

A Conclusion, Limitations, and Future Work

In this work, we tackle the problem of resource-information asymmetry in AI audits and legal cases involving AI. We focus our discussion on Title VII’s “less discriminatory alternative” (LDA) requirement as a case study exhibiting how claimants often face a steep uphill battle in order to meet their evidentiary burden (as noted in previous works [11, 24, 40]). We are motivated by this issue to develop tools and methods to reduce the hurdles claimants face. Our main contributions are (1) to cast the problem of finding an LDA as one of estimating the performance-fairness Pareto frontier (PF), (2) to provide a novel technical result that, to our knowledge, provides a first closed-form expression for the performance-fairness PF, and (3) to show how this result can be used as a scaling law for performance-fairness PFs that directly addresses both the resource and information asymmetry issues posed by the LDA requirement.

Next, we identify the assumptions (both technical and conceptual) of our work, which highlight potential limitations and avenues for future work:

1. Our theoretical analysis is conducted for specific notions of fairness and performance. We justify these choices in Section 3 and Section 6, and we believe that future work tackling other definitions would be valuable. Our work also relies on disparate impact and business necessity being measurable; we do not preclude the use of multiple metrics to measure fairness or performance, and studying the more complex multi-objective optimization problem in which there are more than two metrics of interest would be compelling future work. We are unsure how to address settings in which fairness and performance are not measurable, and we welcome future work that explores the LDA requirement in such settings (e.g., when fairness is ordinal).
2. Our result applies for the DGP given in Section 4.1. A compelling direction for future work would be to determine how well our result holds, for real data and *even* for other DGPs. Although stylized, the DGP we study still allows for significant generality. We also recommend future work, both theoretical and empirical, on the form of the loss-performance PF for other DGPs.
3. Relatedly, our main proof technique that allows us to obtain an analytic expression without strong assumptions on the inputs or the model class is to use a notion of “duality” that approximates Pareto-optimal classifiers with varying fairness characteristics using Bayes optimal estimators on artificially constructed training distributions. To do so, use ζ to “tilt” the (artificial) train distribution. As one explores other DGPs, one could also explore (i) the conditions under which this duality holds and (ii) alternative ways of tilting the training distribution.
4. Assumption 4.1 appears to hold in our experiments. Given that this assumption does not directly depend on the experimental choice of data generating process (DGP), this is strong evidence supporting it. However, we suggest two directions of exploration: (i) further theoretical work to understand when this assumption may not hold, which may result in an additional term in (3) that allows Assumption 4.1 to be removed and therefore strengthens the result; and (ii) empirical investigations to identify when this assumption breaks down in practice.
5. Our result (3) is an upper bound. Although it is tight in some sense (that it holds with equality for some g), we believe a compelling direction for future work is to identify a tighter bound.
6. As mentioned in Section 4.2 the form of B that we adopt is borrowed from the literature on large language models. One could explore alternative forms of B that may be more appropriate for other model classes and sizes.

We identify several other considerations and extensions:

1. A defendant may be able to argue that, in failing to produce a specific less discriminatory alternative (LDA), this approach does not pass the necessary evidentiary standard. As noted in Footnote 4, if the court does not consider this evidence strong enough to satisfy the LDA requirement, we hope that it can be used to support the plaintiff’s requests for further discovery, including data and model access that they may be initially denied.
2. Our experimental results are limited. There are several directions for future work, including conducting experiments on real datasets, further stress testing the limitations of our theoretical results, and running experiments at greater scale.
3. Our approach requires access to some training and test data. We feel that this requirement is unavoidable, and our contribution is to significantly decrease the amount of data that the

claimant needs. However, we acknowledge that this requirement may still pose a hurdle and leave the exploration of techniques that further mitigate the data requirements to future work.

4. The empirical PF is, in a sense, a random variable that depends on the sampled training data and the randomness of the training procedure. One could explore the confidence intervals of the estimated PF that result from different training runs. Similarly, our theoretical result does not have a notion of uncertainty; future work could explore a high-probability version of our result.

B More Information on LDAs

B.1 LDAs Beyond Employment Law

The “less ___ alternative” requirement expands beyond employment to areas including housing, lending, disability, and even environmental justice. The Department of Housing and Urban Development (HUD) has a 2023 rule returning to the 2013 Fair Housing Act standard under which actuarially sound housing-insurance policies are unlawful if an equally effective, less discriminatory practice is available. The Consumer Financial Protection Bureau’s (CFPB) 2023 Fair Lending Report requires lenders to proactively search their credit-scoring models for LDAs even in the absence of litigation. A closely related duty appears in disability law; Equal Employment Opportunity Commission (EEOC) guidance under the American Disabilities Act (ADA) obliges employers to adopt any “reasonable accommodation” that meets business needs without exclusion, effectively functioning as an LDA requirement. In environmental justice, permits to proposed projects may be denied if a less harmful to human health and the environment exists (an “environmentally preferable alternative”) under National Environmental Policy Act (NEPA).

Beyond the US, comparable concepts exist. The EU employs the principle of proportionality in discrimination cases, requiring that measures be appropriate and necessary, but also that the respondent show “there is no practicable alternative.” One may argue that the LDA requirement parallels the EU’s “data minimisation” requirements under the General Data Protection Regulation (GDPR), requiring data processing to be “limited to what is necessary.” In Canada, the Meiorin test examines whether the employer has accommodated affected groups to the point of undue hardship. South Africa, drawing from Section 36 of its Constitution, applies a limitations analysis that considers whether less restrictive means could achieve policy objectives without discriminatory impacts.

B.2 Example: Challenges of finding an LDA

To illustrate the challenges involved in finding an LDA, consider the following example. Consider an AI tool that uses large language and video models to screen application packages, which contain unstructured text and video interviews. Suppose that the plaintiff successfully establishes that the tool disproportionately favors candidates from a specific demographic group using one or more chosen metrics, such as the demographic parity gap (Step 1). Suppose further that the defendant successfully shows that the tool improves their business outcomes by reducing the amount of time to screen applicants while surfacing candidates that are well suited for each job (Step 2).

At this point, the plaintiff can only win the case if they are able to meet the LDA requirement (Step 3). They meet several hurdles in their attempts:

1. The plaintiff first attempts to train their own LDA. They quickly realize that training a comparable model to the defendant’s production, state-of-the-art model requires significant compute.
2. Even if they are able to obtain the necessary compute, training a model requires access to good data. The plaintiff learns that collecting enough training data is prohibitively expensive, so the plaintiff requests it from the defendant (or the developer if the data is owned by a third party). The defendant and/or developer claims that sharing their training data is tantamount to releasing trade secrets and compromises user privacy, and the judge agrees.⁸
3. The plaintiff turns to a third option. They request access to the contested model so that they can use it a starting point to locally search for an LDA by, e.g., fine-tuning or probing its internal

⁸When the owner of the information is a third party that is not a direct party to the lawsuit (e.g., the plaintiff sues an employer, who outsources the AI model development to a third party), it can be even more difficult to obtain this information. Further discussion in Footnote ⁶

representations. The defendant and/or developer claims that this, too, violates trade secrecy, and the judge agrees.

4. Finally, the plaintiff considers establishing the existence of an LDA by characterizing the range of possible classifiers, in the same spirit as analyses in [24, 40], but discovers these approaches work well on finite-dimensional, categorical inputs but not on unstructured ones.

At this point, the plaintiff may throw in the towel, conceding that they cannot meet the LDA requirement because the burden of providing an LDA is too high, especially given the limited resources, data, and model access that they possess.

B.3 On the choice of fairness metric

In Section 3 we note that we choose demographic parity as our fairness metric, but in general this choice is context-dependent and debated upon. History indicates that demographic parity is a likely choice by courts, as selection rates (the backbone of the demographic parity metric) have appeared across multiple contexts, including the Four-Fifths Rule from 1978 to NYC’s Local Law 144 of 2021. Moreover, selection rates do not suffer from the same selection bias issues as other metrics (see Section 6 for further discussion). Similarly, we study BCE loss because it is widely used as the performance metric for classifiers and is precisely what developers/companies optimize in training.

