in \mathbb{R}$. We use data collected from the University of Texas at Austin by the Texas Higher Education Opportunity Project (THEOP). The features we focus on are the students' composite SAT scores, composite ACT scores, and high school class rank (higher is better). We select our minority group to consist of applicants who are identified as "Black, Non-Hispanic" or "Hispanic" ($z = 1$) and our majority group to consist of "White, Non-Hispanic" applicants ($z = 0$). Therefore, $X = [0, H)^d$ and $Z = \{0, 1\}$.

In our first experiment, the set T of criteria consists of all axis-aligned hyperplanes intersecting $[0, H]^d$. Specifically, each criterion $t \in T$ is defined by a dimension $i \in [d]$ and a threshold $b \in [0, H]$. The multi-set $E_t \subseteq E$ consists of those tuples $(\boldsymbol{x}, z)$ with $x[i] \geq b$. For any criterion t such that $E_t \neq \emptyset$, we define $f_m(t, E)$ to be the fraction of individuals accepted by t who are minorities, i.e., $f_m(t, E) = \frac{1}{|E_t|} \sum_{(\boldsymbol{x},z) \in E_t} z$. We define $\texttt{div}$ to measure how far this fraction is from a user-specified diversity level c, i.e., $\texttt{div}(t, E) = 1 - |c - f_m(t, E)|$.

We define the substitutability of two criteria t and t' based on the overlap of the students accepted under the two rules. Using $E_t \Delta E_{t'} = (E_t \cup E_{t'}) \backslash (E_t \cap E_{t'})$ to denote the symmetric difference between E_t and $E_{t'}$, we define

$$\sigma(t, t', E) = \begin{cases} 1 - \frac{|E_t \Delta E_{t'}|}{m} & \text{if } |E_t| \leq |E_{t'}| \leq 2\,|E_t| \\ 0 & \text{otherwise.} \end{cases}$$

By enforcing $|E_t| \leq |E_{t'}| \leq 2\,|E_t|$, the number of applicants admitted by the two criteria is not drastically different.

The multi-set E of feature vectors describes the first 50,000 applicants from the dataset who a) are identified as "White, Non-Hispanic," "Black, Non-Hispanic," or "Hispanic," and b) took the ACT. Of these, 24% are from the minority group. Based on this, we set the fraction c of minority students in the definition of the function $\texttt{div}$ to be $c = 0.24$. Our input criterion t accepts all students whose cumulative ACT score is above 56 (the dotted line in Figure 1b), which includes 11,484 majority students and 1,298 minority students, so $\texttt{div}(t, E) = 0.87$.

Our goal is to find a criterion t' such that $\texttt{div}(t', E) \geq 0.95$ (i.e., the fraction of admitted minority students is in $[0.19, 0.29]$) and $\sigma(t, t', E)$ is as close to 1 as possible. (The threshold 0.95 could be replaced by any other bound.) This is summarized by the following optimization problem:

$$\text{maximize } \sigma(t, t', E) \text{ subject to } \texttt{div}(t', E) \geq 0.95. \quad (1)$$

The new criterion's output will be different from the original criterion's output since it will be more diverse, but the change will be as small possible. Under the optimal solution to Equation (1), students from the top 10% of their class are admitted. This criterion t' is illustrated by the solid line in Figure 1c. It admits 14,582 majority and 5,576 minority students, so $\texttt{div}(t', E) = 0.96$. Moreover, $\sigma(t, t', E) = 0.69$.

We thus rediscover the "top 10% rule" passed in 1997 by the Texas legislature (H.B.588), which guarantees any Texas student in the top 10% of their graduating class admission into any of Texas's public post-secondary institutions.

In the appendix, we describe our experiments in a more general setting where each criterion is represented by the intersection of multiple axis-aligned half-spaces. For example, a criterion might require that an admitted applicant has an ACT score above 50 and is in the top 40% of her class. This generalization allows us to preserve a higher degree of similarity while still increasing diversity. See Figure 2 for a visualization of our experimental results.

Image search

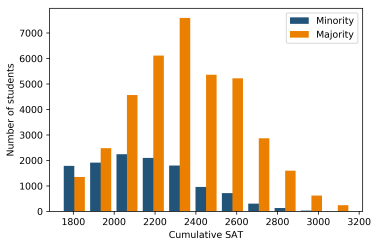
In image search, our goal is to provide users with similar image search terms that yield more diverse results. For example, if a user begins with the Google search "haircut", the resulting images are often surprisingly male-dominated. However, we find that there are often similar queries, such as "popular haircuts", which return more images of women.

Our model applies to any search engine, but we concentrate on Google and Bing in our experiments. We define the set X of observable attributes to be the set of all images indexed by Google and $Z = \{0, 1\}$ to be the set of sensitive attributes indicating gender. In particular, given an example (x, z), $z = 0$ if and only if x is an image of a man. We define the set of input examples E to be the entire set $X \times Z$. Given a search criterion t such as "haircut", we define $t(x)$ to be 1 if the image x is displayed in the first fifty images returned by a Google search for t and 0 otherwise, filtering Google's results so that they only return images of faces.

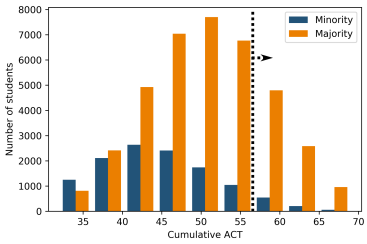
We adopt a simple diversity metric: given a search term t and the top fifty images returned by this search, how far from 50-50 is the gender ratio? In other words, $\texttt{div}(t, E) = 1 - \left| \frac{1}{2} - \frac{1}{50} \sum_{(x,z) \in E_t} z \right|$. Given an even split, $\frac{1}{50} \sum_{(x,z) \in E_t} z = \frac{1}{2}$, so $\texttt{div}(t, E) = 1$, and the closer $\frac{1}{50} \sum_{(x,z) \in E_t} z$ is to 0 or 1, the smaller $\texttt{div}(t, E)$ is.

We base our similarity metric between search terms on Bing's lists of related searches. (We define diversity based on Google and similarity based on Bing so that our results are not concentrated on one search engine.) Whether the search engine suggests a given term depends on the similarity between that term and the original term. We assume that the higher in the list a term is, the more similar it is to the original search. Thus, we define the distance $\texttt{dist}(t, t', E)$ (where $\texttt{dist}(t, t', E) = 1 - \sigma(t, t', E)$) to be the position of t' in the list of related searches given the original search term t, divided by the list's length, or 1 if t' is not in the list.

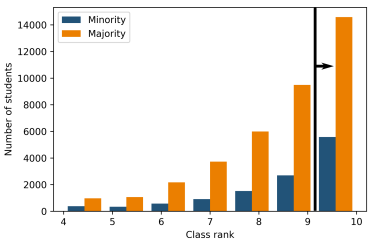
In Figure 3, we display plots for two different searches. For the sake of readability, we only plot a subset of the re-



(a) Cumulative SAT scores.

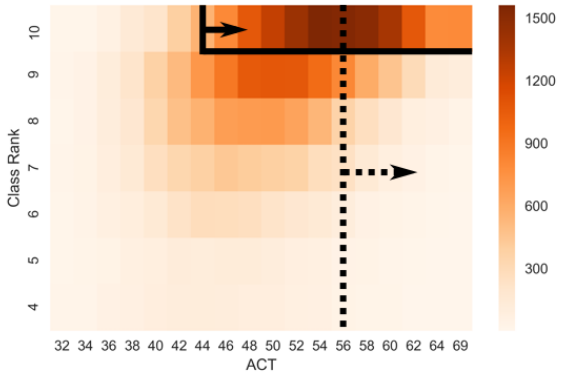


(b) Cumulative ACT scores with original criterion: ACT score above 56.

