
Implications of Modeled Beliefs for Algorithmic Fairness in Machine Learning

Jeff Edmonds Karan Singh Ruth Uner

Department of Electrical Engineering and Computer Science
York University, Toronto, Canada

Abstract

Ethics and societal implications of automated decision making have become a major theme in Machine Learning research. Conclusions from theoretical studies in this area are often stated in general terms (such as affirmative action possibly hurting all groups or fairness measures being incompatible with a decision maker being rational). Our work aims to highlight the degree to which such conclusions are in fact relying on modeled beliefs as well as on the technicalities of a chosen framework of analysis (eg. statistical learning theory, game theory, dynamics etc). We carefully discuss prior work through this lens and then highlight the effect of modeled beliefs by means of a simple statistical model where an observed score X is the result of two unobserved hidden variables ("talent" T and "environment" E). We assume that variable T is identically distributed for two subgroups of a population while E models the disparities between an advantaged and a disadvantaged group. We analyze (Bayes-)optimal decision making under a variety of distributional assumptions and show that even the simple model under consideration exhibits some counterintuitive effects.

1 Introduction

Ethics and societal implications of data-derived decision rules have become a major theme in Machine Learning research as well as the public perception and discussion of Machine Learning (or more generally AI) based technology. Conclusions from theoretical studies in this area are often stated in general terms (such as "subsidy improves only the learner's utility while actually making both candidate groups worse-off" [10], or fairness goals with rational decision makers being only achievable when evaluation scores are not reported to the rational decision maker [11]). However, these are typically derived in very specific mathematical frameworks, under specific mathematical assumptions on a data-generating process. While these assumptions are usually stated clearly in the technical parts, they may get overlooked by a technically less versed reader, or a reader mainly focusing on the main take-aways of some published article in an area of such high societal interest and relevance.

In this work, we would like to highlight the role of mathematically modeled beliefs when framing and analyzing questions of algorithmic fairness for automated decision making. We show how, even in a simple setting (which can be viewed as a simple causal generative model), the conclusions depend heavily on properties of the underlying data-generating process. Our aim with this work is to raise awareness for how a modeled belief will influence conclusions even in seemingly clean mathematical studies.

As is common in the fairness literature, we assume that there are two subgroups of some general population, one being in a disadvantaged position. We denote the advantaged group with A and the disadvantaged group with B and consider the following setup: We assume a statistical causal model

where an observed score X is the result of two unobserved hidden variables T and E . The score X can be thought of as individual’s observable *performance* on screening decisions such as resume, grades, school, letters, interviews etc. It is how good the person looks “on paper”. For ease of reading, we will suggestively name the two influencing variables “talent” T and “environment” E . We assume that variable T is identically distributed for two subgroups A and B of a population while E models the disparities between an advantaged and a disadvantaged group. We emphasize that we are not making any claims about what the variables T and E correspond to in a real world scenario (this will in fact vary with applications). What we model here is a scenario where there are various factors that influence an observed outcome X , some of which are distributed identically for the two subpopulations in question (think talent T), and some of which are reflecting a disparity between the groups (think environment E).

We assume an additive model for how the score gets influenced by talent and environment, namely

$$X_g = T_g + E_g \quad \text{for } g \in \{A, B\}.$$

We assume that the talent variables T_A and T_B for the two groups are distributed identically while the environment variable E_A is distributed identically to a *shift* of the environment variable E_B of the disadvantaged group, namely E_A is distributed identically to $E_B + k$ for some $k > 0$. The shift k models the advantage that individuals from group A enjoys (on average). Now we consider a decision maker gets to make choice on individuals from the two groups at a given score $X_g = x$ (or above a given score $X_g \geq x$). Further, we assume that, from the decision maker’s perspective, the more relevant factor for a positive decision is a high value of T .

Intuitively, one would assume that at a fixed score $X_A = X_B = x$, the expected value for the talent-variable T_g is always higher for the disadvantaged group than for the advantaged group. If an individual that starts out in a disadvantaged position achieves the same score as an individual from an advantaged background, we expect that the disadvantaged individual must have possessed an inert quality to overcome the hurdles. We analyze here whether and when this is also the case in a statistical decision making setup. We identify a large variety of distributions, where this intuitive conclusion is indeed justified. Surprisingly, we also show that this however depends on the specifics of the distributions of T and E and that in some cases, even under our seemingly straightforward additive generative modeling, the expected talent of the advantaged group can be higher. This shows *how subtle differences in assumptions and frameworks can have a significant impact on derived conclusions*.

In this extended abstract version of our work, we will discuss a few illustrative examples to highlight the phenomena and describe the more general versions of our technical results on an intuitive level. The full technical details can be found in the appendix.

2 Related Work

Ethics and societal implications of automated decision making have become a major theme in Machine Learning research. That fairness to various individuals or subpopulations is an important consideration in predictive modeling has first been argued and formalized a decade ago in the seminal work of Dwork et al [7]. Since then, a large body of work has addressed this theme. Various formalizations of what it means for a predictor to act fair have been introduced, the main distinction being between notions of *individual fairness* [7, 12, 18], which typically formalize the requirement that “similar individual should be treated similarly” in some form, and notions of *group fairness* [9] which aim to ensure equal treatment. The bulk of studies have then focused on analyzing the proposed notions, their mutual (in-)compatibility [13, 4] and developed algorithmic tools to facilitate them. For example, a substantial body of work has developed methods for learning or building *fair representations*, that is, feature mappings that are aimed at ensuring fairness (typically in the group-fairness sense) of any predictor based on these features [15, 6]. Another common approach is to develop optimization procedures that adhere to constraints given by fairness considerations. Some recent studies have analyzed dynamics of fairness interventions over processes of repeated decision making [14, 17]. Several recent surveys [5, 16] and a textbook [2] give more comprehensive introductions to the field than we can here.

We view our work in a line of studies aim at analyzing the effects of various fairness interventions in dependence on the properties of the affected populations, and highlighting the role that different *worldviews* may play for recommending one fairness intervention over another [8, 19]. In particular,

we adopt the perspective developed and outlined by Friedler et al. [8], which formalizes the occurrence of unfairness as a discrepancy between equality notions in (unobserved) *construct feature and decision spaces* and their representations in the *observed feature and decision spaces*. Their work argues how a choice of fairness intervention is dependent on (or a reflection of) different beliefs about the distributions of various groups in the unobserved construct feature space. For example, a What-You-See-Is-What-You-Get (WYSIWYG) worldview assumes that measured/observed features are an adequate reflection of the underlying construct feature distributions, and hence an inequality in observed feature distributions of two subpopulations is merely an indication of inherent differences between the two groups. Such a worldview may then lead to advocating for individual fairness notions. On the other hand, a We-Are-All-Equal (WAE) worldview postulates that, in the construct space the relevant qualities (what these are will differ in various contexts) are distributed identically and therefore discrepancies in the observed space are a sign of adversity that one of the groups may have faced. Such a worldview naturally leads to advocating for interventions that ensure notions of group fairness. A recent study analyzes the effects of enforcing statistical parity and equalized odds under the WAE and WYSIWYG worldviews [19]. Another recent study that adopts this perspective analyzes a statistical learning setting and analyzes, under an WAE assumption and specific modelings of disadvantageous bias to one of the groups, whether enforcing those same group fairness requirements will lead to recovering a Bayes optimal prediction rule [1]. This latter work argues purely in terms of frequencies observed in the two groups and ignores the role that specific (observed or unobserved) features may play in the biasing process.

Our modeling can be viewed in response to recent models on (repeated) binary decision making under certain assumptions on the source of discrepancy in the observed scores. A recent work on modeling effects of affirmative action starts out with a similar setup to ours [11]: An admission decision is made based on a score $S_i = T_i + X$ which is a sum of a random *type* variable and a noise variable X . Here, it is assumed that the type variable has a lower mean for the disadvantaged group, while the noise is a zero-mean addition which is equally distributed for both groups. The goal is to recover the type of the individual, which represents an individual’s quality in this study. The study then analyzes whether a university can set up an admission rule so as to ensure fairness for hiring decisions that are made at a second stage by a rational, and ideally group blind employer. The conclusions in the study are mostly pessimistic. We believe that this is due to the modeling, which is based on the disadvantaged group being of *inherently lower quality*. We consider this a reflection of a WYSIWYG worldview and believe this should be made explicit.