C Intuitions for Theoretical Assumptions and Results

C.1 Intuition for Assumption 4.1

This assumption essentially implies that each Pareto-optimal model \hat{f} that belongs to model class \mathcal{F} is symmetric with respect to the DGP, which one would expect to hold for large models that serve as universal function approximators. That is, the “distance” between the loss-minimizing \hat{f} in \mathcal{F} learned on q and the Bayes optimal classifier $q(1|x, a)$ may depend on the model class \mathcal{F} and the number of samples D it is trained on, but not on q . As discussed in the previous section, this may not always hold exactly (e.g., when some q ’s are trivial to learn), but we believe it to be reasonable for deep learning models trained on sufficiently sophisticated tasks. The second condition in Assumption 4.1 implies that \hat{f} is symmetric with respect to A . This condition does *not* imply that \hat{f} fits group $A = 0$ as well as the other group $A = 1$, but that the “distance” between \hat{f} and $q(1|x, a)$ is symmetric with respect to A ; to see this, observe that it “compares” \hat{f} to the Bayes optimal classifier for q , which itself may be skewed.⁹

C.2 Key Proof Intuition for Theorem 4.2

One might try to obtain the PF by solving the constrained minimization $\hat{f}_{\underline{\Delta}} = \arg \min_{f \in \mathcal{F}} \mathcal{L}(p_{X,A,Y}, f)$ subject to $\Delta(p_{X,A,Y}, f) = \underline{\Delta}$, then computing $\mathcal{L}(p_{X,A,Y}, \hat{f}_{\underline{\Delta}})$ for all $\underline{\Delta}$. However, this problem is highly difficult to solve analytically and generally requires strong assumptions on the model class \mathcal{F} . Alternatively, one may wish to iterate over all possible \hat{f} and compute their corresponding loss and fairness gap values (akin to [10, 24]), but this approach is only feasible for finite-dimensional, categorical features where the number of possible classifiers is small [40]; this approach does not scale for large models with high-dimensional, unstructured inputs.

To address these issues, we use a technique that often motivates in-processing methods and fair representation learning [35, 21, 15]: that one can induce a fairness gap $\underline{\Delta}$ by constructing an *artificial train distribution* $q \neq p$. In other words, we can transform the constrained minimization problem $\min_{f \in \mathcal{F}} \mathcal{L}(p_{X,A,Y}, f)$ subject to $\Delta(p_{X,A,Y}, f) = \underline{\Delta}$ into an unconstrained problem $\min_{f \in \mathcal{F}} \mathcal{L}(q_{X,A,Y}, f)$, where q “tilts” p so that the solution \hat{f} to the latter problem has the desired fairness gap $\Delta(p_{X,A,Y}, \hat{f}) = \underline{\Delta}$. Note that “tilting” q does *not* have any bearing on what distribution the developer actually uses to train their model— q is an abstraction.¹⁰ To provide further intuition,

⁹The second condition can be loosened to be equality with a constant shift, but we remove this for simplicity. It will introduce additive constants throughout our proof that are ultimately incorporated into the constants and thus do not affect the main result.

¹⁰We further emphasize that, although this discussion seems to suggest that we are restricting the allowable train distribution, that is not the case. We are simply imposing an *effective* train distribution as a proof technique to analytically characterize the PF.

recall that there is a known equivalence between minimizing an objective function subject to a constraint and minimizing the Lagrangian. Our approach leverages another notion of duality where minimizing $\mathbb{E}_p[\text{loss}]$ subject to a fairness constraint is equivalent to minimizing $\mathbb{E}_q[\text{loss}]$ for some q .

This change of measure allows us to characterize a Pareto-optimal classifier \hat{f} for a given fairness gap Δ in terms of the Bayes optimal classifier $q(1 | x, a)$ for the q used to induce Δ . Then, we apply Theorem G.1 where Assumption 4.1 maps to the condition on S in Theorem G.1. In summary, our approach alleviates the two challenges described above: for (i), we use the Bayes optimal classifier on the effective train distribution q to simulate the constrained loss-minimization problem that is difficult to solve analytically; for (ii), we avoid imposing strong assumptions on the model class \mathcal{F} by using a milder model symmetry assumption.

D Detailed Procedure for Applying the Scaling Law to the LDA Requirement

Here we give details for the high-level procedure described in Section 4.2. Given training data, test data, and a compute budget as well as an estimate of N^+ and D^+ , as described in Section 3, a claimant would apply the scaling law as follows:

1. Given a small model class N^- and small training dataset D^- , empirically trace out the loss-fairness PF for N^- and D^- . To do so, train a set of models, each of which is obtained by minimizing the BCE loss on \mathcal{D}^- plus a regularizer that encourages demographic parity. Varying the weight on the regularizer induces different fairness gaps. For each trained model, compute its loss and fairness gap on the given test data. Thus, each model marks a point on the loss vs. fairness gap plane. The lower convex hull of these points forms the empirical PF. The number of trained models as well as N^- and D^- should be chosen based on one's compute budget.
2. Repeat this for different values of N^- and D^- , as permitted by one's compute budget.
3. Using these experiments, fit the constants C_1 through C_7 to each empirical PF using the following scaling law:

$$\begin{aligned} \text{loss}(\Delta) = & C_1 + C_2/(N^-)^{C_3} + C_2/(D^-)^{C_4} \\ & - C_5 \cdot C_6 \log(1 - C_7 + \Delta) - C_5 \cdot (1 - C_6) \log(C_7 - \Delta) \\ & + (C_7 - \Delta)(1 - C_7 + \Delta) \left(\frac{C_5 \cdot C_6}{2(1 - C_7 + \Delta)^2} + \frac{C_5 \cdot (1 - C_6)}{2(C_7 - \Delta)^2} \right), \end{aligned}$$

with appropriate values for N^- and D^- . Note that C_3 and C_4 are often set to 0.5 [32, 7].

4. Once estimates of C_1 through C_7 have been obtained, one can extrapolate the PF for the contested model by using the same functional form as above, except substituting N^+ and D^+ for N^- and D^- .
5. Alternatively, if one is given a specific loss-fairness pair $(\text{loss}(\Delta), \Delta)$, one can use the functional form above to determine the values of N^+ and D^+ such that $(\text{loss}(\Delta), \Delta)$ lies on the PF. This would tell the claimant how many resources the defendant needs to achieve that pair of values.¹¹

Given that there are only 7 constants to fit, one could fit the scaling law by training as few as 7 small, high-quality models. In practice, the more models, the better (see discussion of Step 1 below).

Note on implementation. We note that our approach “fails gracefully” in that it is *conservative*: if a claimant finds the contested model is δ -far from the PF, then it is *at least* that far. There are two reasons why our approach is conservative. First, the empirical PF is always conservative by definition; there may be (and almost always are) models that Pareto dominate the models one finds empirically. Second, the expression in Theorem 4.2 is an inequality; it gives an upper bound on the loss of a Pareto-optimal model for a given fairness gap.

¹¹Returning to Section 3, this analysis relies on the mild assumption that the PF improves monotonically as N and D increase. By our result (Theorem 4.2), this holds true because $B(\mathcal{F}, D)$, which is the only term that depends on the model class and dataset size, decreases as N and D increase.

Considerations for Step 1. There are multiple ways to trace out the PF empirically. For example, what we describe above is known as linear scalarization. Many works explore other methods (cf. Section 6). There are two main ways that the choice of method affects the claimant’s results: (1) some methods are better at finding *Pareto-optimal* models across a wide range of fairness gaps, and (2) some methods find them more *efficiently* without having to train many models. Thus, a good method will help the claimant get as close to the true PF as possible (which can only help the claimant increase the gap between the contested model and the PF and thus strengthen their case) while training as few models as possible (and thus using as few resources as possible). Our scaling law is plug-and-play: as the methods for finding Pareto-optimal models efficiently improve, so too does our procedure.

Considerations for Step 2. Although not strictly necessary, re-running Step 1 on different values of N^- and D^- will generally improve the estimates of C_1 through C_7 . There is a trade-off here: for a fixed compute budget, one can either run Step 1 multiple times with different (N^-, D^-) values, or run it once with a N^- and D^- that are as close to N^+ and D^+ as possible. One should plan accordingly based on one’s compute budget.