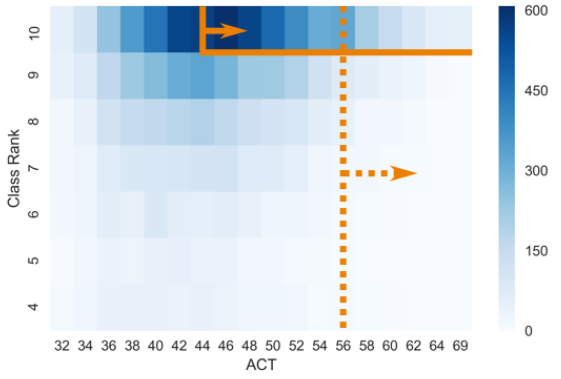


(c) Class rank with suggested criterion: top 10%.

Figure 1: University admissions when each criterion is defined by an axis-aligned half-space. The orange bars indicate the number of white, non-Hispanic students whose features fall within a given range. The blue bars correspond to the black and Hispanic students. The dotted line in Figure 1b represents the original criterion (cumulative ACT score above 56). The solid line in Figure 1c represents the criterion returned by the system (class rank in the top decile).



(a) Majority students' ACT scores versus class ranks.



(b) Minority students' ACT scores versus class ranks.

Figure 2: University admissions when each criterion is defined by the intersection of axis-aligned half-spaces. The plots display the majority (Figure 2a) and minority (Figure 2b) groups' ACT scores versus class ranks. The colorbars indicate the number of students who fall in each bin. The dotted lines represent the original criterion: ACT score above 56. The solid lines represent the suggested criterion: ACT score above 44 and class rank in the top decile.

lated searches. We can see that some searches have nearby searches with diverse search results. For example, "neckerchief" is similar to "scarf", but the images associated with "scarf" are almost all of women whereas the images associate with "neckerchief" have a more balanced gender ratio.

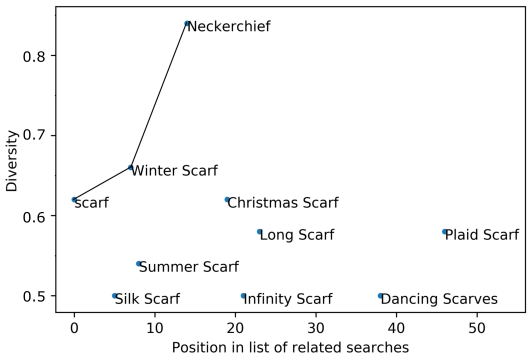
Job search

We next instantiate our general model for job applicant search. Given a recruiter's original query such as "computer programmer", the goal is to provide her with alternative search queries that return diverse sets of applicants.

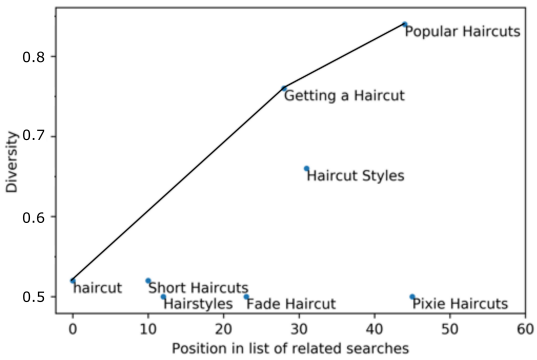
To instantiate our general model, we define the set X of observable attributes to be an abstract set of job applicants' observable attributes. For example, X could be a set of individuals' LinkedIn profiles. Let $Z = \{0, 1\}$ be the set of sensitive attributes which indicate gender. In particular, given an example (x, z), $z = 0$ if and only if x corresponds to a man. In this setting, we define the set of input examples E to be the entire set $X \times Z$. Each criterion $t \in T$ corresponds to an occupation. Given an individual (x, z), we define $t(x)$ to be 1 if the individual is applying for the job t and 0 otherwise.

To measure diversity, we use workforce statistics from the U.S. Census Bureau (2016). Given an occupation t, these statistics tell us the fraction ω_t of full-time workers in the U.S. with that job year-round who are women. We define diversity to measure how far from $\frac{1}{2}$ that fraction is, i.e., $\texttt{div}(t, E) = 1 - \left|\frac{1}{2} - \omega_t\right|$. If there is a 50-50 gender split, then $\omega_t = \frac{1}{2}$ and $\texttt{div}(t, E) = 1$. Meanwhile, the closer ω_t is to 0 or 1, the smaller $\texttt{div}(t, E)$ is.

In order to characterize similarity among jobs, we use the Department of Labor's Standard Occupational Classification (SOC) System, which organizes hundreds of occupations into a single hierarchy. Occupations are grouped together if they have similar duties or require similar skills, education, or training (U.S. Office of Management and Budget, 2018). For a small subset of the hierarchy, see Figure 5 in the appendix. We define the similarity of two jobs based on the length of the path between them in the hierarchy. For two jobs t and t', let $\ell(t, t')$ be the length of this path. The maximum path length between any two leafs in this hierarchy is

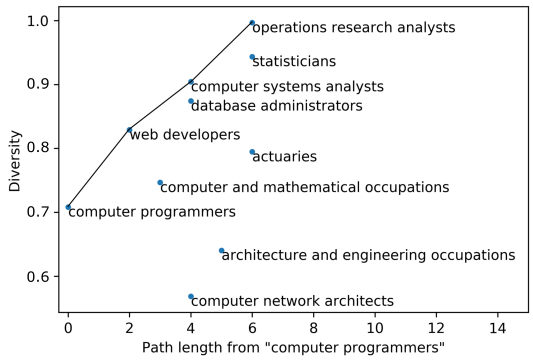


(a) Original search: "scarf"

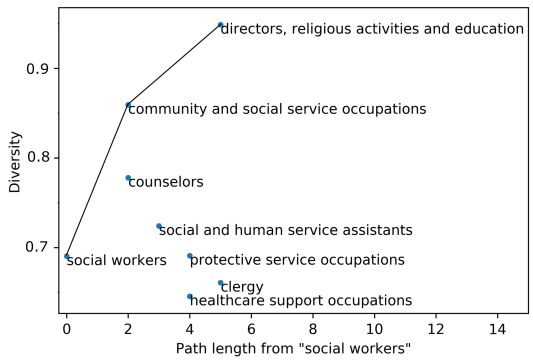


(b) Original search: "haircut"

Figure 3: Plots for image search. The position on the x-axis of a image search t' equals its position in Bing's list of search terms related to the original search t. Its position on the y-axis equals the diversity $\texttt{div}(t, t', E)$. The line illustrates the Pareto frontier of this bi-criteria optimization problem.



(a) Original search: "computer programmers"



(b) Original search: "social workers"

Figure 4: Plots for job search. The position on the x-axis of a job t' equals the path length between it and the original search t' in the SOC hierarchy. Its position on the y-axis equals the diversity $\texttt{div}(t, t', E)$. The line illustrates the Pareto frontier of this bi-criteria optimization problem.

8. Thus, we define the substitutability of two jobs t and t' as $\sigma(t, t', E) = 1 - \frac{\ell(t,t')}{8}$ so that it is in $[0, 1]$.

In Figure 4, we display plots for two occupations. We can see that some occupations have nearby occupations with more diverse workforces. For example, "web developer" is similar to "computer programmer", but there are more female web developers than computer programmers.