Another recent study that considers a score-based decision task assumes a WAE worldview [10]. The work shows that, if the relevant feature is distributed equally between the two groups, while the difference in score distributions is the result of a *manipulation* that the two groups are differently equipped to perform (the motivating example here is that different groups may have different abilities to pay for expensive tutoring sessions for SAT scores), then a group blind thresholding will result in unfairness in the sense that the advantaged group will enjoy more false positive decisions while the disadvantaged group members will suffer from more false negative decision. That work also considers the possibility of an intervention in form of subsidies but with pessimistic conclusions. In that study the advantage is modeled as a deterministic response to a published threshold (if paying for tutoring sessions can get me admitted, I will pay, otherwise I won’t bother). In contrast, in our work the advantage is modeled as a random variable, namely an *environmental factor* that an individual did not choose, and which is distributed differently between the two groups.

Ultimately, the main goal of our work is to argue for making modeling assumptions, that might be influenced by worldviews of authors, more explicit in academic work that analyzes effects of fairness interventions. In this work, we decidedly adopt a We-Are-All-Equal worldview. This is modeled by the relevant contributing factor T being identically distributed among subgroups that are considered for a fairness intervention.

3 Some Illustrative examples

As a warm-up, we will here illustrate the phenomena described in the introduction in a simple, discrete setting. In the next section, we will summarize our more general results. The two illustrative examples in this section are intently simplistic, to highlight the phenomenon. We consider random variables T and E that take values in $\{1, 2, 3, 4\}$ and we condition on the score being $X = 4$ or $X \geq 4$. Figure 1 illustrates the distributions in the two scenarios defined below.

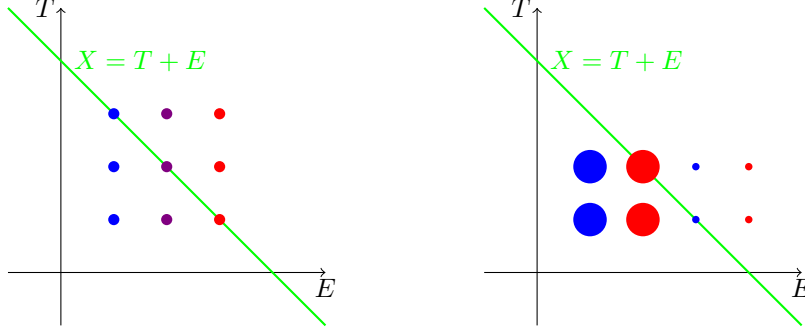


Figure 1: Illustration of the distributions in Scenario 1 (left hand side) and Scenario 2 (right hand side). Blue dots depict the distribution of the disadvantaged group, while the red dots the distribution of the advantaged group.

Scenario 1 Consider the following distribution for the random variables: The talent variable is distributed uniformly over the values $\{1, 2, 3\}$ for both groups, while the environment variables E_B is distributed uniformly over $\{1, 2\}$ and E_A uniformly over $\{2, 3\}$ (that is, we have $k = 1$ here). A simple calculation shows that we obtain the following expected talent scores for these distributions:

$$\begin{aligned}\mathbb{E}[T_B|X_B = 4] &= 2.5 \\ \mathbb{E}[T_A|X_A = 4] &= 1.5 \\ \mathbb{E}[T_B|X_B \geq 4] &= \frac{8}{3} = 2.\bar{6} \\ \mathbb{E}[T_A|X_A = 4] &= \frac{11}{5} = 2.2\end{aligned}$$

Thus, in this scenario we see the intuitively expected effect. If group B has a disadvantaged environment score, while the talent is identically distributed between the groups, then, conditioning on a fixed score or a threshold, the expected talent in the disadvantaged group is higher.

Scenario 2 Consider the following distribution for the random variables: The talent variable is distributed uniformly over the values $\{1, 2\}$ for both groups, while the environment variables E_B takes value 1 with probability 0.9 and value 3 with probability 0.1. The environment variable E_A again is identically distributed to $E_B + 1$, that is, it takes value 2 with probability 0.9 and value 4 with probability 0.1. With these distributions we obtain the following for the expected talent scores:

$$\begin{aligned}\mathbb{E}[T_B|X_B = 4] &= 1 \\ \mathbb{E}[T_A|X_A = 4] &= 2 \\ \mathbb{E}[T_B|X_B \geq 4] &= 1.5 \\ \mathbb{E}[T_A|X_A = 4] &> 1.9\end{aligned}$$

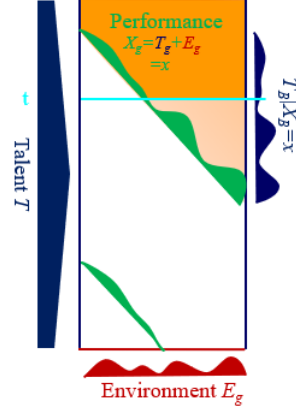
Thus, in this scenario we see the opposite effect. This is an example of the unintuitive case, where the expected talent will be larger for the advantaged group. We will see that this phenomenon can occur generally in the case where the environment range is wider than the talent range (see Section B in the appendix). In this particular case, to illustrate this phenomenon, we set things up so that for the majority of the disadvantaged population, the score is set so that even being "maximally talented" is not sufficient to reach the required level for the score X .

4 Summary of results for more general distribution classes

In this section, we provide an overview on our general technical results. More detailed notes on these (and proofs) can be found in the appendix.

Wide Uniform Talent Distributions We start by identifying broad classes of distributions where the conclusions are as in Scenario 1 above. Namely, given a fixed score, candidates from the disadvantaged group have a higher expected talent. Our first result shows this is the case whenever the talent distribution of T_A and T_B is uniform, the result holds for arbitrary environment distributions.

Consider the illustration on the right. The talents T_A and T_B of individuals in the two groups are random variables drawn independently according to the same arbitrary distribution. The blue shape on the left depicts this distribution rotated. Their environments E_A and E_B are formed by sampling independently from two other arbitrary distributions. The one for group g is depicted at the bottom in red. Both of these curves depicted a height coming up out of the page in the 3rd dimension. In this way, the height of the joint density $d(t, e)$ of any point in the rectangle is the product $d(t, e) = d(t) \cdot d(e)$ of the corresponding probability density functions. The top diagonal green curve represents the probability space of possibilities in which g 's performance $X_g = T_g + E_g$ is conditioned to be a fixed value x . We assume here, that the conditioning (the green line in the figure) is on a value $X = x$ that restricts the talent range but not the environment range, as is the case for the upper green line. This is to be understood in contrast to the case of the the bottom green line, which restricts the talent range and also restricts the environment range.



(See also Figure 2, where the green lines restrict the environment range but not the talent range.) We obtain the following technical result:

Theorem 1. *Let T_A and T_B be distributed identically, uniformly over some range. Let E_B be arbitrarily distributed in some bounded range and let E_A be identically distributed to $E_B + k$ for some $k > 0$. Let $X_A = T_A + E_A$, and $X_B = T_B + E_B$, and let $\bar{k} = \mathbb{E}(E_A) - \mathbb{E}(E_B)$. If x is so that conditioning on $X_g = x$ does not restrict the range of E_A and E_B (while it may restrict the range of T_A and T_B), we get:*

- $\mathbb{P}(T_B \geq t_A | X_B = x) \Rightarrow \mathbb{P}(T_A \geq t_A | X_A = x)$,
- $\mathbb{P}(T_B \geq t_A | X_B = x_A - k) = \mathbb{P}(T_A \geq t_A | X_A = x)$, and $\mathbb{E}(T_B | X_B = x - \bar{k}) = \mathbb{E}(T_A | X_A = x_A)$;
- $\mathbb{P}(T_B \geq t_A + k | X_B = x) = \mathbb{P}(T_A \geq t_A | X_A = x)$, and $\mathbb{E}(T_B | X_B = x) - \mathbb{E}(T_A | X_A = x) = \bar{k}$.

Wide Environment or Non-Uniform Talent We now deal with a set of complementary conditions from the above, namely the case where the talent distributions can be arbitrary (but in a restricted range), while the environment variables are from a restricted class of distributions, namely from a log-concave distribution [3]. In these cases, we again observe the intuitive behavior of the expected talent values. A probability distribution is *log-concave* if the logarithm of its density function is convex and includes concave, linear, and Gaussian distributions. More over, it includes any density function that is sub-exponentially increasing in the first sub-range, then concave, then sub-exponentially decreasing.