Considerations for Step 3. One may find that there are multiple sets of constants that fit the empirical PF well, as we explore in Section 5 and Appendix A. One could choose the set of constants that minimize the distance between the empirical PF and the fitted PF. Unintuitively, this might not be the appropriate choice. Recalling that the PF marks the lowest loss for each fairness gap, no observed point can lie below the true PF. Thus, one may wish to use the PF that is as close to the empirical one as possible while lying entirely below it.

Finally, we note that letting B scale with $N^{-\alpha} + D^{-\beta}$ is well supported by previous works on large models (language models, in particular). As discussed in Section 1, our work is intended for large models, as this is where the LDA requirement imposes the greatest burden. One may wish to adjust the form for B as appropriate.

E Discussion on the Practicality of the LDA Requirement

Several scholars critically examine the efficacy of disparate impact doctrine and practical application of the LDA requirement. Many observe that identifying precise and legally sufficient LDAs is difficult for plaintiffs [1, 55], and others highlight the inconsistent (and sometimes deferential) application of the business necessity standard by courts that can undermine LDA-based claims [26]. Some factors even discourage employers from seeking less discriminatory alternatives, including the risk of reverse discrimination lawsuits and legal uncertainty following *Ricci v. DeStefano* [28, 29]. Together, these critiques reveal a gap between the aspirations of disparate impact theory and the practical barriers faced by litigants. Our work takes the following perspective: while the LDA requirement remains intact, claimants increasingly need methods to demonstrate the existence of feasible alternatives.

F Helpful Lemmas

Lemma F.1. *Let μ and ν be finite probability measures of the same mass on a space \mathcal{Z} such that μ is absolutely continuous with respect to ν , i.e., $\mu \ll \nu$. Let $t(z) := \frac{d\mu}{d\nu}(z)$ denote the corresponding Radon-Nikodým derivative and $R(z) := t(z) - 1$. Let $S : \mathcal{Z} \rightarrow \mathbb{R}$ be absolutely integrable with respect to μ and ν . Then,*

$$\text{Cov}_\nu(R, S) = 0 \iff \mathbb{E}_\nu[RS] = \mathbb{E}_\nu[R]\mathbb{E}_\nu[S] \iff \mathbb{E}_\mu[S] = \mathbb{E}_\nu[S].$$

Proof. First, since μ and ν are probability measures and t is the Radon-Nikodým derivative,

$$\mathbb{E}_\nu[R] = \mathbb{E}_\nu[t - 1] = \mathbb{E}_\nu[t] - 1 = 0.$$

Furthermore, by the definition of R ,

$$\mathbb{E}_\nu[RS] = \mathbb{E}_\nu[(t - 1)S] = \mathbb{E}_\nu[tS] - \mathbb{E}_\nu[S] = \mathbb{E}_\mu[S] - \mathbb{E}_\nu[S].$$

Therefore,

$$\mathbb{E}_\mu[S] = \mathbb{E}_\nu[S] \iff \mathbb{E}_\nu[RS] = 0 = \mathbb{E}_\nu[R]\mathbb{E}_\nu[S],$$

where the last equality is a consequence of having established that $\mathbb{E}_\nu[R] = 0$. This completes the proof, with the first \iff in the lemma statement following from the definition of covariance. \square

Lemma F.2. Consider the setup in Section 3 notation in Section 4 and data generating process in Theorem 4.2. Then,

$$\text{Cov}_{p_{X|A=1}}(p(1 | X, A = 1), q(1 | X, A = 1)) \geq 0.$$

Proof. The covariance is well-defined since $p(1 | X, 1)$ and $q(1 | X, 1)$ have finite second moments by definition of p and q . To show that the expression is non-negative, we appeal to Chebyshev's Association Inequality (see, e.g., [13] Theorem 2.14), which states that if f and h are real-valued functions that are monotonic in the same direction and Z is a real-valued random variable, then $\mathbb{E}[f(Z)h(Z)] \geq \mathbb{E}[f(Z)]\mathbb{E}[h(Z)]$. Thus, $\text{Cov}(i(Z), j(Z)) = \mathbb{E}[i(Z)j(Z)] - \mathbb{E}[i(Z)]\mathbb{E}[j(Z)] \geq 0$.

To get the final result, we map Z , f , and h to quantities in our setup. Let the expectations be taken over $p(X|A = 1)$, let $\phi(X) = p(1|X, 1) = \sigma(g(X) - \zeta^p)$, and let $\psi(X) = q(1|X, 1) = \sigma(g(X) - \zeta)$.

Next, set $U = g(X)$ and denote by $\mu = p_{X|A=1} \circ g^{-1}$ the pushforward measure of $p_{X|A=1}$ by g . Introduce the deterministic functions $f(u) = \sigma(u - \zeta^p)$ and $h(u) = \sigma(u - \zeta)$ for $u \in \mathbb{R}$. Now, since the standard logistic (sigmoid) function σ is monotonically increasing in its argument, both f and h are hence nondecreasing real-valued functions. So, Chebyshev's Association Inequality applies and gives that $\text{Cov}_\mu(f(U), h(U)) \geq 0$.

Finally, because $\phi(X) = f(U)$ and $\psi(X) = h(U)$, we consequently have that

$$\text{Cov}_{p_{X|A=1}}(\phi(X), \psi(X)) = \text{Cov}_\mu(f(U), h(U)) \geq 0,$$

which proves the claim. \square

Note that the lemma above is where the inequality in Theorem 4.2 arises. It holds with equality when $g(X)$ is almost surely constant.

Lemma F.3. Suppose $Z \in (0, 1)$ is a random variable. Then,

$$\mathbb{E}[\log(1 - Z)] = \log(1 - \mathbb{E}[Z]) - \frac{\text{Var}(Z)}{2(1 - \mathbb{E}[Z])^2} + \mathcal{O}(\mathbb{E}[(Z - \mathbb{E}[Z])^3])$$

and

$$\mathbb{E}[\log(Z)] = \log(\mathbb{E}[Z]) - \frac{\text{Var}(Z)}{2\mathbb{E}[Z]^2} + \mathcal{O}(\mathbb{E}[(Z - \mathbb{E}[Z])^3])$$

Proof. Since the function $\log(1 - z)$, $z \in (0, 1)$ is (at least) three times differentiable, we can compute a second order Taylor expansion about a value $\mu \in (0, 1)$:

$$\begin{aligned} \log(1 - z) &= \log(1 - \mu) - \frac{z - \mu}{1 - \mu} - \frac{(z - \mu)^2}{2(1 - \mu)^2} - \frac{1}{6} \cdot \frac{2}{(1 - \xi)^3} (z - \mu)^3 \\ &= \log(1 - \mu) - \frac{z - \mu}{1 - \mu} - \frac{(z - \mu)^2}{2(1 - \mu)^2} - \frac{(z - \mu)^3}{3(1 - \xi)^3} \end{aligned}$$

for some ξ between μ and z . Then, plugging in the random variable Z for z and $\mathbb{E}[Z]$ for μ , and taking an expectation, we have

$$\begin{aligned} \mathbb{E}[\log(1 - Z)] &= \log(1 - \mathbb{E}[Z]) - \frac{\mathbb{E}[(Z - \mathbb{E}[Z])^2]}{2(1 - \mathbb{E}[Z])^2} - \frac{\mathbb{E}[(Z - \mathbb{E}[Z])^3]}{3(1 - \xi_1)^3} \\ &= \log(1 - \mathbb{E}[Z]) - \frac{\text{Var}[Z]}{2(1 - \mathbb{E}[Z])^2} + \mathcal{O}(\mathbb{E}[(Z - \mathbb{E}[Z])^3]) \end{aligned}$$

where the linear term disappears from the first equality because its expectation is zero. Additionally, all expectations are well-defined because bounded random variables have finite moments of all orders.