Diversity estimation guarantees

In this section, we provide provable guarantees for finding similar criteria with more diverse results. Here, our goal is to estimate the diversity of any given criterion using historical data. Calculating reliable diversity estimates is crucial because in some settings, it may be preferable or even mandatory to avoid using individuals' protected attributes to measure diversity. For example, an admissions officer may not know the ethnicity of any college applicant from the current year. However, they may have access to historical data, such as the University of Texas data we use to run our experiments, which they can use to form these estimates. Furthermore, given estimates that reflect the diversity of any criterion over society as a whole, a user can introduce policies, such as a fixed college admissions criterion, that can stay in place for years at a time. For example, they can be confident that admitting students based on class rank will consistently yield more diverse student bodies than admitting students based on standardized test scores. Finally, using diversity estimates rather than true diversity scores in order to choose among criteria means that no individual will be favored based on his or her sensitive attributes.

Throughout this section, we consider a setting where a criterion t is a map $t : X \to \{0, 1\}$. Given an example (x, z), $z \in [0, 1]$ is the degree to which x belongs to the minority group. Given a multi-set $E = \{(x_1, z_1), \ldots, (x_N, z_N)\}$ of examples, we define $f_m(t, E)$ to be the degree to which an average instance $(x, z) \in E_t$ belongs to the minority group. In other words, $f_m(t, E) = \frac{1}{|E_t|} \sum_{(x,z) \in E_t} z$. Recall that in the university admissions setting, we define $\texttt{div}(t, E)$ to measure how far $f_m(t, E)$ is from a user-specified $c \in [0, 1]$, i.e., $\texttt{div}(t, E) = 1 - |c - f_m(t, E)|$.

To justify this choice of $\texttt{div}(t, E)$, we prove that $f_m(t, E)$ estimates the protected attribute of an example $(x, z) \sim \mu$ conditioned on $t(x) = 1$. In other words, $f_m(t, E)$ approximates $\mathbb{E}_{(x,z)\sim\mu}[z \mid t(x) = 1]$ for all $t \in T$ so long as E has a sufficiently large number of candidates. That is, we show that with probability at least $1 - \delta$ over the draw of $E = \{(x_1, z_1), \ldots, (x_N, z_N)\} \sim \mu^N$, for all $t \in T$ such that $\Pr_{(x,z)}[t(x) = 1] \geq c$, we have that $\left|f_m(t, E) - \mathbb{E}_{(x,z)\sim\mu}[z \mid t(x) = 1]\right| \leq \epsilon$ when N is sufficiently large with respect to ϵ, δ, and c.

We use the learning-theoretic notion of VC dimension (Vapnik and Chervonenkis, 1971) to provide this sample complexity guarantee. VC dimension measures the *intrinsic complexity* of binary-valued function classes, or in other words, classes of functions that map to $\{0, 1\}$. Given a set $\mathcal{G}$ of functions which map an abstract domain $\mathcal{A}$ to $\{0, 1\}$ and a distribution μ' over $\mathcal{A}$, bounding the VC dimension of $\mathcal{G}$ allows us to bound the number of samples $a_1, \ldots, a_N \sim \mu'$ sufficient to ensure that the difference between the average value $\frac{1}{N}\sum_{i=1}^{N} g(a_i)$ of any function $g \in \mathcal{G}$ over the samples and its expected value $\mathbb{E}_{a\sim\mu'}[g(a)]$ is small. VC dimension applies specifically to binary-valued functions, so it does not immediately apply to our setting because the protected attributes z are real-valued and we are concerned with the conditional expectation $\mathbb{E}[z \mid t(x) = 1]$, not the expectation $\mathbb{E}[z]$. Nonetheless, we show how to use the VC dimension of T to bound the number of samples sufficient to ensure the difference between $f_m(t, E)$ and $\mathbb{E}[z \mid t(x) = 1]$ is small, for any criterion $t \in T$. In many applications, bounding the VC dimension of T is simple, so applying this sample complexity bound is straight-forward.

Below, we define VC dimension in terms of an abstract set of functions $\mathcal{G}$ which map a domain $\mathcal{A}$ to $\{0, 1\}$.

Definition 1 (VC dimension)**.** We say that $\mathcal{G}$ shatters the set $\mathcal{S} = \{a_1, \ldots, a_M\} \subseteq \mathcal{A}$ if for all binary vectors $\boldsymbol{b} \in \{0, 1\}^M$, there is a function $g_{\boldsymbol{b}} \in \mathcal{G}$ such that $g_{\boldsymbol{b}}(a_i) = b[i]$ for all $i \in [M]$. The VC dimension of $\mathcal{G}$, denoted $\text{VCdim}(\mathcal{G})$, is the size of the largest set that $\mathcal{G}$ shatters.

For example, recall the college admissions criteria we study: axis-aligned half-spaces and the intersection of axis-aligned half-spaces in $\mathbb{R}^d$. For both, $\text{VCdim}(T) \in O(d)$.

We now present our sample complexity bound. The full proof is in the appendix.

Theorem 1. *For any $\epsilon \in (0, 1/2)$ and $\delta \in (0, 1)$, with probability $1-\delta$ over the draw $E = \{(x_1, z_1), \ldots, (x_N, z_N)\} \sim \mu^N$, for all $t \in T$ such that $\Pr_{(x,z)}[t(x) = 1] \geq c$, $\left|f_m(t, E) - \mathbb{E}_{(x,z)\sim\mu}[z \mid t(x) = 1]\right| \leq \epsilon$ when $N \geq N_0 \in O\left(\frac{1}{c\epsilon^2}\left(\text{VCdim}(T)\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$.*

Estimation with correlated examples

Theorem 1 applies when the examples (x, z) are identically and independently distributed. In this section, we provide diversity estimation guarantees even when the examples are correlated. Rather than the i.i.d. assumption, we assume there is a distribution $\bar{\mu}$ over datasets E of a fixed size m, so the support of $\bar{\mu}$ is a subset of $(X \times Z)^m$. We show how to use a set $\mathcal{S} = \left\{E^{(1)}, \ldots, E^{(N)}\right\} \sim \bar{\mu}^N$ to estimate $\mathbb{E}_{E\sim\bar{\mu}}[f_m(t, E)]$ for any criterion $t \in T$. One way to interpret this result is in the setting where we wish to design admissions criteria for a network of universities, such as all U.S. public community colleges. If we have a criterion that admits diverse students over a random set of N colleges, we can be confident the criterion also admits diverse students on average nationwide. Colleges may have correlated applicants depending on geography, and this section's results allow for this correlation. Specifically, we prove the following guarantee. The full proof is in the appendix.

Theorem 2. *For any $\epsilon, \delta \in (0, 1)$, $N \geq N_0 \in \Theta\left(\frac{1}{\epsilon^2}\left(\text{VCdim}(T)\log\left(m\text{VCdim}(T)\right) + \log\frac{1}{\delta}\right)\right)$ samples are sufficient to ensure that with probability $1 - \delta$ over the draw of $E^{(1)}, \ldots, E^{(N)} \sim \bar{\mu}$, for all $t \in T$, $\left|\frac{1}{N}\sum_{i=1}^{N} f_m\left(t, E^{(i)}\right) - \mathbb{E}_{E\sim\bar{\mu}}[f_m(t, E)]\right| \leq \epsilon$.*

Limitations

We now describe several limitations of our work, some of which suggest directions for future research.

Scholars (Fryer Jr, Loury, and Yuret, 2007) have argued that *color-blind admissions criteria*, like those we study, are less effective at admitting students with strong academic performance compared to race-aware criteria. However, race-aware criteria may not be legal in many scenarios, in which case our approach presents a promising alternative.

A second limitation is that it may be computationally expensive to search for similar criteria yielding more diverse results. For example, this is the case in our university admissions application when each criterion is defined by the intersection of half-spaces. In fact, we prove that this problem is NP-complete in the appendix. Formulating an approximation algorithm is a promising direction for future work.

Lastly, if there is no clear notion of substitutability among criteria, then our framework will not apply. However, we believe our examples illustrate that there are often straightforward ways to define substitutability.