Theorem 2. *Let talent distribution for T_A and T_B be arbitrarily, but identically distributed within a narrow range $[t_{min}, t_{max}]$. Let environments E'_A and E_B be identically distributed with a log-concave density function, and let $E_A = E'_A + k$. Condition on the both of their performance scores $X_g = T_g + E_g$ being some fixed value x . Suppose this condition restricts the environment range but not the talent range. Then we have:*

$$\begin{aligned}
 \mathbb{P}(T_B \geq t | X_B = x - k) &= \mathbb{P}(T_A \geq t | X_A = x) \\
 \mathbb{E}(T_B | X_B = x_A - k) &= \mathbb{E}(T_A | X_A = x_A) \\
 \mathbb{P}(T_B \geq t | X_B = x) &\geq \mathbb{P}(T_A \geq t | X_A = x) \\
 \mathbb{E}(T_B | X_B = x) &\geq \mathbb{E}(T_A | X_A = x).
 \end{aligned}$$

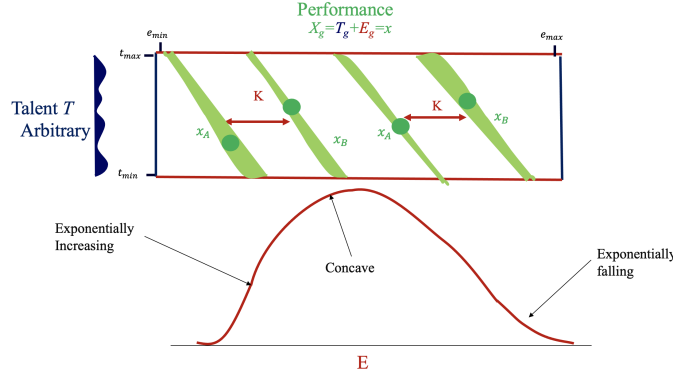


Figure 2: The range of the environment is wider than that of talent and hence both talents T_A and T_B conditioned on $X_A = X_B = x$ lie within the same narrow talent range $[t_{min}, t_{max}]$.

We explain this result as follows. In Figure 2, the range of the environment is wider than that of talent and hence both talents T_A and T_B conditioned on $X_A = X_B = x$ lie within the same narrow talent range $[t_{min}, t_{max}]$. The second diagonal green curve represents the probability space of possibilities in which B 's performance $X_B = T_B + E_B$ is conditioned to be a fixed value of x . The first one represents the probability space of possibilities in which A 's performance $X_A = T_A + E_A = T_A + E'_A + k$ is conditioned to be a fixed value x or equivalently in which $T_A + E'_A$ is conditioned to be a fixed to $x - k$. This $-k$ is what shifts this green line to the right. The left most green trapezoid corresponds to the possibilities for which $X_A = x$. Projecting the first green line down, one sees that conditioned on this, larger values of e have higher probability than lower ones. This is why this green trapezoid is higher/fatter on the bottom. Projecting this green trapezoid left to the talent range $[t_{min}, t_{max}]$, one can see that this makes the $\mathbb{E}(T_A|X_A = x)$ small. This is depicted by the green dot. In contrast, $X_B = x$ is depicted by the second green trapezoid. Because the environment distribution is flatter in this region, $\mathbb{E}(T_B|X_B = x)$ is larger. The same effect of $\mathbb{E}(T_B|X_B = x) > \mathbb{E}(T_A|X_A = x)$ is shown in the right two green trapezoids. The formal proof appears in Section D.3.

The counter-intuitive case Finally, let us highlight that the conclusion $\mathbb{E}(T_B|X_B = x) > \mathbb{E}(T_A|X_A = x)$ is not valid for all combinations of talent and environment distributions. If the shape of the environmental distribution curve E_B in the previous section was “flipped upside down”, then it would be convex. Figure 3, illustrates the case that the environment distribution E_B is convex. The figure illustrates how for every x , $\mathbb{E}(T_B|X_B = x) < \mathbb{E}(T_A|X_A = x)$. The proof is identical to that for Theorem 4 except “increases” is changed to “decreases” and each $\geq$ is changed to a $\leq$.

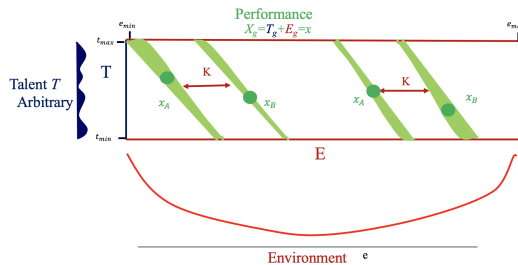


Figure 3: An Environment Distribution that is not log-concave.

Figure 4 below provides the two examples that originally motivated this work. These are very simple cases where the conclusion about expected talent is contrary to what one may expect from our additive model.

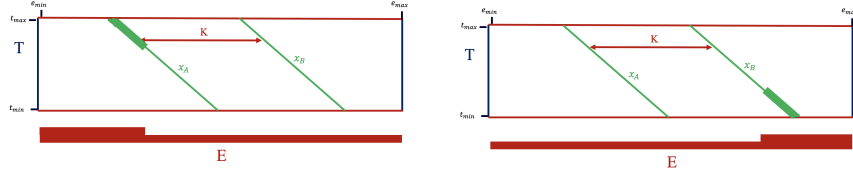


Figure 4: Though the environment distribution is linear every where, except at the one quick transition, at this point curve is not log-concave. The figure then shows how for x values that very carefully take advantage of this bump, there is a reversal of our result, namely $\mathbb{E}(T_B | X_B = x) < \mathbb{E}(T_A | X_A = x)$.

Limitations In this section we discuss the limitations of our model in detail. Our main underlying modeling assumption is that the performance score of an individual is equal to the sum of their “talent” and “environment”. This is clearly a limiting model of how these variables influence each other and how observable quantities relate to unobservables. Studying more complex interactions between hidden and observed variables is left out of our study, but an important direction for further investigation. Here our focus is on distilling out how modeling assumptions may influence conclusions, and show that this influence exists even in a seemingly simple model.

Furthermore, we here handle the wide talent case when the talent distribution is uniform or normal. We include some relaxations of this assumption in the appendix. However our results here are currently not giving a full picture of cases yet. Providing a full characterization of cases is an intriguing direction for future work.

5 Conclusions

In this work, we analyze a simple additive model, where two unobserved variables T and E result in an observed performance score X . We assume that, the “talent” component T is identically distributed for two subpopulations A and B while the “environmental” component E is not, with the A distribution assumed to be advantageous. A person’s performance level is computed as $X = T + E$ level. As a decision maker focuses on specific candidates, we condition on this performance level having a particular level, $X = x$ level. We assume that the decision preference of candidates is heavily determined by the talent component. Hence, within this conditioning, we analyze the resulting difference in talent distributions and identify a wide range of distributional properties, where the decision should be made in favor of an disadvantaged individual. Somewhat surprisingly, we also exhibit situations, where despite the observed score being the sum and the E_A distribution being an additive shift of E_B , there may be scores for which the conclusion would be reversed.

We leave it up to the reader to map the distributions we analyze to societal scenarios that would be adequately modeled by these (such as, for example, there being a large middle-class versus opportunity being shifted to extremal end cases). We acknowledge and emphasize that any inferred statements about reality of any society, depends crucially on empirical evaluations of modeling assumptions. This is beyond the scope of our work here.

What we are aiming at here, instead, is to *highlight the role of technical modeling assumptions* and, crucially, the *degree to which these technical assumptions* (and consequently any conclusions derived in the resulting model) are *reflections of underlying worldviews*. We found that this latter aspect (the connection to worldviews) is rarely made explicit in studies that analyze, purely mathematically, the effects of algorithmic fairness interventions under some statistical modelings of the sources of discrepancy between an advantaged and a disadvantaged group. Nevertheless, the conclusions are often stated in general, plain language. Our concern here is, that the degree to which such conclusions are relying on modeled beliefs as well as the technicalities of a framework of analysis chosen (statistical learning theory, game theory, dynamics etc), are not easily recovered by a reader that is less mathematically versed. We would like to advocate for making it a responsibility of authors in this research community to explicitly discuss the influence of modeled beliefs and frameworks of analysis on obtained results, rather than hiding them under mathematical obscurity.

References

- [1] Blum Avrim and Stangl Kevin. Recovering from Biased Data: Can Fairness Constraints Improve Accuracy? In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*, Leibniz International Proceedings in Informatics (LIPIcs), pages 3:1–3:20, 2020.
- [2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairml-book.org, 2019. <http://www.fairmlbook.org>.
- [3] Ted Bergstrom and Mark Bagnoli. Log-concave probability and its applications. *Economic Theory*, 26:445–469, 08 2005.
- [4] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5, 10 2016.
- [5] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM*, 63(5):82–89, 2020.
- [6] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A. Weis, Kevin Swersky, Toniann Pitassi, and Richard S. Zemel. Flexibly fair representation learning by disentanglement. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pages 1436–1445, 2019.
- [7] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012*, pages 214–226, 2012.
- [8] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM*, 64:136–143, 2021.
- [9] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS’16*, page 3323–3331. Curran Associates Inc., 2016.
- [10] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* ’19*, page 259–268. Association for Computing Machinery, 2019.
- [11] Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* ’19*, page 240–248. Association for Computing Machinery, 2019.
- [12] Michael J. Kearns, Aaron Roth, and Zhiwei Steven Wu. Meritocratic fairness for cross-population selection. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, pages 1828–1836, 2017.
- [13] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017*, pages 43:1–43:23, 2017.
- [14] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6196–6200. International Joint Conferences on Artificial Intelligence Organization, 2019.
- [15] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pages 3381–3390, 2018.
- [16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6):115:1–115:35, 2021.