The expansion for $\mathbb{E}[\log(Z)]$ is done similarly. \square

Lemma F.4. Let $Z \in [0, 1]$, $\mu = \mathbb{E}[Z]$, $\mu_0 \in \mathbb{R}$, and $\Gamma \in \mathbb{R}$ subject to $\mu = \mu_0 - \Gamma$. Then,

$$\text{Var}[Z] \leq (\mu_0 - \Gamma)(1 - \mu_0 + \Gamma).$$

Proof. Since $Z \in [0, 1]$, $Z^2 \leq Z$, $\mathbb{E}[Z^2] \leq \mathbb{E}[Z]$. This implies that

$$\text{Var}[Z] = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 \leq \mathbb{E}[Z](1 - \mathbb{E}[Z]) = \mu(1 - \mu).$$

Plugging in $\mu = \mu_0 - \Gamma$ gives the result. \square

G Intermediate result

We provide an intermediate result that characterizes BCE loss for any DGP. We present this intermediate result for two reasons. First, it lends intuition for how our main result is obtained. Second, it does not depend on a specific definition of fairness or the DGP; it can therefore be used as a building block for future works that examine different notions of fairness and DGPs.

We use the same notation as in Section 4.1 and repeat it here for convenience. Let p and q denote joint distributions over random variables X , A , and Y . Let $p(1 | x, a)$ and $q(1 | x, a)$ denote the Bayes optimal classifiers under p and q , respectively. Let $\text{KL}(\cdot, \cdot)$ denote the Kullback-Liebler divergence. The result below simply decomposes the loss of a classifier \hat{f} on a test distribution p into three components. For reasons that will become clear in the following section, we state this result in terms of p and an auxiliary distribution q .

Lemma G.1. *Consider a classifier $\hat{f} : \mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$. Consider arbitrary joint distributions p and q over X , A , and Y , where $p_{X,A,Y} \ll q_{X,A,Y}$. Let $S(x, a, y) := y \log(q(1 | x, a) / \hat{f}(x, a)) + (1 - y) \log((1 - q(1 | x, a)) / (1 - \hat{f}(x, a)))$ be absolutely integrable. We assume $\mathbb{E}_{p_{X,A,Y}}[S(x, a, y)] = \mathbb{E}_{q_{X,A,Y}}[S(x, a, y)]$. Then,*

$$\begin{aligned} \mathcal{L}(p_{X,A,Y}, \hat{f}) &= (\mathcal{L}(q_{X,A,Y}, \hat{f}) - \mathcal{L}(q_{X,A,Y}, q_{1|X,A})) \\ &\quad + \mathbb{E}_{x,a \sim p_{X,A}} [\text{KL}(p_{Y|x,a} \parallel q_{Y|x,a})] + \mathcal{L}(p_{X,A,Y}, p_{1|X,A}). \end{aligned} \quad (4)$$

Interpretation. This result decomposes the loss of an arbitrary classifier \hat{f} on a test distribution p into three components: (1) the misspecification loss due to the choice of model class \mathcal{F} ; (2) the loss due to distribution shift between p and q ; and (3) the irreducible test loss given by the BCE loss of the Bayes optimal classifier $p_{1|X,A}$ on $p_{X,A,Y}$, where $p_{1|X,A}$ denotes the classifier that returns $p_{1|X,A}(x, a)$ on (x, a) . This result follows from a simple telescoping sum that utilizes the definitions of information theoretic terms and the stated condition involving S .

Why is this useful? Recall that our goal is to derive a scaling law of the loss-fairness PF. That is, our goal is to produce a closed-form expression of loss, which is the left-hand side of (4), in terms of the fairness gap of \hat{f} on p . Theorem G.1 gets us part of the way there; the next step is to write the right-hand side of (4) in terms of the chosen fairness gap for a given DGP.

Understanding the condition on \hat{f} . We briefly comment on the condition $\mathbb{E}_{p_{X,A,Y}}[S(x, a, y)] = \mathbb{E}_{q_{X,A,Y}}[S(x, a, y)]$, which is a condition on \hat{f} . We observe that S is the log likelihood ratio between Bernoulli models $q(1 | x, a)$ and $\hat{f}(x, a)$, such that $\mathbb{E}_{q_{X,A,Y}}[S(x, a, y)]$ is the expected conditional KL divergence between them. In the next section, Assumption 4.1 implicitly asks that this condition hold across all q 's that we consider, which implies that $\mathbb{E}_{q_{X,A,Y}}[S(x, a, y)]$, i.e., that the “distance” (as given by an expected log likelihood ratio) between \hat{f} and the Bayes optimal $q(1 | x, a)$ is constant.

When would we expect this to hold true and is it restrictive? One might expect this to hold true if \hat{f} is the model that results from training on q . If \hat{f} is the result of training on q , then the condition asks that the training procedure and model class result in an \hat{f} that replicates the patterns in its train distribution equally well, regardless of train distribution. In practice, we do not expect this condition to hold exactly; the training procedure and model class may fit some q 's better than others, e.g., if there exists a q that is trivial to learn. However, we believe this is a reasonable condition for large models trained on sufficiently sophisticated tasks because deep learning models are designed to serve as universal function approximators across different distributions.

Importantly, we emphasize that q is *not* the true train distribution. As discussed in Section 4.1, q is an artificial, auxiliary distribution that gives rise to our main proof technique.

H Proof of Lemma G.1

Recall that we denote the binary cross-entropy (BCE) loss ℓ of a classifier f with respect to a sample (x, a, y) by

$$\ell(f, (x, a, y)) = -(y \log(f(x, a)) + (1 - y) \log(1 - f(x, a))).$$

Note that f can be a soft classifier that returns values in $[0, 1]$. Further recall that we denote the expected BCE of f with respect to a distribution $p_{X,A,Y}$ by

$$\mathcal{L}(p_{X,A,Y}, f) = -\mathbb{E}_{x,a,y \sim p_{X,A,Y}} [y \log(f(x, a)) + (1 - y) \log(1 - f(x, a))],$$

and this notation $\mathcal{L}(\cdot, \cdot)$ is used analogously for other distributions and classifiers.

Proof. We split the proof into four steps.

Step 1: Decomposition. We can decompose the BCE loss of a classifier $\hat{f} \in \mathcal{F}$ on a distribution $p_{X,A,Y}$ as follows

$$H(p_{X,A,Y}, \hat{f}) = \underbrace{(H(p_{X,A,Y}, \hat{f}) - H(q_{XAY}, \hat{f}))}_{\text{loss of } \hat{f} \text{ due to distribution shift}} \quad (5)$$

$$+ \underbrace{(H(q_{XAY}, \hat{f}) - H(q_{XAY}, q_{1|X,A}))}_{\text{train loss of } \hat{f} \text{ relative to Bayes optimal}} \quad (6)$$

$$+ \underbrace{(H(q_{XAY}, q_{1|X,A}) - H(p_{X,A,Y}, q_{1|X,A}))}_{\text{loss of Bayes optimal } q_{1|X,A} \text{ due to distribution shift}} \quad (7)$$

$$+ \underbrace{(H(p_{X,A,Y}, q_{1|X,A}) - H(p_{X,A,Y}, p_{1|X,A}))}_{\text{difference in test loss of Bayes optimal classifiers}} \quad (8)$$

$$+ \underbrace{H(p_{X,A,Y}, p_{1|X,A})}_{\text{irreducible test loss}} \quad (9)$$

by simply adding and subtracting terms, where we slightly abuse notation to let $q_{1|X,A}$ denote a (soft) classifier where the prediction for (x, a) is given by $q_{Y|X,A}(1|x, a)$.

As written below each term, (5) can be viewed as the loss of \hat{f} due to distribution shift between the train distribution $q_{X,A,Y}$ and test distribution $p_{X,A,Y}$; (6) is the difference in loss between \hat{f} on the train distribution and the Bayes optimal classifier (also relative to the train distribution), which can be viewed as the loss due to the choice of model family \mathcal{F} ; (7) also captures loss due to distribution shift, but for the Bayes optimal classifier of $q_{1|X,A}$; (8) gives difference in test loss between the Bayes optimal classifier that is optimal with respect to the train distribution and one that is optimal with respect to the test distribution, which can be viewed in some sense as the irreducible generalization loss; and (9) gives the irreducible test loss of the Bayes optimal classifier $p_{1|X,A}$ (that is optimal with respect to the test distribution).