Conclusion

We study the problem of suggesting similar criteria that yield more diverse results. In our setting, a domain expert chooses some selection criterion, such as an admissions criterion to choose among a set of college applicants. We do not assume she knows the true quality of any given criterion, or that the notion of "true quality" is well-defined. Rather, she uses heuristics and intuition to select a criterion she believes will return high-quality results. At the same time, this expert would like to ensure the results are diverse, but optimizing for diversity and quality may be difficult without access to ground-truth quality metrics. We introduce automated techniques for suggesting similar criteria yielding more diverse results. We demonstrate its strong performance in three critical domains: college admissions, image search, and job search. Finally, we complement our experiments with theoretical guarantees, analyzing the amount of historical data required to estimate criteria diversity.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. *Proceedings of the International Conference on Machine Learning (ICML)*.

Anthony, M., and Bartlett, P. L. 2009. *Neural network learning: Theoretical foundations*. Cambridge University Press.

Biega, A. J.; Gummadi, K. P.; and Weikum, G. 2018. Equity of attention: Amortizing individual fairness in rankings. *Conference on Research and Development in Information Retrieval (SIGIR)*.

Bolukbasi, T.; Chang, K.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–186.

Celis, L. E.; Keswani, V.; Straszak, D.; Deshpande, A.; Kathuria, T.; and Vishnoi, N. K. 2018. Fair and diverse DPP-based data summarization. *arXiv preprint arXiv:1802.04023*.

Celis, L. E.; Straszak, D.; and Vishnoi, N. K. 2018. Ranking with Fairness Constraints. In *Proceedings of the International Colloquium on Automata, Languages and Programming (ICALP)*.

Chierichetti, F.; Kumar, R.; Lattanzi, S.; and Vassilvitskii, S. 2017. Fair clustering through fairlets. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.

Dudley, R. M. 1967. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis* 1(3):290 – 330.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the ACM Conference on Innovations in Theoretical Computer Science (ITCS)*.

Dwork, C.; Immorlica, N.; Kalai, A. T.; and Leiserson, M. D. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *Proceedings of the Conference on Fairness, Accountability and Transparency (FAT*)*.

Fryer Jr, R. G.; Loury, G. C.; and Yuret, T. 2007. An economic analysis of color-blind affirmative action. *The Journal of Law, Economics, & Organization* 24(2):319–355.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.

Jones, R.; Rey, B.; Madani, O.; and Greiner, W. 2006. Generating query substitutions. In *Proceedings of the International World Wide Web Conference (WWW)*.

Karako, C., and Manggala, P. 2018. Using image fairness representations in diversity-based re-ranking for recommendations. In *Proceedings of the Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML)*.

Kay, M.; Matuszek, C.; and Munson, S. A. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*.

Kennedy, L. S., and Naaman, M. 2008. Generating diverse and representative image search results for landmarks. In *Proceedings of the International World Wide Web Conference (WWW)*.

Otterbacher, J.; Bates, J.; and Clough, P. 2017. Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*.

Pollard, D. 1984. *Convergence of Stochastic Processes*. Springer.

Sauer, N. 1972. On the density of families of sets. *Journal of Combinatorial Theory, Series A* 13(1):145–147.

Singh, A., and Joachims, T. 2018. Fairness of exposure in rankings. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Texas Higher Education Opportunity Project. 2008. Administrative college application data: Documentation for public use data files. https://theop.princeton.edu/admin_doc.html.

U.S. Census Bureau. 2016. Full-time, year-round workers and median earnings: 2000 and 2013-2016. https://www.census.gov/data/tables/time-series/demo/industry-occupation/median-earnings.html.

U.S. Office of Management and Budget. 2018. Standard occupational classification manual. https://www.bls.gov/soc/2018/soc_2018_manual.pdf.

Vapnik, V., and Chervonenkis, A. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications* 16(2):264–280.

Wang, M.; Yang, K.; Hua, X.-S.; and Zhang, H.-J. 2010. Towards a relevant and diverse search of social images. *IEEE Transactions on Multimedia* 12(8):829–842.

Yang, K., and Stoyanovich, J. 2017. Measuring fairness in ranked outputs. In *Proceedings of the Conference on Scientific and Statistical Database Management (SSDBM)*.

Yang, K.; Stoyanovich, J.; Asudeh, A.; Howe, B.; Jagadish, H.; and Miklau, G. 2018. A nutritional label for rankings. In *Proceedings of the Conference on Management of Data (SIGMOD)*.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2017. Fairness constraints: Mechanisms for fair classification. *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Zehlike, M.; Bonchi, F.; Castillo, C.; Hajian, S.; Megahed, M.; and Baeza-Yates, R. 2017. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the Conference on Information and Knowledge Management (CIKM)*.

Additional related work

From a computational perspective, researchers studying algorithmic fairness have developed many algorithms assuming access to ground truth quality scores of individuals or outcomes. For example, in fair binary classification, a learning algorithm receives as input a training set of examples (x, y, z) where x is a set of observable attributes, z is a set of protected attributes, and y is a ground truth binary label. The goal is to learn a binary classifier which nearly matches the training set's ground truth labels, and which satisfies a given notion of fairness (e.g., (Agarwal et al., 2018; Dwork et al., 2018; Hardt, Price, and Srebro, 2016; Zafar et al., 2017)). Thus, the quality of a classifier is clear: it measures how well that classifier matches the training set or a test set of fresh examples, potentially with the addition of a fairness loss function. In our setting however, the true quality of any given criterion may not even be defined.

Several papers have studied fair machine learning in the context of unsupervised learning, though in settings that are orthogonal to ours. These include works on fair clustering (Chierichetti et al., 2017) and bias in word embeddings (Bolukbasi et al., 2016; Caliskan, Bryson, and Narayanan, 2017).

Ethical data usage

In this section, we discuss the ethics of our data usage. We use human data in two settings: college admissions and image search.

College admissions data. We use college admissions data collected by the Texas Higher Education Opportunity Project (THEOP), which is a part of Princeton University's Office of Population Research. We use a public use dataset which consists of college application data collected from the University of Texas at Austin, spanning from 1991 to 2003. Princeton's Institutional Review Board approved the administrative data collection and also the public release files. We requested access to these files and were granted it.

According to the data documentation (Texas Higher Education Opportunity Project, 2008), THEOP has taken steps to de-identify the data, including:

- Eliminating university-assigned applicant identification numbers and high school names, cities, and states.
- Eliminating small frequency (less than 20) cells by collapsing multiple values into range categories. For example, individual test score values are collapsed into test score ranges.
- Setting some values to missing to preserve the well-established categories but to hide individual values.

Nevertheless, it might be possible to re-identify individuals in this dataset. However, since this data is publicly available and (to the best of our knowledge) has been used by researchers in hundreds of academic works, we believe that our use of the dataset will not increase the risk of re-identification.

Image search. The other setting in which we use human data is image search. We do not download or store any images – we only store statistics in the form of Figure 3. We do not believe there is any risk of re-identification from these general statistics.

Additional results for university admissions

In this section, we generalize the setup in the main body by studying a setting where each criterion is represented by the intersection of multiple axis-aligned hyperplanes. For example, a criterion might require that an admitted applicant has an ACT score above 50 and is in the top 40% of her class. By generalizing the setup in the main body, we are able to preserve a higher degree of similarity while still increasing diversity.

Given a d-dimensional feature space $[0, H]^d$, every criterion t in the set T is defined by d thresholds $b_1, \ldots, b_d$. We use the notation $t = (b_1, \ldots, b_d)$. The set E_t of admitted applicants consists of all those whose feature vectors $\boldsymbol{x}$ satisfy $x[i] \geq b_i$ for all $i \in [d]$. In other words, $E_t = \{(\boldsymbol{x}, z) : (\boldsymbol{x}, z) \in E, x[i] \geq b_i, \forall i \in [d]\}$.