- [17] Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, page 359–368. Association for Computing Machinery, 2019.
- [18] Saeed Sharifi-Malvajerdi, Michael J. Kearns, and Aaron Roth. Average individual fairness: Algorithms, generalization and experiments. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 8240–8249, 2019.
- [19] Samuel Yeom and Michael Tschantz. Discriminative but not discriminatory: A comparison of fairness definitions under different worldviews. In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, pages 33–41, 08 2018.

Checklist

1. For all authors...
 - (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [\[Yes\]](#)
 - (b) Did you describe the limitations of your work? [\[Yes\]](#) We discuss the limitations of our modeling and the scope of our conclusions within the model in detail.
 - (c) Did you discuss any potential negative societal impacts of your work? [\[No\]](#) Our work is of theoretical nature. The main goal of our work is to advocate for more transparency in technical modeling assumptions and their connections to underlying belief systems. We believe that more transparency is ultimately beneficial to readers, implied societal perceptions, and any potential applications built based on technical modeling assumptions.
 - (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [\[Yes\]](#)
2. If you are including theoretical results...
 - (a) Did you state the full set of assumptions of all theoretical results? [\[Yes\]](#)
 - (b) Did you include complete proofs of all theoretical results? [\[Yes\]](#)
3. If you ran experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [\[N/A\]](#)
 - (b) Did you specify all the training details (e.g., data splits, hyper-parameters, how they were chosen)? [\[N/A\]](#)
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [\[N/A\]](#)
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [\[N/A\]](#)
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
 - (a) If your work uses existing assets, did you cite the creators? [\[N/A\]](#)
 - (b) Did you mention the license of the assets? [\[N/A\]](#)
 - (c) Did you include any new assets either in the supplemental material or as a URL? [\[N/A\]](#)
 - (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [\[N/A\]](#)
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [\[N/A\]](#)
5. If you used crowdsourcing or conducted research with human subjects...
 - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [\[N/A\]](#)
 - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [\[N/A\]](#)

- (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Wide Uniform Talent Distributions

Consider Figure 1. The talents T_A and T_B of individuals in the two groups are random variables drawn independently according to the same arbitrary distribution. The blue shape on the left depicts this distribution rotated. Their environments E_A and E_B are formed by sampling independently from two other arbitrary distributions. The one for group g is depicted at the bottom in red. Both of these curves depicted a height coming up out of the page in the 3^{rd} dimension. In this way, the height $Pr(t, e)$ of any point in the rectangle is the product $Pr(t) \times Pr(e)$ of the corresponding probabilities. The top diagonal green curve represents the probability space of possibilities in which g 's performance $X_g = T_g + E_g$ is conditioned to be a fixed value x , namely $\bigcup_e Pr(T_g = x - e \ \& \ E_g = e)$.

For this section, as shown in the figure, the green line restricts the talent range but not the environment range.¹

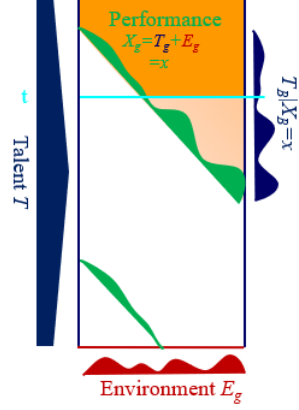


Figure 1

Instead, we focus on $X_g = x$. $Pr(X_g = x)$ is the area out of the page along the green diagonal line. $Pr(T_g \geq t \ \& \ X_g = x)$ is that that is also above the blue horizontal line. $Pr(T_g \geq t | X_g = x)$ is then the ratio of these. We struggled to compute this, without impressive results. (See the Appendix.) Instead, we focus on the talent distribution being uniform. In this case, the green curve is just a linear transformation of the red curve. This is easy to integrate.

The top right blue curve is green curve projected onto talent line, namely $Pr(T_g | X_g = x)$. Though a little care needs to be done to makes sure we are projecting the distributions correctly, this picture, then argues for the following lemma. Its proof is in the Section D.2.

Lemma 1. *Let $P_T(t)$ denote the density function of the talent distribution T and $P_{E_g}(e_g)$ for the environment distribution E_g . Condition on g 's performance score $X_g = T_g + E_g$ being some fixed value x . Then the density function of $Pr(T_g \in [t_g, t_g + \delta t] | X_g = x_g)$ is $P_{x_g, g}(t_g) = c_g P_T(t_g) P_{E_g}(x_g - t_g)$ for some constant c_g .*

We will now use this lemma to prove the main result of this section. Though we did prove a result for talent distributions that are not quite uniform, like the right blue one in the figure, the statement of the result was too complicated to include. Hence, let's again assume that the talent distribution is uniform.

The talents T_A and T_B of individuals in the two groups are random variables drawn independently according to the same uniform distribution. Their environments E_A and E_B are random variables drawn independently from two other arbitrary distributions with the only requirements being that the first is privileged over the second and both have ranges smaller than that for talent. Their performance is defined by $X_g = T_g + E_g$. Consider only the people whose performance is some fixed value $X_g = x$. Suppose this condition restricts the talent range but not the environment range. The following follows.

Let t_A be any talent level that the company is hoping to hire. Let x_A be any performance value. Let $e_A = x_A - t_A$. The company pulls all the resumes of the group A people with this level. The talent level can't be determined from the resume, but the expected fraction of these that will have the required talent is denoted $Pr(T_A \geq t_A | X_A = x_A)$.

Define r to be the ratio and k and $\bar{k}$ the environment gaps such that $Pr(E_B \leq e_A) = r \cdot Pr(E_A \leq e_A)$, $Pr(E_B \leq e_A - k) = Pr(E_A \leq e_A)$, and $\bar{k} = Exp(E_A) - Exp(E_B)$. Because group A is privileged over group B , the A environment values are likely higher, and hence r is likely greater than one and k and $\bar{k}$ are likely positive. The result is strengthened by the fact that both r and k depend on $e_A = x_A - t_A$.

One way to understand this is to sort N random people from each group according to their environment into two parallel lines. Search for the A person in the line who has environment value e_A . Let $p \in [0, 1]$ be the percentile such that he is pN^{th} in line. Then talk to the B person who is in the same place in line. Let e_B denote her environment value. Let $k = e_A - e_B$ denote the environment gap.

One definition of fairness is to keep all the requirements the same. In this case

$$Pr(T_B \geq t_A | X_B = x_A) = r \cdot Pr(T_A \geq t_A | X_A = x_A).$$

People in group B actually have a higher probability of meeting the talent threshold. Remember that $r > 1$.

One reasonable definition of fairness is to insist that the same fraction of B meets that same talent t_A requirement. One achieves this by changing the performance value considered for B 's to $x_B = x_A - k$, giving

$$\begin{aligned} Pr(T_B \geq t_A | X_B = x_A - k) &= Pr(T_A \geq t_A | X_A = x_A). \\ Exp(T_B | X_B = x_A - k) &= Exp(T_A | X_A = x_A). \end{aligned}$$

From this perspective, the affirmative action of lowering the performance requirement of the disadvantaged group makes sense.

Another reasonable definition of fairness is to take the resumes of the B people with the same performance level and shift the talent expectation to $t_B = t_A + k$ so that the probabilities of achieving it is still the same, namely

$$\begin{aligned} Pr(T_B \geq t_A + k | X_B = x_A) &= Pr(T_A \geq t_A | X_A = x_A). \\ Exp(T_B | X_B = x_A) - Exp(T_A | X_A = x_A) &= \bar{k}. \end{aligned}$$

From this perspective, if the B person looks the same on paper, then you should hire her because she is likely considerably more talented. To understand this last statement about probability, take N random people with performance x_A from each group and sort them by talent into two parallel lines. Consider the two people shaking hands a p fraction down the line. Then the talent of the B person will be k higher.