Step 2: Expanding (5) and (7). Focusing on just two of the terms above, we have

$$\begin{aligned} (5) + (7) &= -\mathbb{E}_{x,a,y \sim p_{X,A,Y}} [y \log \hat{f}(x, a) + (1 - y) \log(1 - \hat{f}(x, a))] \\ &\quad + \mathbb{E}_{x,a,y \sim q_{X,A,Y}} [y \log \hat{f}(x, a) + (1 - y) \log(1 - \hat{f}(x, a))] \\ &\quad - \mathbb{E}_{x,a,y \sim q_{X,A,Y}} [y \log q_{Y|X,A}(1|x, a) + (1 - y) \log(1 - q_{Y|X,A}(1|x, a))] \\ &\quad + \mathbb{E}_{x,a,y \sim p_{X,A,Y}} [y \log q_{Y|X,A}(1|x, a) + (1 - y) \log(1 - q_{Y|X,A}(1|x, a))] \\ &= - \int_{x,a,y} (p_{X,A,Y}(x, a, y) - q_{X,A,Y}(x, a, y)) \\ &\quad \left(y \log \hat{f}(x, a) + (1 - y) \log(1 - \hat{f}(x, a)) \right) dx da dy \\ &\quad + \int_{x,a,y} (p_{X,A,Y}(x, a, y) - q_{X,A,Y}(x, a, y)) \\ &\quad \left(y \log q_{Y|X,A}(1|x, a) + (1 - y) \log(1 - q_{Y|X,A}(1|x, a)) \right) dx da dy \end{aligned}$$

$$\begin{aligned}
& \left(y \log q_{Y|X,A}(1|x, a) + (1-y) \log(1 - q_{Y|X,A}(1|x, a)) \right) dx da dy \\
&= \int_{x,a,y} (p_{X,A,Y}(x, a, y) - q_{X,A,Y}(x, a, y)) \\
&\quad \left(y \log \left(\frac{q_{Y|X,A}(1|x, a)}{\hat{f}(x, a)} \right) + (1-y) \log \left(\frac{1 - q_{Y|X,A}(1|x, a)}{1 - \hat{f}(x, a)} \right) \right) dx da dy \\
&= \int_{x,a,y} q_{X,A,Y}(x, a, y) (w(x, a, y) - 1) \\
&\quad \left(y \log \left(\frac{q_{Y|X,A}(1|x, a)}{\hat{f}(x, a)} \right) + (1-y) \log \left(\frac{1 - q_{Y|X,A}(1|x, a)}{1 - \hat{f}(x, a)} \right) \right) dx da dy \\
&= \mathbb{E}_{q_{X,A,Y}} \left[(w(x, a, y) - 1) \left(y \log \left(\frac{q_{Y|X,A}(1|x, a)}{\hat{f}(x, a)} \right) \right. \right. \\
&\quad \left. \left. + (1-y) \log \left(\frac{1 - q_{Y|X,A}(1|x, a)}{1 - \hat{f}(x, a)} \right) \right) \right],
\end{aligned}$$

where $w(x, a, y) = p_{X,A,Y}(x, a, y)/q_{X,A,Y}(x, a, y)$. Let

$$R := w(x, a, y) - 1 \quad \text{and} \quad S := y \log \left(\frac{q_{Y|X,A}(1|x, a)}{\hat{f}(x, a)} \right) + (1-y) \log \left(\frac{1 - q_{Y|X,A}(1|x, a)}{1 - \hat{f}(x, a)} \right).$$

Then, by invoking Theorem F.1—which applies because S is absolutely integrable under the condition given in the lemma statement and because $p_{X,A,Y} \ll q_{X,A,Y}$ —we have that

$$\mathbb{E}_{q_{X,A,Y}}[RS] = \mathbb{E}_{q_{X,A,Y}}[R] \mathbb{E}_{q_{X,A,Y}}[S] \iff \mathbb{E}_{q_{X,A,Y}}[S] = \mathbb{E}_{p_{X,A,Y}}[S].$$

By the condition given in the lemma statement, $\mathbb{E}_{q_{X,A,Y}}[S] = \mathbb{E}_{p_{X,A,Y}}[S]$ and hence we obtain

$$(5) + (7) = \mathbb{E}_{q_{X,A,Y}}[RS] = \mathbb{E}_{q_{X,A,Y}}[R] \mathbb{E}_{q_{X,A,Y}}[S] = 0.$$

where the last equality is because $\mathbb{E}_{q_{X,A,Y}}[R] = 0$, as shown in the proof of Theorem F.1

Step 3: Expanding (8). The term (8) can be written as

$$- \int_{x,a,y} p_{X,A,Y}(x, a, y) \left[y \log \frac{q_{Y|X,A}(1|x, a)}{p_{Y|X,A}(1|x, a)} + (1-y) \log \frac{q_{Y|X,A}(0|x, a)}{p_{Y|X,A}(0|x, a)} \right] dx da dy.$$

Substituting for possible values of $y \in \{0, 1\}$ gives

$$\begin{aligned}
& - \int_{x,a} p_{X,A}(x, a) p_{Y|X,A}(0|x, a) \left[\log \frac{q_{Y|X,A}(0|x, a)}{p_{Y|X,A}(0|x, a)} \right] dx da \\
& - \int_{x,a} p_{X,A}(x, a) p_{Y|X,A}(1|x, a) \left[\log \frac{q_{Y|X,A}(1|x, a)}{p_{Y|X,A}(1|x, a)} \right] dx da \\
&= - \int_{x,a} p_{X,A}(x, a) (1 - p_{Y|X,A}(1|x, a)) \left[\log \frac{q_{Y|X,A}(0|x, a)}{p_{Y|X,A}(0|x, a)} \right] dx da \\
& - \int_{x,a} p_{X,A}(x, a) p_{Y|X,A}(1|x, a) \left[\log \frac{q_{Y|X,A}(1|x, a)}{p_{Y|X,A}(1|x, a)} \right] dx da \\
&= - \int_{x,a} p_{X,A}(x, a) \left[(1 - p_{Y|X,A}(1|x, a)) \left[\log \frac{1 - q_{Y|X,A}(1|x, a)}{1 - p_{Y|X,A}(1|x, a)} \right] \right. \\
&\quad \left. + p_{Y|X,A}(1|x, a) \left[\log \frac{q_{Y|X,A}(1|x, a)}{p_{Y|X,A}(1|x, a)} \right] \right] \\
&= \mathbb{E}_{x,a \sim p_{X,A}} [\text{KL}(p_{Y|X,A}(1|x, a) \parallel q_{Y|X,A}(1|x, a))],
\end{aligned}$$

where we use $\text{KL}(r \parallel s)$ to denote the KL divergence between the Bernoulli(r) and Bernoulli(s) distributions.

Step 4: Putting it together. Combining Steps 1-3,

$$\begin{aligned}
& H(p_{X,A,Y}, \hat{f}) \\
&= \textcircled{5} + \textcircled{6} + \textcircled{7} + \textcircled{8} + \textcircled{9} \\
&= \textcircled{6} + \textcircled{8} + \textcircled{9} \\
&= H(q_{XAY}, \hat{f}) - H(q_{XAY}, q_{1|X,A}) \\
&\quad + \mathbb{E}_{x,a \sim p_{X,A}} [\text{KL}(p_{Y|x,a} \parallel q_{Y|x,a})] + H(p_{X,A,Y}, p_{1|X,A}),
\end{aligned}$$

which concludes the proof. \square

I Proof of Theorem 4.2

Proof. In this proof, we use Theorem G.1 to express the BCE loss in terms of the train distribution parameter ζ , then likewise express the demographic parity gap in terms of ζ . Combining these representations allows us to express the BCE loss directly in terms of the demographic parity gap. We note that, by definition of our data-generating process, $p_{X,A,Y} \ll q_{X,A,Y}^\zeta$ for all ζ .