In our experiments, we use the same notion of diversity and similarity as in the main body. Namely,

$$\texttt{div}(t, E) = 1 - \left| 0.24 - \frac{1}{|E_t|} \sum_{(\boldsymbol{x}, z) \in E_t} z \right|$$

and

$$\sigma(t, t', E) = \begin{cases} 1 - \frac{|E_t \Delta E_{t'}|}{m} & \text{if } |E_t| \leq |E_{t'}| \leq 2\,|E_t| \\ 0 & \text{otherwise.} \end{cases}$$

Given a criterion $t = (b_1, \ldots, b_d)$, the goal is to find an alternative criterion $t' = (b'_1, \ldots, b'_d) \in T$ such that $\texttt{div}(t', E) \geq 0.95$ and $\sigma(t, t', E)$ is as close to 1 as possible. (Of course, 0.24 and 0.95 could be replaced by any user-specified constants.) As we prove in Theorem 3 via a reduction from the set cover problem, this problem is NP-complete. However, if the number of dimensions d is a constant, then it is possible to solve the problem in poly$(|E|)$ time since we can assume without loss of generality that $b'_i \in \{0, x_1[i], \ldots, x_m[i], H + 1\}$ for all $i \in [d]$.

We use the same set E of applicants as described in the main body and again set the input criterion t to automatically accept all students with a cumulative ACT score above 56. In other words, $t = (56, 0, 0)$, where the first component corresponds to ACT score, the second corresponds to SAT score, and the third corresponds to class rank. Figure 2 in the main body depicts our algorithm's output $t' = (44, 2140, 10)$, projected onto a 2-dimensional plane. The criterion t' suggests relaxing the ACT threshold but setting a stricter threshold on class rank and SAT score: students from the top 10% of their class with a cumulative SAT score above 2140 and a cumulative ACT score above 44 should be admitted. The number of white, non-Hispanic students returned by this criterion is 12,802 and the number of black or Hispanic students returned is 3,016. Therefore, the fraction of students admitted under this criterion who are black or Hispanic is 0.19. In other words $\texttt{div}(t', E) = 0.95$. Finally, $\sigma(t, t', E) = 0.75$.

Next, we prove that the above problem is NP-complete. We formalize the decision version of the problem below.

Definition 2 (Hyperplane intersection suggestion problem (HISP))**.** Given a set $E = \{(\boldsymbol{x}_1, z_1), \ldots (\boldsymbol{x}_m, z_m)\} \subseteq X \times Z$, a criteria $t = (b_1, \ldots, b_d)$ of examples, a diversity threshold $c \in [0, 1]$, and a symmetric difference bound s, is there an alternative criteria $t' = (b'_1, \ldots, b'_d)$ such that $|E_t \Delta E_{t'}| \leq s$ and $\frac{1}{|E_{t'}|} \sum_{(\boldsymbol{x},z) \in E_{t'}} z \geq c$?

We prove that the HICP is NP-complete via a reduction from the set cover problem, which we define below.

Definition 3 (Set cover problem (SCP))**.** Given a ground set $[n]$, a family $\mathcal{S} = \{S_1, \ldots, S_N\} \subseteq 2^{[n]}$ of subsets such that their union equals $[n]$ $\left(\text{i.e., } \bigcup_{i=1}^{N} S_i = [n]\right)$ and a bound k, are there at most k sets $S_{\ell_1}, \ldots, S_{\ell_k} \in \mathcal{S}$ such that $\bigcup_{j=1}^{k} S_{\ell_j} = [n]$?

Theorem 3. *The hyperplane intersection suggestion problem (HISP) is NP-complete.*

Proof. We give a reduction from the set cover problem to the HISP which operates in polynomial time. The reduction maps an arbitrary SCP input I_{SCP} to an HISP input I_{HISP}. The input I_{SCP} consists of a family $\mathcal{S} = \{S_1, \ldots, S_N\} \subseteq 2^{[n]}$ of N subsets of $[n]$ such that $\bigcup_{i=1}^{N} S_i = [n]$ and a bound k. The input I_{HISP} is defined as follows.

- The feature space X is N-dimensional, i.e., $X = \mathbb{R}^N$.
- The set E of input examples consists of $N + n + 1$ elements $(\boldsymbol{x}_1, z_1), \ldots, (\boldsymbol{x}_{N+n}, x_{N+n+1})$. For $i \leq n$, $\boldsymbol{x}_i$ indicates which sets in $\mathcal{S}$ the point i falls in. Specifically,
$$x_i[j] = \begin{cases} 0 & \text{if } i \in S_j \\ 1 & \text{otherwise.} \end{cases}$$
For $i \in \{1, \ldots, N\}$, $\boldsymbol{x}_{n+i}$ is all-ones vector minus the i^{th} standard basis vector. Specifically, $x_{n+i}[i] = 0$ and $x_{n+i}[j] = 1$ for all $j \neq i$. Finally, $\boldsymbol{x}_{N+n+1} = (1, \ldots, 1)$. The sensitive attributes $z_1, \ldots, z_{N+n+1} \in [0, 1]$ are defined such that if $i \leq n$, $z_i = 0$, and otherwise, $z_i = \frac{1}{2}$.
- The input criteria t is the all-zeros vector $(0, \ldots, 0)$.
- The diversity constant is $c = \frac{1}{2}$.
- The symmetric difference bound is $s = n + k$.

In Claims 1 and 2, we prove that the answer to the SCP given input I_{SCP} is "yes" if and only if the answer to the HICP given input I_{HICP} is "yes", and thus the reduction holds.

Claim 1. *If the answer to the SCP given input I_{SCP} is "yes", then the answer to the HICP given input I_{HICP} is "yes".*

Proof of Claim 1. Since the answer to the SCP is "yes", there must be k set $S_{\ell_1}, \ldots, S_{\ell_k} \in \mathcal{S}$ such that $\bigcup_{j=1}^{k} S_{\ell_j} = [n]$. Define the new criteria $t' = (b'_1, \ldots, b'_N)$ such that $b'_i = 1$ if $i \in \{\ell_1, \ldots, \ell_k\}$, and otherwise $b'_i = 0$. We claim that $|E_t \Delta E_{t'}| = s = k + n$ and $\frac{1}{|E_{t'}|} \sum_{(\boldsymbol{x},z) \in E_{t'}} z \geq c = \frac{1}{2}$, and thus the answer to the HICP given input I_{HICP} is "yes".

First, we will characterize which examples fall in the set $E_{t'}$, which will allow us to show that $|E_t \Delta E_{t'}| = s = k + n$. We claim that for all $\boldsymbol{x}_i$ such that $i \leq n$, $t'(\boldsymbol{x}_i) = 0$, and thus $(\boldsymbol{x}_i, z_i) \notin E_{t'}$. This is because there is some $j \in [k]$ such that $i \in S_{\ell_j}$. After all, $S_{\ell_1}, \ldots, S_{\ell_k}$ is a set cover. By definition of $b'_1, \ldots, b'_N$, we know that $b'_{\ell_j} = 1$, but by definition of $\boldsymbol{x}_i$, we know that $x_i[\ell_j] = 0$. Therefore, $t'(\boldsymbol{x}_i) = 0$ for all $i \leq n$. Next, we claim that for all vectors $\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{n+N}$, $t'(\boldsymbol{x}_{n+i}) = 0$ if and only if $i \in \{\ell_1, \ldots, \ell_k\}$. After all, suppose $i \in \{\ell_1, \ldots, \ell_k\}$. Then $x_{n+i}[i] = 0$, but $b'_i = 1$, which means that $t'(\boldsymbol{x}_{n+i}) = 0$. Next, suppose $i \notin \{\ell_1, \ldots, \ell_k\}$. Then for all $j \in \{\ell_1, \ldots, \ell_k\}$, $x_{n+i}[j] = 1$ and $b'_j = 1$, and for all $j \notin\in \{\ell_1, \ldots, \ell_k\}$, $x_{n+i}[j] \in \{0, 1\}$ and $b'_j = 0$. Therefore, $t'(x_{n+i}) = 1$. Finally, it's clear that the all-ones vector $t'(x_{N+n+1}) = 1$. In total, we have that

$$E_{t'} = \{(\boldsymbol{x}_{n+i}, z_{n+i}) : i \notin \{\ell_1, \ldots, \ell_k\}\} \cup \{(\boldsymbol{x}_{N+n+1}, z_{N+n+1})\}. \tag{2}$$

Since $E_t = E$, we have that $E_t \Delta E_{t'} = \{(\boldsymbol{x}_i, z_i) : i \leq n\} \cup \{(\boldsymbol{x}_{n+i}, z_{n+i}) : i \in \{\ell_1, \ldots, \ell_k\}\}$. Therefore, $|E_t \Delta E_{t'}| = s = k + n$.