Theorem 3. *Let T_A and T_B be distributed identically, uniformly over some range. Let E_B be arbitrarily distributed in some bounded range and let E_A be identically distributed to $E_B + k$ for some $k > 0$. Let $X_A = T_A + E_A$, and $X_B = T_B + E_B$, and let $\bar{k} = \mathbb{E}(E_A) - \mathbb{E}(E_B)$. If x is so that conditioning on $X_g = x$ does not restrict the range of E_A and E_B (while it may restrict the range of T_A and T_B), we get:*

- $Pr(T_B \geq t_A | X_B = x) = r \cdot Pr(T_A \geq t_A | X_A = x)$, for some $r > 1$;
- $Pr(T_B \geq t_A | X_B = x_A - k) = Pr(T_A \geq t_A | X_A = x)$, and $Exp(T_B | X_B = x - k) = Exp(T_A | X_A = x)$;
- $Pr(T_B \geq t_A + k | X_B = x) = Pr(T_A \geq t_A | X_A = x)$, and $Exp(T_B | X_B = x) - Exp(T_A | X_A = x) = \bar{k}$.

Proof of Theorem 3: If the value of k is fixed for all environments e_A , then the two group's environments E_A and E_B are distributed according to the same distribution shifted by k , i.e. $E_A = E'_A + k$. In this case, conditioning on $X_B = x_A - k$, shifts the distribution back so that the distribution for T_A and T_B are trivially the same. Here, however, we are allowing E_A and E_B to have completely different distributions, allowing k to vary with e_A and making the result harder to prove.

Lemma 1 states that $Pr(T_g \in [t'_g, t'_g + \delta t] | X_g = x_g) = c'_g P_T(t'_g) P_{E_g}(x_g - t'_g) \delta t$. Because the talent distribution is uniform, the value $P_T(t'_g)$ is a constant, which can be blended in with c_g . This allows us to integrate for $t'_g \in [t_g, t_{max}]$ obtaining the result $Pr(T_g \geq t_g | X_g = x_g) = c_g Pr(x_g - t_{max} \leq E_g \leq x_g - t_g)$. Setting $t_g = t_{min}$ gives $Pr(T_g \geq t_{min} | X_g = x_g) = 1$. The conditions $T_g + E_g = X_g = x_g$ and $T_g \in [t_{min}, t_{max}]$ together limit the environment to $E_g \in [x_g - t_{max}, x_g - t_{min}]$. However, the lemma states that the condition $X_g = x_g$ restricts the talent range but not the environment range. $Pr(x_g - t_{max} \leq E_g \leq x_g - t_{min}) = 1$. This gives $c_g = 1$ and

$$Pr(T_g \geq t_g | X_g = x_g) = Pr(E_g \leq x_g - t_g).$$

The results follow.

$$\begin{aligned} Pr(T_B \geq t_A | X_B = x_A) &= Pr(E_B \leq x_A - t_A) = r \cdot Pr(E_A \leq x_A - t_A) = r \cdot Pr(T_A \geq t_A | X_A = x_A). \\ Pr(T_B \geq t_A | X_B = x_A - k) &= Pr(E_B \leq x_A - t_A - k) = Pr(E_A \leq x_A - t_A) = Pr(T_A \geq t_A | X_A = x_A). \\ Pr(T_B \geq t_A + k | X_B = x_A) &= Pr(E_B \leq x_A - t_A - k) = Pr(E_A \leq x_A - t_A) = Pr(T_A \geq t_A | X_A = x_A). \end{aligned}$$

We will now prove the results about expectation. Lemma 4 states that if $\forall t Pr(T_B \geq t | X_B = x_B) > Pr(T_A \geq t | X_A = x_A)$, then $Exp(T_B | X_B = x_B) > Exp(T_A | X_A = x_A)$. At first glance, one might

think that this converts these results about probabilities into the ones about expectations. However, it is much harder because the value of k might depend on e_A , which depends on t_A , which we are integrating over. Hence, we will have to work harder.

Plugging $c_g = 1$ back into the statement from Lemma 1 gives $Pr(T_g \in [t, t+\delta t] | X_g = x_g) = P_{E_g}(x_g - t)\delta t$. We can then calculate the expected value as follows. $Exp(T_g | X_g = x_g) = \int_t t \cdot Pr(T_g \in [t, t+\delta t] | X_g = x_g) = \int_t t \cdot P_{E_g}(x_g - t)\delta t = x_g \cdot \int_t P_{E_g}(x_g - t)\delta t - \int_t (x_g - t) \cdot P_{E_g}(x_g - t)\delta t = x_g \cdot 1 - Exp(E_g)$.

This gives $Exp(T_B | X_B = x_B) - Exp(T_A | X_A = x_A) = Exp(E_A) - Exp(E_B) + x_B - x_A$.

Letting $\bar{k} = Exp(E_A) - Exp(E_B)$ and $x_B = x_A - \bar{k}$, gives $Exp(T_B | X_B = x_A - \bar{k}) - Exp(T_A | X_A = x_A) = 0$.

Letting $x_A = x_B = x$, gives $Exp(T_B | X_B = x) - Exp(T_A | X_A = x) = Exp(E_A) - Exp(E_B) = \bar{k}$. ■

B Wide Environment or Non-Uniform Talent

We now cover the complement condition from Section A.

Theorem 4. Let talent distribution for T_A and T_B be arbitrary within a narrow range $[t_{min}, t_{max}]$. Let environments E'_A and E_B be independently chosen from the same arbitrary distribution. Let $E_A = E'_A + k$. Condition on the both of their performance scores $X_g = T_g + E_g$ being some fixed value x . Suppose this condition restricts the environment range but not the talent range. Groups A and B having the same conditioned talent range, means that there are no immediate reasons that $Exp(T_A | X_A = x)$ and $Exp(T_B | X_B = x)$ are different. However, if the environment distribution is log-concave, then we still have the following:

$$\begin{aligned} Pr(T_B \geq t | X_B = x_A - k) &= Pr(T_A \geq t | X_A = x_A) \\ Exp(T_B | X_B = x_A - k) &= Exp(T_A | X_A = x_A) \\ Pr(T_B \geq t | X_B = x) &\geq Pr(T_A \geq t | X_A = x) \\ Exp(T_B | X_B = x) &\geq Exp(T_A | X_A = x). \end{aligned}$$

When $E_A = E'_A + k$ and $X_B = x_A - k$, then the distribution T_A and T_B are trivially the same. Hence, the result is easy.

We explain this result as follows. In Figure 2, the range of the environment is wider than that of talent and hence both talents T_A and T_B conditioned on $X_A = X_B = x$ lie within the same narrow talent range $[t_{min}, t_{max}]$.

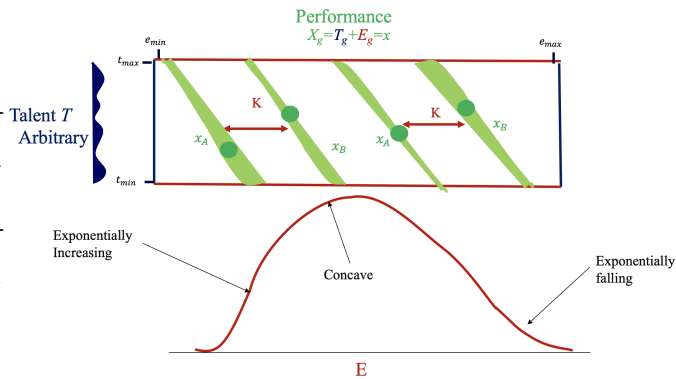


Figure 2

Having the same range, means that there are no immediate reasons that $Exp(T_A | X_A = x)$ and $Exp(T_B | X_B = x)$ are different. Here the talent distribution can be arbitrary. E'_A and E_B are independently chosen from the same *log-concave* [3] environment distributions and $E_A = E'_A + k$.

The definition of log-concave is that after you take the log of the curve, it is concave. Note when you take the log of an exponential you get linear which is just on the edge of being concave. Though the

Gaussian curve begins and ends convex, its equation is e^{-x^2} , the log of which is the concave function $-x^2$.

The second diagonal green curve represents the probability space of possibilities in which B 's performance $X_B = T_B + E_B$ is conditioned to be a fixed value of x . The first one represents the probability space of possibilities in which A 's performance $X_A = T_A + E_A = T_A + E'_A + k$ is conditioned to be a fixed value x or equivalently in which $T_A + E'_A$ is conditioned to be a fixed to $x - k$. This $-k$ is what shifts this green line to the right. The left most green trapezoid corresponds to the possibilities for which $X_A = x$. Projecting the first green line down, one sees that conditioned on this, larger values of e have higher probability than lower ones. This is why this green trapezoid is higher/fatter on the bottom. Projecting this green trapezoid left to the talent range $[t_{min}, t_{max}]$, one can see that this makes the $Exp(T_A|X_A = x)$ small. This is depicted by the green dot. In contrast, $X_B = x$ is depicted by the second green trapezoid. Because the environment distribution is flatter in this region, $Exp(T_B|X_B = x)$ is larger. The same effect of $Exp(T_B|X_B = x) > Exp(T_A|X_A = x)$ is shown in the right two green trapezoids. The formal proof appears in Section D.3.