Step 1: Simplifying BCE loss. Recall from Theorem G.1 that

$$\begin{aligned}
H(p_{X,A,Y}, \hat{f}^\zeta) &= (H(q_{XAY}^\zeta, \hat{f}^\zeta) - H(q_{XAY}^\zeta, q_{1|X,A}^\zeta)) \\
&\quad + \mathbb{E}_{x,a \sim p_{X,A}} [\text{KL}(p_{Y|x,a} \parallel q_{Y|x,a}^\zeta)] + H(p_{X,A,Y}, p_{1|X,A}).
\end{aligned}$$

We begin by characterizing each component of this expression.

First, by Assumption 4.1, $H(q_{XAY}^\zeta, \hat{f}^\zeta) - H(q_{XAY}^\zeta, q_{1|X,A}^\zeta)$ can be written as a constant $c_1(\mathcal{F}, D)$ that depends only on the model class \mathcal{F} and the amount of training data D .

Second, we note that $H(p_{X,A,Y}, p_{1|X,A})$ can also be treated as a constant $c_2(p)$ that depends only on the test distribution p and not on the model class \mathcal{F} or the train distribution q^ζ . This gives

$$H(p_{X,A,Y}, \hat{f}^\zeta) = c_1(\mathcal{F}, D) + \mathbb{E}_{x,a \sim p_{X,A}} [\text{KL}(p_{Y|x,a} \parallel q_{Y|x,a}^\zeta)] + c_2(p), \quad (10)$$

where

$$\begin{aligned}
c_1(\mathcal{F}, D) &:= H(q_{XAY}^\zeta, \hat{f}^\zeta) - H(q_{XAY}^\zeta, q_{1|X,A}^\zeta), \\
c_2(p) &:= H(p_{X,A,Y}, p_{1|X,A}).
\end{aligned}$$

Thus, it remains only to characterize the expected KL divergence term and write it in terms of the demographic parity gap. That will be the goal of the following steps.

Step 2: Rewriting the KL divergence term.

$$\begin{aligned}
& \mathbb{E}_{x,a \sim p_{X,A}} [\text{KL}(p_{Y|x,a} \parallel q_{Y|x,a}^\zeta)] \\
&= - \int_{x,a} p_{X,A}(x, a) \left[(1 - p_{Y|X,A}(1|x, a)) \log \frac{1 - q_{Y|X,A}^\zeta(1|x, a)}{1 - p_{Y|X,A}(1|x, a)} \right. \\
&\quad \left. + p_{Y|X,A}(1|x, a) \log \frac{q_{Y|X,A}^\zeta(1|x, a)}{p_{Y|X,A}(1|x, a)} \right] \\
&= - \mathbb{E}_{x,a \sim p_{X,A}} [(1 - p_{Y|X,A}(1|x, a)) \log(1 - q_{Y|X,A}^\zeta(1|x, a)) \\
&\quad + p_{Y|X,A}(1|x, a) \log q_{Y|X,A}^\zeta(1|x, a)] + c_3(p), \quad (11)
\end{aligned}$$

where

$$c_3(p) := \int_{x,a} p_{X,A}(x, a) [(1 - p_{Y|X,A}(1|x, a)) \log(1 - p_{Y|X,A}(1|x, a))$$

$$+ p_{Y|X,A}(1|x, a) \log(p_{Y|X,A}(1|x, a)) \Big],$$

is a constant that depends only on the test distribution p . Recalling from our data generating process given in Section 4 that $p(A = 1) = \pi$, (11) becomes

$$\begin{aligned} & \mathbb{E}_{x, a \sim p_{X,A}} \left[\text{KL} \left(p_{Y|x,a} \parallel q_{Y|x,a}^\zeta \right) \right] \\ &= -\mathbb{E}_{x, a \sim p_{X,A}} \left[(1 - p_{Y|X,A}(1|x, a)) \log(1 - q_{Y|X,A}^\zeta(1|x, a)) \right. \\ &\quad \left. + p_{Y|X,A}(1|x, a) \log q_{Y|X,A}^\zeta(1|x, a) \right] + c_3(p) \\ &= \pi \left(-\mathbb{E}_{x \sim p_{X|A}} \left[(1 - p_{Y|X,A}(1|x, A)) \log(1 - q_{Y|X,A}^\zeta(1|x, A)) \right. \right. \\ &\quad \left. \left. + p_{Y|X,A}(1|x, A) \log q_{Y|X,A}^\zeta(1|x, A) \mid A = 1 \right] \right) + c_4(p), \quad (12) \end{aligned}$$

where the last line follows from the fact that, when $A = 0$, the q distribution no longer depends on ζ (cf. our data generating process for Y given in Section 4) such that

$$\begin{aligned} c_4(p) &:= c_3(p) + (1 - \pi) \left(-\mathbb{E}_{x \sim p_{X|A}} \left[(1 - p_{Y|X,A}(1|x, A)) \log(1 - q_{Y|X,A}^\zeta(1|x, A)) \right. \right. \\ &\quad \left. \left. + p_{Y|X,A}(1|x, A) \log q_{Y|X,A}^\zeta(1|x, A) \mid A = 0 \right] \right). \end{aligned}$$

Step 3: Upper bounding KL divergence term. By the definition of covariance,

$$\begin{aligned} & \mathbb{E}_{x \sim p_{X|A}} \left[(1 - p_{Y|X,A}(1|x, A)) \log(1 - q_{Y|X,A}^\zeta(1|x, A)) \mid A = 1 \right] \\ &= \mathbb{E}_{x \sim p_{X|A}} \left[(1 - p_{Y|X,A}(1|x, A)) \mid A = 1 \right] \mathbb{E}_{x \sim p_{X|A}} \left[\log(1 - q_{Y|X,A}^\zeta(1|x, A)) \mid A = 1 \right] \\ &\quad + \text{Cov}_{x \sim p_{X|A}} \left((1 - p_{Y|X,A}(1|x, A)), \log(1 - q_{Y|X,A}^\zeta(1|x, A)) \mid A = 1 \right). \end{aligned}$$

By Lemma F.2 and our data-generating process as given in the theorem statement and Section 4, the covariance term is non-negative. Therefore,

$$\begin{aligned} & \mathbb{E}_{x \sim p_{X|A}} \left[(1 - p_{Y|X,A}(1|x, A)) \log(1 - q_{Y|X,A}^\zeta(1|x, A)) \mid A = 1 \right] \\ &\geq \mathbb{E}_{x \sim p_{X|A}} \left[(1 - p_{Y|X,A}(1|x, A)) \mid A = 1 \right] \mathbb{E}_{x \sim p_{X|A}} \left[\log(1 - q_{Y|X,A}^\zeta(1|x, A)) \mid A = 1 \right]. \end{aligned} \quad (13)$$

One can apply analogous reasoning to show, again, by Theorem F.2 that

$$\begin{aligned} & \mathbb{E}_{x \sim p_{X|A}} \left[p_{Y|X,A}(1|x, A) \log q_{Y|X,A}^\zeta(1|x, A) \mid A = 1 \right] \\ &\geq \mathbb{E}_{x \sim p_{X|A}} \left[p_{Y|X,A}(1|x, A) \mid A = 1 \right] \mathbb{E}_{x \sim p_{X|A}} \left[\log q_{Y|X,A}^\zeta(1|x, A) \mid A = 1 \right]. \end{aligned}$$

(Note that the two applications of Theorem F.2 above are where the inequality in Theorem 4.2 arises. Therefore, if one wishes to remove the inequality or provide a high-probability version of our result that holds with equality, one should address Theorem F.2.) Therefore, from (12),

$$\begin{aligned} & \mathbb{E}_{x, a \sim p_{X,A}} \left[\text{KL} \left(p_{Y|x,a} \parallel q_{Y|x,a}^\zeta \right) \right] \\ &\leq -\pi \left(\mathbb{E}_{x \sim p_{X|A}} \left[(1 - p_{Y|X,A}(1|x, A)) \mid A = 1 \right] \mathbb{E}_{x \sim p_{X|A}} \left[\log(1 - q_{Y|X,A}^\zeta(1|x, A)) \mid A = 1 \right] \right. \\ &\quad \left. + \mathbb{E}_{x \sim p_{X|A}} \left[p_{Y|X,A}(1|x, A) \mid A = 1 \right] \mathbb{E}_{x \sim p_{X|A}} \left[\log q_{Y|X,A}^\zeta(1|x, A) \mid A = 1 \right] \right) + c_4(p) \\ &= -\pi(1 - c_5(p)) \mathbb{E}_{x \sim p_{X|A}} \left[\log(1 - q_{Y|X,A}^\zeta(1|x, A)) \mid A = 1 \right] \\ &\quad - \pi c_5(p) \mathbb{E}_{x \sim p_{X|A}} \left[\log q_{Y|X,A}^\zeta(1|x, A) \mid A = 1 \right] + c_4(p), \end{aligned}$$

where $c_5(p) := \mathbb{E}_{x \sim p_{X|A}} [p_{Y|X,A}(1|x, A) \mid A = 1]$.