Next, we claim that $\frac{1}{|E_{t'}|} \sum_{(\boldsymbol{x},z) \in E_{t'}} z \geq c = \frac{1}{2}$. This follows Equation (2) and the fact that $z_i = \frac{1}{2}$ for all $i > n$.

Therefore, the answer to the HICP given input I_{HICP} is "yes". □

Claim 2. *If the answer to the HICP given input I_{HICP} is "yes", then the answer to the SCP given input I_{SCP} is "yes".*

Proof of Claim 2. Since the answer to the HICP is "yes", there must be an alternative criteria $t' = (b'_1, \ldots, b'_N)$ such that $|E_t \Delta E_{t'}| \leq s = n + k$ and $\frac{1}{|E_{t'}|} \sum_{(\boldsymbol{x},z) \in E_{t'}} z \geq c = \frac{1}{2}$. Since $z_i = 0$ for all $i \leq n$, it must be that $(\boldsymbol{x}_i, z_i) \notin E_{t'}$ for all $i \leq n$. Since $E = E_t$, this means these n examples are elements of the symmetric difference $E_t \Delta E_{t'}$, so there can only be at most k other examples that fall in this set. In particular, there are at most k vectors $\boldsymbol{x}_i \in \{\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{n+N}\}$ such that $t'(\boldsymbol{x}_i) = 0$. Let K be the set $K = \{i : b'_i > 0\}$. We claim that if $i \in K$, then $t'(\boldsymbol{x}_{n+i}) = 0$. This is because $x_{n+i}[i] = 0$. Therefore, the size of K is bounded by the number of vectors $\boldsymbol{x}_{n+i}$ such that $t'(\boldsymbol{x}_{n+i}) = 0$, which we know is at most k. In other words, $|K| \leq k$.

Next, we claim that $\bigcup_{i \in K} S_i = [n]$. To see why this is, consider an arbitrary element $j \in [n]$. Since $t'(\boldsymbol{x}_j) = 0$, we know there exists $i \in K$ such that $b'_i > 0$ and $x_j[i] = 0$. After all, $\boldsymbol{x}_j \in \{0,1\}^N$, so the only other option is that for some $i \in K$, $b'_i > 1$. But this cannot be the case because if $b'_i > 1$ for some i, then $E_{t'} = \emptyset$, which is impossible given that $|E_t \Delta E_{t'}| \leq n + k < n + N + 1 = |E_t|$ and $|E_t \Delta \emptyset| = |E_t|$. Therefore, there exists $i \in K$ such that $b'_i > 0$ and $x_j[i] = 0$. By definition of $\boldsymbol{x}_j$, this means that $j \in S_i$. Therefore, $\bigcup_{i \in K} S_i = [n]$, so the answer to the SCP is "yes". □

□

The Standard Occupation Classification System

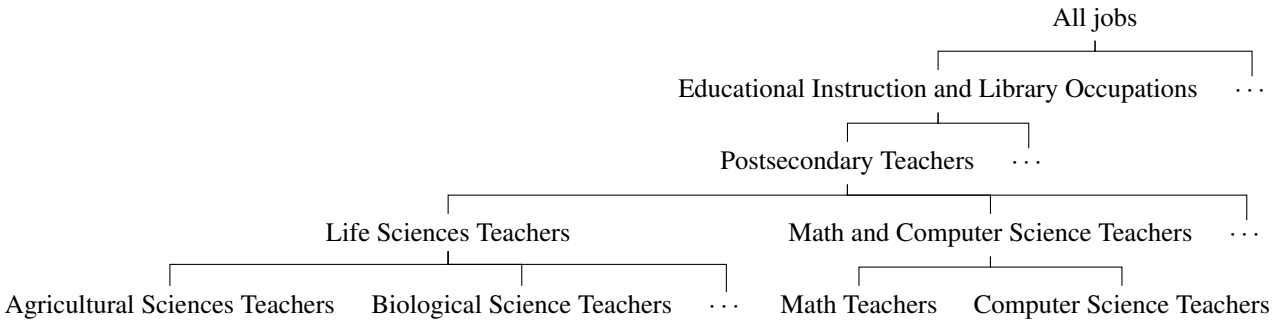


Figure 5: A small subset of the Department of Labor's Standard Occupational Classification System.

Proofs of diversity estimation guarantees

It is possible to provide sample complexity guarantees for any real-valued function class $\mathcal{H}$ using VC dimension, even though VC dimension is a complexity measure that only applies to binary function classes. To derive these guarantees, it is sufficient to bound the VC dimension of the class $\mathcal{G}_{\mathcal{H}}$ of *below-the-graph indicator functions*, defined as follows.

Definition 4 (Below-the-graph indicator functions)**.** Let $\mathcal{H}$ be a class of functions mapping a domain $\mathcal{A}$ to $[0,1]$. The class $\mathcal{G}_{\mathcal{H}} = \{g_h : h \in \mathcal{H}\}$ of *below-the-graph indicator functions* is defined such that $g_h : \mathcal{A} \times [0,1] \to \{0,1\}$ and $g_h(a,y) = \text{sign}(h(a) - y)$.

Lemmas 1 and 2 demonstrate how the VC dimension of the class of below-the-graph indicator functions can be used to provide sample complexity guarantees for real-valued function classes.

Lemma 1 (Theorems 18.4 and 19.7 of (Anthony and Bartlett, 2009))**.** *Let $\mathcal{H}$ be a class of functions mapping an abstract domain $\mathcal{A}$ to $[0,1]$ and let $\mathcal{G}_{\mathcal{H}}$ be the class of below-the-graph indicator functions. Then for any $\alpha, \beta, \delta \in (0,1)$ and any distribution μ' over $\mathcal{A}$, with probability at least $1 - \delta$ over the draw of $a_1, \ldots, a_N \sim \mu'$, for all $h \in \mathcal{H}$, $(1-\alpha)\mathbb{E}_{a \sim \mu'}[h(a)] - \beta < \frac{1}{N}\sum_{i=1}^{N} h(a_i) < (1+\alpha)\mathbb{E}_{a \sim \mu'}[h(a)] + \beta$ when $N \geq N_0 \in O\left(\frac{1}{\alpha\beta}\left(\text{VCdim}\left(\mathcal{G}_{\mathcal{H}}\right)\log\frac{1}{\beta} + \log\frac{1}{\delta}\right)\right)$.*

Lemma 2 (Pollard (1984); Dudley (1967))**.** *Let $\mathcal{H}$ be a class of functions mapping an abstract domain $\mathcal{A}$ to $[0,1]$ and let $\mathcal{G}_{\mathcal{H}}$ be the class of below-the-graph indicator functions. For any $\epsilon, \delta \in (0,1)$ and any distribution μ' over $\mathcal{A}$, with probability at least $1 - \delta$ over the draw of $a_1, \ldots, a_N \sim \mu'$, for any $h \in \mathcal{H}$, $\left|\frac{1}{N}\sum_{i=1}^{N} h(a_i) - \mathbb{E}_{a \sim \mu'}[h(a)]\right| \leq \epsilon$ when $N \geq N_0 \in \Theta\left(\frac{1}{\epsilon^2}\left(\text{VCdim}\left(\mathcal{G}_{\mathcal{H}}\right) + \log\frac{1}{\delta}\right)\right)$.*

Proof of Theorem 1

In this section, we prove Theorem 1. To do so, we use the following helpful lemma.