If the red environmental distribution curve E_B was flipped upside down, then it would be convex. The rich are rich, the poor are poor, and there is no middle class. The same proof by picture, gives that for every x , $Exp(T_B|X_B = x) < Exp(T_A|X_A = x)$.

A probability distribution is *log-concave* if the logarithm of its density function is convex [3]. It includes concave, linear, and Gaussian distributions among many others. More over, it includes any density function that is "sub"-exponentially increasing in the first sub-range, then concave, then "sub"-exponentially decreasing.

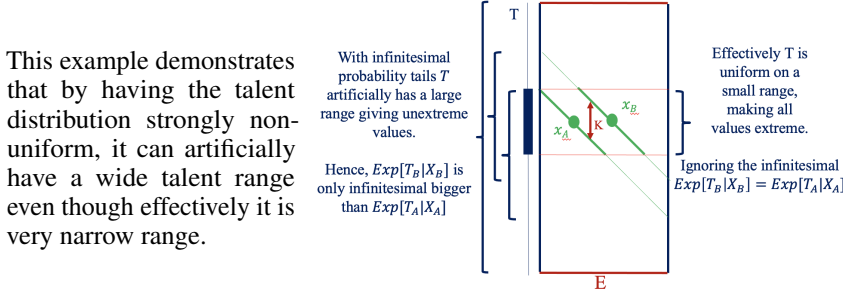


Figure 3

C Gaussian Distributions

In this section, we will redo the proof with Gaussian distributions. We have seen that $Exp(T|X = x) = x - Exp(E)$ is not always true. What is fun, however, is that we can use Bayes' rule to show that the conditional distribution is also Gaussian.

Theorem 5. Let the talent T_A and T_B be drawn from distribution $\mathcal{N}(\mu_T, \sigma_T^2)$ and group g 's environment E_g from $\mathcal{N}(\mu_{E_g}, \sigma_{E_g}^2)$ with $\mu_{E_A} - \mu_{E_B} = k$. Their conditional talent is then given by:

$$\begin{aligned} [T|X=x] &= \mathcal{N}\left(\frac{(x-\mu_E)\sigma_T^2 + \mu_T\sigma_E^2}{\sigma_T^2 + \sigma_E^2}, \frac{\sigma_T^2\sigma_E^2}{\sigma_T^2 + \sigma_E^2}\right) \\ [T_B - T_A|X_A = x_A \& X_B = x_A - k] &= \mathcal{N}\left(0, \frac{2\sigma_T^2\sigma_E^2}{\sigma_T^2 + \sigma_E^2}\right) \\ [T_B - T_A|X_A = X_B = x] &= \mathcal{N}\left(k\frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}, \frac{2\sigma_T^2\sigma_E^2}{\sigma_T^2 + \sigma_E^2}\right) \end{aligned}$$

Proof of Theorem 5: Lemma 2 uses Bayes' rule to give the first line. Then the mean and variance of the difference between two variables, is the difference between their means and the sum of their

²We could bang out these when the environment variances $\sigma_{E_A}^2$ and $\sigma_{E_B}^2$ are different, but it is too ugly to be worth it.

variances. The second line sets performance $X_B = X_A - k$ in order to zero the expected difference in talent. The third sets performance $X_B = X_A$ and computes the expected difference in talent. ■

It is worth noting the two extremes. When the talent range is much wider than that of environment, i.e. when $\sigma_E^2 \ll \sigma_T^2$, then the difference in expected talent values is maximized, namely $\text{Exp}(T_B - T_A | X_B = X_A) \rightarrow k$. In contrast, when the environment range is much wider than that of talent, i.e. when $\sigma_E^2 \gg \sigma_T^2$, then the difference in expected talent values is minimized, namely $\text{Exp}(T_B - T_A | X_B = X_A) \rightarrow 0$. It is also interesting that if the variance $\sigma_E^2 \rightarrow 0$, then E becomes a constant along with X and hence $T = X - E$ also becomes a constant. This is why $\frac{\sigma_T^2 \sigma_E^2}{\sigma_T^2 + \sigma_E^2} \rightarrow 0$. The proof is Section D.1.

D Longer Proofs

We give the longer math proofs in this section.

D.1 Proof for Gaussian Distributions

Lemma 2. *Let a person's talent be distributed as $T = \mathcal{N}(\mu_t, \sigma_t^2)$, their environment as $E_g = \mathcal{N}(\mu_E, \sigma_E^2)$, and their performance, defined by $X_g = T_g + E_g$, is conditioned to be x . It follow that the person's conditional talent is also normally distributed, namely*

$$[T|X=x] = \mathcal{N}\left(\frac{(x-\mu_E)\sigma_T^2 + \mu_T\sigma_E^2}{\sigma_T^2 + \sigma_E^2}, \frac{\sigma_T^2\sigma_E^2}{\sigma_T^2 + \sigma_E^2}\right)$$

Proof of Lemma 2: We will use Bayes' rule to show that the conditional distribution $[T|X = T + E = x]$ is also Gaussian. Specifically the rule gives

$$\text{Pr}(T \in [t, t + \delta t] | X \in [x, x + \delta x]) = \text{Pr}(X \in [x, x + \delta x] | T \in [t, t + \delta t]) \times \frac{\text{Pr}(T \in [t, t + \delta t])}{\text{Pr}(X \in [x, x + \delta x])}$$

Conditioned on T having some fixed value of $t' \in [t, t + \delta t]$, we independently choose a value for E and then compute $X = t' + E$. Hence, we get $X \in [x, x + \delta x]$ iff $E \in [x - t', x - t' + \delta x]$. Because E , T , and E are normally distributed we get

$$\begin{aligned} \text{Pr}(E \in [x - t', x - t' + \delta x]) &= \frac{1}{\sqrt{2\pi}\sigma_E} e^{-(x - \mu_E)^2 / (2\sigma_E^2)} \times \delta x \text{ where } e = x - t \\ \text{Pr}(T \in [t, t + \delta t]) &= \frac{1}{\sqrt{2\pi}\sigma_T} e^{-(t - \mu_T)^2 / (2\sigma_T^2)} \times \delta t \\ \text{Pr}(X \in [x, x + \delta x]) &= \frac{1}{\sqrt{2\pi}\sigma_X} e^{-(x - \mu_X)^2 / (2\sigma_X^2)} \times \delta x \text{ where } \mu_X = \mu_T + \mu_E, \sigma_X^2 = \sigma_T^2 + \sigma_E^2 \end{aligned}$$

Plugging these into Bayes' rule gives the exponential power to be this ugly thing. With help from maple this simplifies to

$$\frac{(x - t - \mu_E)^2}{2\sigma_E^2} + \frac{(t - \mu_T)^2}{2\sigma_T^2} - \frac{(x - \mu_T - \mu_E)^2}{2(\sigma_T^2 + \sigma_E^2)} = \frac{\left(t - \frac{(x - \mu_E)\sigma_T^2 + \mu_T\sigma_E^2}{\sigma_T^2 + \sigma_E^2}\right)^2}{2\frac{\sigma_T^2\sigma_E^2}{\sigma_T^2 + \sigma_E^2}}$$

This gives that the conditional talent distribution is

$$[T|X=x] = \mathcal{N}\left(\frac{(x-\mu_E)\sigma_T^2 + \mu_T\sigma_E^2}{\sigma_T^2 + \sigma_E^2}, \frac{\sigma_T^2\sigma_E^2}{\sigma_T^2 + \sigma_E^2}\right)$$

■

D.2 Proof Wide Uniform Talent Distributions

Lemma 1 *Let $P_T(t)$ denote the density function of the talent distribution T and $P_{E_g}(e_g)$ for the environment distribution E_g . Condition on g 's performance score $X_g = T_g + E_g$ being some fixed value x . Suppose this condition restricts the talent range but not the environment range. Then*

the density function of $Pr(T_g \in [t_g, t_g + \delta t] | X_g = x)$ is $P_{x,g}(t_g) = c_g P_T(t_g) P_{E_g}(x - t_g)$ for some constant c_g .

Proof of Lemma 1: We will drop the subscript $g \in \{A, B\}$, because the statements apply to either group. Because these random variables T and E_g are independent, we can define the cross density function $P(t, e) = P_T(t) \times P_E(e)$ so that $Pr(T \in [t, t + \delta t] \& E \in [e, e + \delta e]) = \delta t \cdot P_T(t) \times \delta e \cdot P_E(e) = \delta t \delta e \cdot P(t, e)$. We could imagine raising a third dimension coming out of the page on the $\langle T, E \rangle$ rectangle in the figure, so that its height at location $\langle t, e \rangle$ is $P(t, e)$.