Step 4: Moving the expectation inside the log. Then, we apply Theorem F.3, using $q_{Y|X,A}^\zeta(1|x, 1)$ as the Z in the statement of the lemma, and noting that it is a random variable that takes values between 0 and 1, as required by the lemma. For ease of notation, let

$$\begin{aligned}\bar{q}_1 &:= \mathbb{E}_{x \sim p_{X|A}} \left[q_{Y|X,A}^\zeta(1|x, A) \mid A = 1 \right], \\ \bar{q}_0 &:= \mathbb{E}_{x \sim p_{X|A}} \left[q_{Y|X,A}^\zeta(1|x, A) \mid A = 0 \right], \\ V &:= \text{Var}_{x \sim p_{X|A}} \left[q_{Y|X,A}^\zeta(1|x, A) \mid A = 1 \right].\end{aligned}\tag{14}$$

Then, Theorem F.3 gives

$$\begin{aligned}\mathbb{E}_{x,a \sim p_{X,A}} \left[\text{KL} \left(p_{Y|x,a} \parallel q_{Y|x,a}^\zeta \right) \right] \\ \leq -\pi(1 - c_5(p)) \left(\log(1 - \bar{q}_1) - \frac{V}{2(1 - \bar{q}_1)^2} \right) - \pi c_5(p) \left(\log(\bar{q}_1) - \frac{V}{2\bar{q}_1^2} \right) \\ + \mathcal{O} \left(\mathbb{E}_{x \sim p_{X|A}} \left[(q_{Y|X,A}^\zeta(1|x, A) - \bar{q}_1)^3 \mid A = 1 \right] \right) + c_4(p).\end{aligned}\tag{15}$$

Now that the KL divergence term is upper bounded, we proceed to characterize the demographic parity gap, and then substitute it into the KL divergence term.

Step 5: Characterizing demographic parity gap. The definition of demographic parity gap is

$$\Delta(p_{X,A,Y}, \hat{f}^\zeta) := \left| \mathbb{E}_{x \sim p_{X|A}} [\hat{f}^\zeta(x, A) \mid A = 1] - \mathbb{E}_{x \sim p_{X|A}} [\hat{f}^\zeta(x, A) \mid A = 0] \right|.$$

We can rewrite it as

$$\begin{aligned}\mathbb{E}_{x \sim p_{X|A}} [\hat{f}^\zeta(x, A) \mid A = 1] - \mathbb{E}_{x \sim p_{X|A}} [\hat{f}^\zeta(x, A) \mid A = 0] \\ = \mathbb{E}_{x \sim p_{X|A}} [q_{Y|X,A}^\zeta(1|x, A) \mid A = 1] - \mathbb{E}_{x \sim p_{X|A}} [q_{Y|X,A}^\zeta(1|x, A) \mid A = 0] \\ + (\mathbb{E}_{x \sim p_{X|A}} [\hat{f}^\zeta(x, A) \mid A = 1] - \mathbb{E}_{x \sim p_{X|A}} [q_{Y|X,A}^\zeta(1|x, A) \mid A = 1])\end{aligned}\tag{16}$$

$$+ (\mathbb{E}_{x \sim p_{X|A}} [q_{Y|X,A}^\zeta(1|x, A) \mid A = 0] - \mathbb{E}_{x \sim p_{X|A}} [\hat{f}^\zeta(x, A) \mid A = 0]).\tag{17}$$

By Assumption 4.1

$$\begin{aligned}\mathbb{E}_{x \sim p_{X|A}} [\hat{f}^\zeta(x, A) \mid A = 1] - \mathbb{E}_{x \sim p_{X|A}} [\hat{f}^\zeta(x, A) \mid A = 0] \\ = \mathbb{E}_{x \sim p_{X|A}} [q_{Y|X,A}^\zeta(1|x, A) \mid A = 1] - \mathbb{E}_{x \sim p_{X|A}} [q_{Y|X,A}^\zeta(1|x, A) \mid A = 0].\end{aligned}$$

By our data generating process given in Theorem 4.2 in which A mediates the effect of ζ and larger values of ζ result in lower positive classification rates,

$$\mathbb{E}_{x \sim p_{X|A}} [q_{Y|X,A}^\zeta(1|x, A) \mid A = 0] \geq \mathbb{E}_{x \sim p_{X|A}} [q_{Y|X,A}^\zeta(1|x, A) \mid A = 1].$$

for $\zeta \geq 0$. Therefore,

$$\Delta(p_{X,A,Y}, \hat{f}^\zeta) = \bar{q}_0 - \bar{q}_1.\tag{18}$$

Note that if we relaxed the second condition in Assumption 4.1 to allow for an additive constant, this would allow us to complete our analysis but with more bookkeeping, as the expression above would have two cases to ensure $\Delta \geq 0$.

Step 6: Combining and rewriting BCE loss in terms of $\Delta(p_{X,A,Y}, \hat{f}^\zeta)$. We can now substitute this expression for $\Delta(p_{X,A,Y}, \hat{f}^\zeta)$ from (18) into the BCE loss. From (15), we have

$$\begin{aligned}\mathbb{E}_{x,a \sim p_{X,A}} \left[\text{KL} \left(p_{Y|x,a} \parallel q_{Y|x,a}^\zeta \right) \right] \\ \leq -\pi(1 - c_5(p)) \log(1 - \bar{q}_0 + \Delta(p_{X,A,Y}, \hat{f}^\zeta)) + \frac{\pi(1 - c_5(p))V}{2(1 - \bar{q}_0 + \Delta(p_{X,A,Y}, \hat{f}^\zeta))^2}\end{aligned}$$

$$\begin{aligned}
& -\pi c_5(p) \log(\bar{q}_0 - \Delta(p_{X,A,Y}, \hat{f}^\zeta)) + \frac{\pi c_5(p)V}{2(\bar{q}_0 - \Delta(p_{X,A,Y}, \hat{f}^\zeta))^2} \\
& + c_4(p) + \mathcal{O}\left(\mathbb{E}_{x \sim p_{X|A}} \left[(q_{Y|X,A}^\zeta(1|x,A) - \bar{q}_1)^3 \mid A=1 \right]\right).
\end{aligned}$$