Lemma 3. *Let* $\left\{E^{(1)}, \ldots, E^{(M)}\right\}$ *be a set of elements in* $(X \times Z)^m$ *and let* $v = \text{VCdim}(T)$. *Then*

$$\left| \left\{ \begin{pmatrix} f_m\left(t, E^{(1)}\right) \\ \vdots \\ f_m\left(t, E^{(M)}\right) \end{pmatrix} : t \in T \right\} \right| \leq (mM+1)^v.$$

Proof. Let $\mathcal{S}' = \{x_1, \ldots, x_{Mm}\}$ be the set of all observable attributes x such that $(x, z) \in E^{(i)}$ for some $E^{(i)}$ and some $z \in [0, 1]$. Classic results from learning theory (Sauer, 1972) allow us to bound the number of ways criteria in T can label the set $\mathcal{S}'$. In particular,

$$\left| \left\{ \begin{pmatrix} t\left(x_1\right) \\ \vdots \\ t\left(x_{mM}\right) \end{pmatrix} : t \in T \right\} \right| \leq (mM+1)^v.$$

Suppose t and t' are two criteria such that $t\left(x_i\right) = t'\left(x_i\right)$ for all $i \in [mM]$. We claim that $f_m\left(t, E^{(i)}\right) = f_m\left(t', E^{(i)}\right)$ for all $i \in [M]$. For a contradiction, suppose

$$\begin{aligned} f_m\left(t, E^{(i)}\right) &= \frac{1}{\left|E_t^{(i)}\right|} \sum_{(x,z) \in E_t^{(i)}} z \\ &\neq \frac{1}{\left|E_{t'}^{(i)}\right|} \sum_{(x,z) \in E_{t'}^{(i)}} z = f_m\left(t', E^{(i)}\right) \end{aligned}$$

for some $i \in [M]$. This means that $E_t^{(i)} = \left\{(x, z) : (x, z) \in E^{(i)}, t(x) = 1\right\} \neq \left\{(x, z) : (x, z) \in E^{(i)}, t'(x) = 1\right\} = E_{t'}^{(i)}$. Therefore, there is some example $(x, z) \in \mathbb{E}^{(i)}$ such that $t(x) \neq t'(x)$, which is a contradiction, so the claim holds. □

Theorem 1. *For any* $\epsilon \in (0, 1/2)$ *and* $\delta \in (0, 1)$, *with probability* $1 - \delta$ *over the draw* $E = \{(x_1, z_1), \ldots, (x_N, z_N)\} \sim \mu^N$, *for all* $t \in T$ *such that* $\Pr_{(x,z)}[t(x) = 1] \geq c$, $\left|f_m(t, E) - \mathbb{E}_{(x,z)\sim\mu}[z \mid t(x) = 1]\right| \leq \epsilon$ *when* $N \geq N_0 \in O\left(\frac{1}{c\epsilon^2}\left(\text{VCdim}(T) \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right)\right)$.

Using Lemma 1, we show that the numerator and the denominator of $f_m(t, E) = \dfrac{\sum_{i=1}^N z_i t(x_i)}{\sum_{i=1}^N t(x_i)}$ are concentrated around their means for all $t \in T$.

Claim 3. *For any* $\epsilon \in (0, 1)$ *and* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$ *over the draw of* $(x_1, z_1), \ldots, (x_N, z_N) \sim \mu$, *for all* $t \in T$ *such that* $\Pr_{(x,z)}[t(x) = 1] \geq c$, *we have*

$$(1 - \epsilon) \Pr_{(x,z)\sim\mu}[t(x) = 1] \leq \frac{1}{N} \sum_{i=1}^N t(x_i) \leq (1 + \epsilon) \Pr_{(x,z)\sim\mu}[t(x) = 1].$$

when $N \geq N_0 \in O\left(\frac{1}{c\epsilon^2}\left(\text{VCdim}(T) \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right)\right)$.

Proof. This follows immediately from Lemma 1 by setting $\alpha = \frac{\epsilon}{2}$ and $\beta = \frac{c\epsilon}{2}$. □

Claim 4. *For any* $\epsilon \in (0, 1)$ *and* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$ *over the choice of* $(x_1, z_1), \ldots, (x_N, z_N) \sim \mu$, *for all* $t \in T$,

$$(1 - \epsilon)\, \mathbb{E}_{(x,z)\sim\mu}[zt(x)] - c\epsilon < \frac{1}{N} \sum_{i=1}^N z_i t(x_i)$$

and

$$(1 + \epsilon)\, \mathbb{E}_{(x,z)\sim\mu}[zt(x)] + c\epsilon > \frac{1}{N} \sum_{i=1}^N z_i t(x_i),$$

when $N \geq N_0 \in O\left(\frac{1}{c\epsilon^2}\left(\text{VCdim}(T) \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right)\right)$.

Proof. Let $\alpha = \frac{\epsilon}{2}$ and $\beta = \frac{c\epsilon}{2}$. Using Lemma 1, it suffices to show that the VC dimension of class of functions

$$\mathcal{G} = \{g_t : (x, z, y) \mapsto \text{sign}(zt(x) - y)\}_{t \in T}$$

is at most VCdim(T). To do this, we show that if there is a set of examples $(x_1, z_1), \ldots, (x_M, z_M)$, and threshold values $y_1, \ldots, y_M$, such that

$$\left\{ \begin{pmatrix} \text{sign}(z_1 \cdot t(x_1) - y_1) \\ \vdots \\ \text{sign}(z_M \cdot t(x_m) - y_M) \end{pmatrix} : t \in T \right\} = \{0, 1\}^M, \tag{3}$$

then $M \leq$ VCdim(T).

Consider such a set of M examples and thresholds that satisfy Equation (3). Note that for this to be true, it must be that $y_i < z_i$ for all $i \in [M]$. Under such conditions, $\text{sign}(z_i \cdot t(x_i) - y_i) = t(x_i)$. Therefore,

$$\left\{ \begin{pmatrix} t(x_1) \\ \vdots \\ t(x_M) \end{pmatrix} : t \in T \right\} = \{0, 1\}^M,$$

as well. This implies that T shatters the set of M examples $(x_1, \ldots, x_M)$, hence, $M \leq$ VCdim(T). Therefore, VCdim$(\mathcal{G}) \leq$ VCdim(T). □

We now prove Theorem 1.