The performance of a person with talent T and environment E is given by $X = T + E$. Fix a performance value x of which we will require of all group g people that we are considering for acceptance. By the statement in the lemma, this restricts the talent from $[t^{min}, t^{max}]$ to $[x - e^{max}, x - e^{min}]$.

Our goal is prove that the probability density function $P_x(t) = Pr(T \in [t, t + \delta t] | X = x) / \delta t$ of the distribution on talents t that arise under this condition is simply $P(t, x - t)$. The standard conditional probability formula $Pr(T \in [t, t + \delta t] \& X = x) / Pr(X = x)$ is awkward because the later is zero. Conditioning on our probability space amounts to narrowing our $\langle T, E \rangle$ rectangle of possibilities to the 1-dimensional line defined by $\{\langle T, E \rangle | x = T + E\}$. Lets us define the infinitesimal rectangle of possibilities $S_t = \{T \in [t, t + \delta t]\} \times \{E \in [x - t - \delta t, x - t]\}$. Within this, X is sufficiently close to x , the density function $P(t, e)$ is sufficiently constant. Hence, we will approximate $Pr(T \in [t, t + \delta t] \& X = x)$ with $Pr(S_t)$, which is $P(t, x - t) \cdot (\delta t)^2$. Lets return to the awkward fact that $Pr(X = x)$ is zero. Let's define $S_x = \bigcup_t S_t$ to be the union of all of our rectangles within which X is sufficiently close to x . Then we will replacing $Pr(X = x)$ with $Pr(S_x)$. Lets denote this probability with $p_x \cdot \delta t$. We are now able to determine the probability $Pr(T \in [t, t + \delta t] | X = x) = Pr(S_t | S_x) = Pr(S_t) / Pr(S_x) = [P(t, x - t) \cdot (\delta t)^2] / (p_x \cdot \delta t)$. Our density function $P_x(t)$ is this divided by δt . It follows that the density function of $Pr(T_g \in [t_g, t_g + \delta t] | X_g = x)$ is $P_{x,g}(t_g) = c_g P_T(t_g) P_{E_g}(x - t_g)$. Here c_g is the multiplicative constant needed to make the area under the density functions one. ■

D.3 Proof of Wide Environment or Non-Uniform Talent

Theorem 4 Let talent distribution for T_A and T_B be arbitrary within a narrow range $[t_{min}, t_{max}]$. Let environments E'_A and E_B be independently chosen from the same arbitrary distribution. Let $E_A = E'_A + k$. Condition on the both of their performance scores $X_g = T_g + E_g$ being some fixed value x . Suppose this condition restricts the environment range but not the talent range. Groups A and B having the same conditioned talent range, means that there are no immediate reasons that $Exp(T_A | X_A = x)$ and $Exp(T_B | X_B = x)$ are different. However, if the environment distribution is log-concave, then we still have the following.

$$Pr(T_B \geq t | X_B = x) \geq Pr(T_A \geq t | X_A = x)$$

$$Exp(T_B | X_B = x) \geq Exp(T_A | X_A = x)$$

Proof of Theorem 4: Let $P_T(t)$ denote the density function of the distribution with which talent T is chosen and $P_{E_g}(e)$ denote the same for environment E_g . Lets compare $P_{E_A}(e)$ and $P_{E_B}(e)$. Having $E_A = E'_A + k$ with E'_A chosen using $P_{E_B}(e)$ gives $P_{E_A}(e) \delta e = Pr(E_A \in [e, e + \delta e]) = Pr(E'_A + k \in [e, e + \delta e]) = Pr(E'_A \in [e - k, e - k + \delta e]) = P_{E_B}(e - k) \delta e$. Define $\hat{E}_g(t)$ to be $P_{E_g}(x - t)$. Note that $\hat{E}_A(t) = P_{E_A}(x - t) = P_{E_B}((x - t) - k) = P_{E_B}(x - (t + k)) = \hat{E}_B(t + k)$. Define $H(t) = \hat{E}_B(t) / \hat{E}_A(t)$. One property of E_B being log-concave is that $H(t)$ is increasing.

Lemma 1 states that the density function of $Pr(T_g \in [t, t + \delta t] | X_g = x)$ is $P_{x,g}(t) = c_g P_T(t) P_{E_g}(x - t)$. Fix some threshold $\hat{t}$. This gives

$$Pr(T_g \geq \hat{t} | X_g = x) = \frac{\int_{t \in [\hat{t}, t_{max}]} P_T(t) \hat{E}_g(t) \delta t}{\int_{t \in [t_{min}, t_{max}]} P_T(t) \hat{E}_g(t) \delta t}$$

To simplify notation, let's define a functional $\mathcal{F}$, which for any function $R(t)$ gives the fraction of the area under the curve that is to the right of the $t=\hat{t}$, namely

$$\mathcal{F}(R) = \frac{\int_{t \in [\hat{t}, t_{max}]} R(t) \delta t}{\int_{t \in [t_{min}, t_{max}]} R(t) \delta t} \quad (1)$$

Hence we can write

$$Pr(T_g \geq \hat{t} | X_g = x) = \mathcal{F}(P_T \hat{E}_g)$$

Recall that $H(t) = \hat{E}_B(t)/\hat{E}_A(t)$ is increasing. Let c be the constant $H(\hat{t})$. Hence $\forall t \in [t_{min}, \hat{t}]$ we have $H(t)/c \leq 1$ and $\forall t \in [\hat{t}, t_{max}]$ we have $H(t)/c \geq 1$. Hence, for any function $R(t)$, multiplying $R(t)$ by $H(t)/c$ decreases $R(t)$ for those t before $\hat{t}$ and increase those after. It follows that the fraction of the area under the curve $R(t)$ that is right of the $t=\hat{t}$ increases, i.e. $\mathcal{F}(R) < \mathcal{F}(H/c \cdot R)$. Similarly, $\mathcal{F}(R) = \mathcal{F}(cR)$. The result follows

$$\begin{aligned} Pr(T_A \geq \hat{t} | X_A = x) &= \mathcal{F}(P_T \hat{E}_A) = \mathcal{F}(c P_T \hat{E}_A) < \mathcal{F}(H \cdot P_T \hat{E}_A) \\ &= \mathcal{F}(\hat{E}_B / \hat{E}_A \cdot P_T \hat{E}_A) = \mathcal{F}(P_T \hat{E}_B) = Pr(T_B \geq \hat{t} | X_B = x) \end{aligned}$$

Once we have this, Lemma 4 converts this result about probabilities into one about expectations. ■

Lemma 3. *One property of E_B being log-concave is that $H(t)$ is decreasing.*

Proof of Lemma 3: We need $H(e) = \frac{E(e)}{E'(e+k)}$ to be increasing. When E is differentiable, we need $H'(e) = \frac{E'(e)E'(e+k) - E(e)E''(e+k)}{E'^2(e+k)} > 0$ or $\frac{E'(e+k)}{E'(e)} > \frac{E(e)}{E'(e)}$ or that $Q(e) = \frac{E(e)}{E'(e)}$ is increasing, namely $Q'(e) = \frac{(E'(e))^2 - E(e)E''(e)}{(E'(e))^2} = 1 - \frac{E(e)E''(e)}{(E'(e))^2} > 0$ or that $E(e)E''(e) < (E'(e))^2$. This is exactly the last property that Wikipedia lists for functions $E(e)$ that are log-convex. ■

Lemma 4. *If $\forall t$ $Pr(T_B \geq t | X_B = x_B) > Pr(T_A \geq t | X_A = x_A)$, then $Exp(T_B | X_B = x_B) > Exp(T_A | X_A = x_A)$.*

Proof of Lemma 4:

$$\begin{aligned} Exp(T_B | X_B = x) &= \int_{t \geq 0} t \cdot Pr(T_B \in [t, t+dt] | X_B = x_B) \delta t = \int_{t \geq 0} \int_{\hat{t} \in [0, t]} Pr(T_B \in [t, t+dt] | X_B = x_B) \delta \hat{t} \delta t \\ &= \int_{\hat{t} \geq 0} \int_{t \geq \hat{t}} Pr(T_B \in [t, t+dt] | X_B = x_B) \delta t \delta \hat{t} = \int_{\hat{t} \geq 0} Pr(T_B \geq \hat{t} | X_B = x_B) \delta \hat{t} > \int_{\hat{t} \geq 0} Pr(T_A \geq \hat{t} | X_A = x_A) \delta \hat{t} = Exp(T_A | X_A = x) \quad \blacksquare \end{aligned}$$

E Appendix: Almost Uniform Talent

Theorem 3 requires that the talent distribution is uniform and proves $Pr(T_B \geq t_A + k | X_B = x) = Pr(T_A \geq t_A | X_A = x)$. This section is our unimpressive results when this restriction is loosened slightly and the link between E_A and E_B tightened.