Proof of Theorem 1. First, note that

$$\begin{aligned} \frac{\mathbb{E}_{(x,z)\sim\mu}[zt(x)]}{\mathbb{E}_{(x,z)\sim\mu}[t(x)]} &= \frac{\mathbb{E}[zt(x)]}{\Pr[t(x) = 1]} \\ &= \frac{\mathbb{E}\left[z\mathbb{1}_{\{t(x)=1\}}\right]}{\Pr[t(x) = 1]} \\ &= \frac{\mathbb{E}\left[z \mid t(x) = 1\right] \cdot \Pr[t(x) = 1]}{\Pr[t(x) = 1]} \\ &= \mathbb{E}\left[z \mid t(x) = 1\right]. \end{aligned}$$

Next, by Claims 3 and 4, for $N \geq N_0$,

$$\begin{aligned} \frac{\sum_{i=1}^N z_i t(x_i)}{\sum_{i=1}^N t(x_i)} &\leq \frac{(1 + \epsilon/4)\mathbb{E}[zt(x)] + c\epsilon/4}{(1 - \epsilon/4)\mathbb{E}[t(x)]} \\ &= \frac{(1 + \epsilon/4)\mathbb{E}[zt(x)]}{(1 - \epsilon/4)\mathbb{E}[t(x)]} + \frac{c\epsilon/4}{(1 - \epsilon/4)\mathbb{E}[t(x)]} \\ &\leq \frac{1 + \epsilon/4}{1 - \epsilon/4} \cdot \frac{\mathbb{E}[zt(x)]}{\mathbb{E}[t(x)]} + \frac{c\epsilon/4}{(1 - \epsilon/4)c} && (\mathbb{E}[t(x)] \geq c) \\ &\leq \left(1 + \frac{\epsilon/2}{1 - \epsilon/4}\right) \frac{\mathbb{E}[zt(x)]}{\mathbb{E}[t(x)]} + \frac{\epsilon/4}{1 - \epsilon/4} \\ &\leq \frac{\mathbb{E}[zt(x)]}{\mathbb{E}[t(x)]} + \frac{\epsilon/2 + \epsilon/4}{1 - \epsilon/4} && \left(\frac{\mathbb{E}[zt(x)]}{\mathbb{E}[t(x)]} \leq 1\right) \\ &\leq \frac{\mathbb{E}[zt(x)]}{\mathbb{E}[t(x)]} + \frac{\epsilon/2 + \epsilon/4}{1 - 1/8} && \left(\epsilon \leq \frac{1}{2}\right) \\ &< \frac{\mathbb{E}[zt(x)]}{\mathbb{E}[t(x)]} + \epsilon \\ &= \mathbb{E}\left[z \mid t(x) = 1\right] + \epsilon. \end{aligned}$$

From the other direction,

$$\begin{aligned}
\frac{\sum_{i=1}^{N} z_i t(x_i)}{\sum_{i=1}^{N} t(x_i)} &\geq \frac{(1-\epsilon/4)\mathbb{E}[zt(x)] - c\epsilon/4}{(1+\epsilon/4)\mathbb{E}[t(x)]} \\
&= \frac{(1-\epsilon/4)\mathbb{E}[zt(x)]}{(1+\epsilon/4)\mathbb{E}[t(x)]} - \frac{c\epsilon/4}{(1+\epsilon/4)\mathbb{E}[t(x)]} \\
&\geq \frac{1-\epsilon/4}{1+\epsilon/4} \cdot \frac{\mathbb{E}[zt(x)]}{\mathbb{E}[t(x)]} - \frac{c\epsilon/4}{(1+\epsilon/4)c} && (\mathbb{E}[t(x)] \geq c) \\
&\geq \left(1 - \frac{\epsilon/2}{1+\epsilon/4}\right) \frac{\mathbb{E}[zt(x)]}{\mathbb{E}[t(x)]} - \frac{\epsilon/4}{1+\epsilon/4} \\
&\geq \frac{\mathbb{E}[zt(x)]}{\mathbb{E}[t(x)]} - \frac{\epsilon/2+\epsilon/4}{1+\epsilon/4} && \left(\frac{\mathbb{E}[zt(x)]}{\mathbb{E}[t(x)]} \leq 1\right) \\
&\geq \frac{\mathbb{E}[zt(x)]}{\mathbb{E}[t(x)]} - \frac{\epsilon/2+\epsilon/4}{1+1/8} && \left(\epsilon \leq \frac{1}{2}\right) \\
&> \frac{\mathbb{E}[zt(x)]}{\mathbb{E}[t(x)]} - \epsilon \\
&= \mathbb{E}\left[z \mid t(x) = 1\right] - \epsilon.
\end{aligned}$$

Therefore, the theorem holds. □

Proof of Theorem 2

We now prove Theorem 2, which we restate below.

Theorem 2. *For any $\epsilon, \delta \in (0,1)$, $N \geq N_0 \in \Theta\left(\frac{1}{\epsilon^2}\left(\text{VCdim}(T)\log\left(m\text{VCdim}(T)\right) + \log\frac{1}{\delta}\right)\right)$ samples are sufficient to ensure that with probability $1-\delta$ over the draw of $E^{(1)}, \ldots, E^{(N)} \sim \bar{\mu}$, for all $t \in T$, $\left|\frac{1}{N}\sum_{i=1}^{N} f_m\left(t, E^{(i)}\right) - \mathbb{E}_{E\sim\bar{\mu}}[f_m(t,E)]\right| \leq \epsilon$.*

Proof. By Lemma 2, we only need to bound the VC dimension of the class of below-the-graph indicator functions corresponding to the set of functions $f_m(t,\cdot)$, which take as input any set $E \in (X \times Z)^m$ of m examples and return the diversity measurement $f_m(t,E)$. We use the notation $\mathcal{D} = \{f_m(t,\cdot) \mid t \in T\}$ to denote this set of real-valued functions and $\mathcal{G}_{\mathcal{D}} = \{g_t : (E,y) \mapsto \text{sign}(f_m(t,E) - y) \mid t \in T\}$ to denote the corresponding set of below-the-graph indicator functions. Indeed, in order to prove this theorem, it is enough to prove that $\text{VCdim}\left(\mathcal{G}_{\mathcal{D}}\right) = O\left(\text{VCdim}(T)\log\left(m\text{VCdim}(T)\right)\right)$.

Suppose $\text{VCdim}\left(\mathcal{G}_{\mathcal{D}}\right) = M$. This means that there exists a set $\left\{\left(E^{(1)}, y^{(1)}\right), \ldots, \left(E^{(M)}, y^{(M)}\right)\right\} \subset (X \times Z)^m \times [0,1]$ such that for all binary vectors $\boldsymbol{b} \in \{0,1\}^M$, there exists a criterion $t_{\boldsymbol{b}} \in T$ such that $\text{sign}\left(f_m\left(t_{\boldsymbol{b}}, E^{(i)}\right) - y^{(i)}\right) = b[i]$ for all $i \in [M]$.

Let $\mathcal{S}' = \{x_1, \ldots, x_{Mm}\}$ be the set of all observable attributes x such that $(x,z) \in E^{(i)}$ for some $E^{(i)}$ and some $z \in [0,1]$. Let $v = \text{VCdim}(T)$. Classic results from learning theory (Sauer, 1972) allow us to bound the number of ways criteria in T can label the set $\mathcal{S}'$. In particular, $\left|\left\{\left(t\left(x_1\right), \ldots, t\left(x_{mM}\right)\right) : t \in T\right\}\right| \leq (mM+1)^v$. This fact allows us to bound the number of ways functions in $\mathcal{D}$ can label $E^{(1)}, \ldots, E^{(M)}$. Specifically, in Lemma 3, we prove that $\left|\left\{\left(f_m\left(t, E^{(1)}\right), \ldots, f_m\left(t, E^{(M)}\right)\right) : t \in T\right\}\right| \leq (mM+1)^v$. Thus,

$$2^M = \left|\left\{\begin{pmatrix} \text{sign}\left(f_m\left(t, E^{(1)}\right) - y^{(1)}\right) \\ \vdots \\ \text{sign}\left(f_m\left(t, E^{(M)}\right) - y^{(M)}\right) \end{pmatrix} : t \in T\right\}\right| \leq (mM+1)^v.$$

This means that $M = O(v\log(vm))$, and since $M = \text{VCdim}\left(\mathcal{G}_{\mathcal{D}}\right)$ and $v = \text{VCdim}(T)$, we know that

$$\text{VCdim}\left(\mathcal{G}_{\mathcal{D}}\right) = O\left(\text{VCdim}(T)\log\left(m\text{VCdim}(T)\right)\right),$$

as desired. □