Theorem 6. *The talents T_A and T_B of individuals in the two groups are random variables drawn independently according to the same $\langle \epsilon_1, \epsilon_2 \rangle$ -uniform talent distributions. Let environments E'_A and E_B be independently chosen from the same arbitrary distribution. Let $E_A = E'_A + k$ and $t_B = t_A + k$ or more generally let $E_A = d \cdot E'_A + k$ and $t_B = t_A + k + (d-1)(x - t_B)$.³ Condition on the both of their performance scores $X_g = T_g + E_g$ being some fixed value x . Suppose this condition restricts the talent range but not the environment range. The following follows.*

$$Pr(T_A \geq t_A | X_A = x) - Pr(T_B \geq t_B | X_B = x) \leq (1 + \epsilon_1) \epsilon_2 \frac{t_B - t_A}{t_{max} - t_{min}}$$

This is small as long as the talent range is much bigger than the difference in their talents or the uniform parameters make the talent distribution almost uniform.

Defⁿ $\langle \epsilon_1, \epsilon_2 \rangle$ -uniform: The section requires that the talent distribution T is what we call $\langle \epsilon_1, \epsilon_2 \rangle$ -uniform. We know that something similar is needed because Figure 3 gives an example of a

³Having t_B on both sides is odd, but we are assuming environment $e_B = x - t_B > 0$ making it clear that $t_B \geq t_A + k$ when $d \geq 1$.

non-uniform talent distribution for which the section's result does not hold. Let the range of talent values be denoted by $[t_{min}, t_{max}]$. Because the area under its density function $P_T(t)$ is one, the average value of $P_T(t)$ is $\frac{1}{t_{max}-t_{min}}$. Our almost uniform condition requires that its minimum value at least an $(1+\epsilon_1)$ factor of this, namely $\min_t P_T(t) \geq \frac{1}{(1+\epsilon_1)(t_{max}-t_{min})}$. If its density function $P_T(t_g)$ was a line from zero to a maximum probability, then this maximum value would be $\frac{2}{t_{max}-t_{min}}$ and the slope would be $\frac{2}{(t_{max}-t_{min})^2}$. Our almost uniform condition requires that this slope is never more than $\frac{\epsilon_2}{(t_{max}-t_{min})^2}$. Define $p_\Delta = (1+\epsilon_1)\epsilon_2 \frac{t_B-t_A}{t_{max}-t_{min}}$. $\langle \epsilon_1, \epsilon_2 \rangle$ -almost uniform distributions are almost uniform and the restriction and the result are the strongest when the ϵ s are close to zero. Though we want the group's talent difference $t_B - t_A$ to be as large as possible, we are assuming that it is small relative to the size of the range $[t_{min}, t_{max}]$.

Proof of Theorem 6: Integrating the probabilities given in Lemma 1 gives $Pr(T_g \geq t_g | X_g = x) = \int_{T_g \geq t_g} c_g P_T(T_g) P_{E_g}(x - T_g) \delta T_g$. However, we do not know how to integrate this. Instead we will convert piece of it from person A to B , i.e. $P_{E_A}(e_A) \delta e = d^{-1} P_{E_B}(e_B) \delta e$, $P_T(T_A) - P_T(T_B) \leq \Delta$, and $\delta t_A = d \delta t_B$. Then we are sorry to say the constant c_g gives us a hard time.

Lets compare the density functions of the two groups environmental distributions. Having $E_A = d \cdot E_B + K$ gives $P_{E_A}(e_A) \delta e = Pr(E_A \in [e_A, e_A + \delta e]) = Pr(d \cdot E_B + K \in [e_A, e_A + \delta e]) = Pr(E_B \in [e_B, e_B + d^{-1} \cdot \delta e]) = d^{-1} P_{E_B}(e_B) \delta e$.

We also need to bound $P_T(T_A) - P_T(T_B) \leq \Delta$. If the talent distribution T is uniform, then this is zero. If its density function $P_T(t_g)$ was a line from zero to a maximum probability, then this maximum value would be $\frac{2}{t_{max}-t_{min}}$ and the slope would be $\frac{2}{(t_{max}-t_{min})^2}$. Our almost uniform condition requires that this slope is never more than $\frac{\epsilon_2}{(t_{max}-t_{min})^2}$. This gives $\Delta \leq \frac{\epsilon_2(T_B-T_A)}{(t_{max}-t_{min})^2}$.

The form $F_x(t_A) = t_B = t_A + k + (d-1)e_B$ is the best form for seeing that $T_B \geq T_A + k$. However, in order to be able to take the derivative, lets go farther, namely $t_B = t_A + k + (d-1)[d^{-1}(e_A - k)] = t_A + \frac{K}{d} + \frac{d-1}{d}e_A = t_A + \frac{K}{d} + \frac{d-1}{d}(x - t_A) = \frac{t_A + K}{d} + \frac{d-1}{d}x$. From this we get $\frac{\delta t_B}{\delta t_A} = \frac{1}{d}$.

With this we can continue to compare our group conditional probabilities.

$$\begin{aligned}
Pr(T_A \geq t_A | X_A = x) &= \int_{T_A \geq t_A} c_A P_T(T_A) P_{E_A}(x - T_A) \delta T_A \\
&\leq c_A \int_{T_B \geq t_B} [P_T(T_B) + \Delta] [d^{-1} P_{E_B}(x - T_B)] [d \delta T_B] \\
&\leq c_A \left[\int_{T_B \geq t_B} P_T(t_B) P_{E_B}(x - t_B) \delta T_B + \Delta \int_{T_B} P_{E_B}(x - T_B) \delta T_B \right] \\
&= c_A \left[c_B^{-1} \int_{T_B \geq t_B} P_{x,B}(T_B) \delta T_B + \Delta \cdot 1 \right] \\
&= c_A [c_B^{-1} Pr(T_B \geq t_B | X_B = x) + \Delta].
\end{aligned}$$

Suppose $t_A = t_B = \infty$, then the probabilities are one. This gives $1 = c_A [c_B^{-1} \cdot 1 + \Delta]$. Setting $C_A = c$ and solving gives $c_B^{-1} = \frac{1-c\Delta}{c}$ from which we get

$$P_A = c_A [c_B^{-1} P_B + \Delta] = c \left[\frac{1-c\Delta}{c} P_B + \Delta \right] = (1+c\Delta) P_B - c\Delta \leq P_B - c\Delta$$

We were disappointed that the constant c_g appears in this result, We are not really sure why it does. It can be viewed in two ways. First, it appears in the density function $P_{x,g}(t_g) = c_g P_T(t_g) P_{E_g}(x - t_g)$ to make the area under the function one. Second, $\frac{1}{c_g} = Pr(X_g = x)$. In Figure 3 because the B person is completely in the infinitesimal part of T , the probability of this occurring is infinitesimal. The makes c_B infinitely large. In this case, at least, this is not a problem because conditioned on this event happens, his expected T value is K bigger than A 's.

In order to deal with the c in $P_A \leq P_B - c\Delta$, lets get a rough bound on it. Let the range of talent values be denoted by $[t_{min}, t_{max}]$. Because the area under its density function is one, the average value of $P_T(t_g)$ is $\frac{1}{t_{max}-t_{min}}$. Our almost uniform condition requires that its minimum value at least

an $(1+\epsilon_1)$ factor of this. Using this to bound the entire area gives

$$\begin{aligned}
1 &= \int_{T_g} c_g P_T(T_g) P_{E_g}(x-T_g) \delta T_g \\
&\geq \int_{T_g} c_g \frac{1}{(1+\epsilon_1)(t_{max}-t_{min})} P_{E_g}(x-T_g) \delta T_g \\
&\geq c_g \frac{1}{(1+\epsilon_1)(t_{max}-t_{min})} \int_{T_g} P_{E_g}(x-T_g) \delta T_g \\
&\geq c_g \frac{1}{(1+\epsilon_1)(t_{max}-t_{min})} \times 1 \\
c_g &\leq (1+\epsilon_1)(t_{max}-t_{min})
\end{aligned}$$

From this we can conclude.

$$\begin{aligned}
P_A &\leq P_B - c\Delta \\
&= c_A - (1+\epsilon_1)(t_{max}-t_{min}) \times \frac{\epsilon_2(t_B-t_A)}{(t_{max}-t_{min})^2} \\
&\leq P_B - (1+\epsilon_1)\epsilon_2 \frac{t_B-t_A}{t_{max}-t_{min}}
\end{aligned}$$

When the talent distribution is uniform, $\epsilon_2=0$ giving $P_A = P_B$. ■