We use Theorem F.4 with $Z = q_{Y|X,A}^\zeta(1|x,A)$, $\mu = \bar{q}_1$, $\mu_0 = \bar{q}_0$, $\Gamma = \Delta = \bar{q}_0 - \bar{q}_1$ by (I8), and that the expectations in Theorem F.4 are taken with respect to $p_{X|A=1}$ to get

$$V \leq (c_6(p) - \Delta)(1 - c_6(p) + \Delta(p_{X,A,Y}, \hat{f}^\zeta)),$$

where $c_6(p) = \bar{q}_0$. Therefore,

$$\begin{aligned}
& \mathbb{E}_{x,a \sim p_{X,A}} \left[\text{KL} \left(p_{Y|x,a} \parallel q_{Y|x,a}^\zeta \right) \right] \\
& \leq -\pi(1 - c_5(p)) \log(1 - c_6(p) + \Delta(p_{X,A,Y}, \hat{f}^\zeta)) \\
& \quad + \frac{\pi(1 - c_5(p))(c_6(p) - \Delta(p_{X,A,Y}, \hat{f}^\zeta))(1 - c_6(p) + \Delta(p_{X,A,Y}, \hat{f}^\zeta))}{2(1 - c_6(p) + \Delta(p_{X,A,Y}, \hat{f}^\zeta))^2} \\
& \quad - \pi c_5(p) \log(c_6(p) - \Delta(p_{X,A,Y}, \hat{f}^\zeta)) \\
& \quad + \frac{\pi c_5(p)(c_6(p) - \Delta(p_{X,A,Y}, \hat{f}^\zeta))(1 - c_6(p) + \Delta(p_{X,A,Y}, \hat{f}^\zeta))}{2(c_6(p) - \Delta(p_{X,A,Y}, \hat{f}^\zeta))^2} \\
& \quad + c_4(p) + \mathcal{O}\left(\mathbb{E}_{x \sim p_{X|A}} \left[(q_{Y|X,A}^\zeta(1|x,A) - \bar{q}_1)^3 \mid A=1 \right]\right), \tag{19}
\end{aligned}$$

where we recall all constants:

$$\begin{aligned}
c_2(p) &= H(p_{X,A,Y}, p_{1|X,A}), \\
c_3(p) &= \mathbb{E}_{x,a \sim p_{X,A}} \left[(1 - p_{Y|X,A}(1|x,a)) \log(1 - p_{Y|X,A}(1|x,a)) \right. \\
& \quad \left. + p_{Y|X,A}(1|x,a) \log(p_{Y|X,A}(1|x,a)) \right], \\
c_4(p) &= c_3(p) - (1 - \pi) \mathbb{E}_{x \sim p_{X|A}} \left[(1 - p_{Y|X,A}(1|x,A)) \log(1 - q_{Y|X,A}^\zeta(1|x,A)) \right. \\
& \quad \left. + p_{Y|X,A}(1|x,A) \log q_{Y|X,A}^\zeta(1|x,A) \mid A=0 \right], \\
c_5(p) &= \mathbb{E}_{x \sim p_{X|A}} [p_{Y|X,A}(1|x,A) \mid A=1] \\
c_6(p) &= \mathbb{E}_{x \sim p_{X|A}} [q_{Y|X,A}^\zeta(1|x,A) \mid A=0]
\end{aligned}$$

Putting it all together, combining (I0) and (I9),

$$\begin{aligned}
H(p_{X,A,Y}, \hat{f}^\zeta) & \leq B(\mathcal{F}, D) - c \cdot c' \log(1 - c'' + \underline{\Delta}) - c \cdot (1 - c') \log(c'' - \underline{\Delta}) \\
& \quad + (c'' - \underline{\Delta})(1 - c'' + \underline{\Delta}) \left(\frac{c \cdot c'}{2(1 - c'' + \underline{\Delta})^2} + \frac{c \cdot (1 - c')}{2(c'' - \underline{\Delta})^2} \right) \\
& \quad + \mathcal{O}\left(\mathbb{E}_{x \sim p_{X|A}} \left[(q_{Y|X,A}^\zeta(1|x,A) - \bar{q}_1)^3 \mid A=1 \right]\right),
\end{aligned}$$

where

$$\begin{aligned}
B(\mathcal{F}, D) &:= c_1(\mathcal{F}, D) + c_2(p) + c_4(p) \\
&= c_1(\mathcal{F}, D) + H(p_{X,A,Y}, p_{1|X,A}) \\
& \quad - (1 - \pi) \mathbb{E}_{x \sim p_{X|A}} \left[(1 - p_{Y|X,A}(1|x,A)) \log(1 - q_{Y|X,A}^\zeta(1|x,A)) \right. \\
& \quad \left. + p_{Y|X,A}(1|x,A) \log q_{Y|X,A}^\zeta(1|x,A) \mid A=0 \right] \\
& \quad + \mathbb{E}_{x,a \sim p_{X,A}} \left[(1 - p_{Y|X,A}(1|x,a)) \log(1 - p_{Y|X,A}(1|x,a)) \right. \\
& \quad \left. + p_{Y|X,A}(1|x,a) \log(p_{Y|X,A}(1|x,a)) \right]
\end{aligned}$$

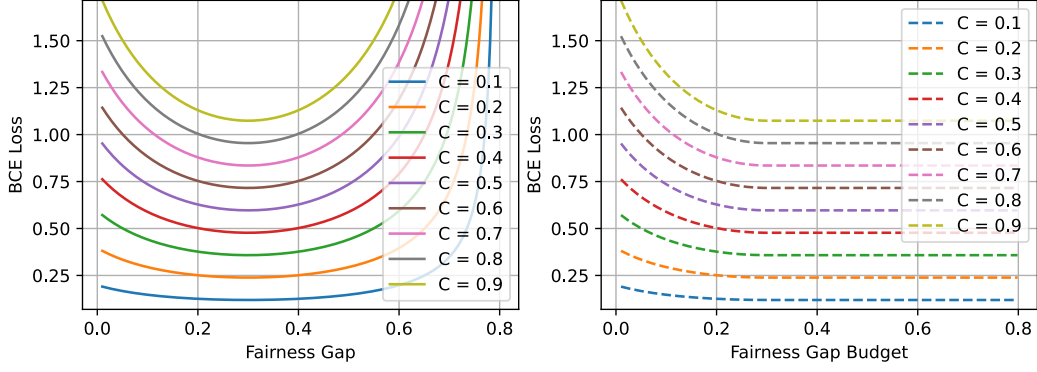


Figure 4: Simulations showing the shape of the closed-form expression for the Pareto frontier given in Theorem 4.2 for fixed values of $c' = 0.5$ and $c'' = 0.8$ while varying c . The left plot shows the precise closed-form, which is the lowest achievable loss among classifiers that have *exactly* the fairness gap value on the x-axis. The right plot shows lowest achievable loss among classifiers that satisfy the fairness gap *budget* on the x-axis.

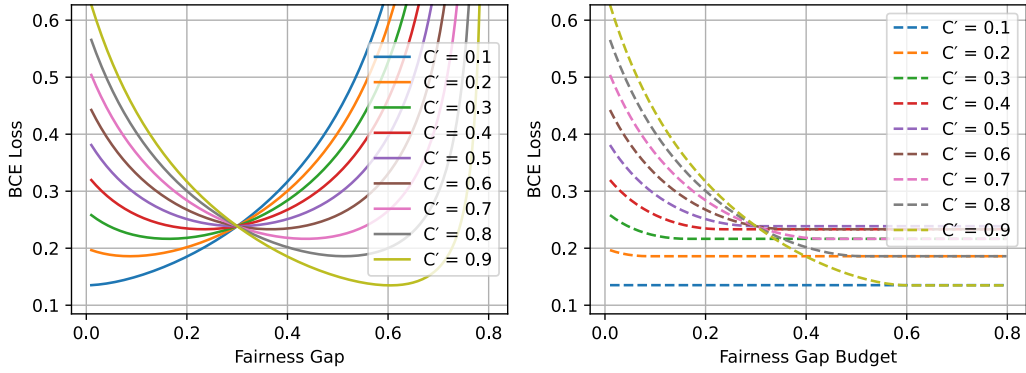


Figure 5: Simulations showing the shape of the closed-form expression for the Pareto frontier given in Theorem 4.2 for fixed values of $c = 0.2$ and $c'' = 0.8$ while varying c' . The left plot shows the precise closed-form, which is the lowest achievable loss among classifiers that have *exactly* the fairness gap value on the x-axis. The right plot shows lowest achievable loss among classifiers that satisfy the fairness gap *budget* on the x-axis.

$$\begin{aligned}
 c &:= \pi \\
 c' &:= 1 - \mathbb{E}_{x \sim p_{X|A}}[p_{Y|X,A}(1 | x, A) | A = 1] \in [0, 1] \\
 c'' &:= \mathbb{E}_{x \sim p_{X|A}}[q^\zeta(1|x, A)|A = 0] \in [0, 1]
 \end{aligned}$$

$c, c', c'' \geq 0$. Note that the constants may not correspond exactly to the stated quantities above, e.g., due to the note below (18). \square

J Additional Experimental Details and Results

J.1 Simulations

In this section, we provide further simulations.

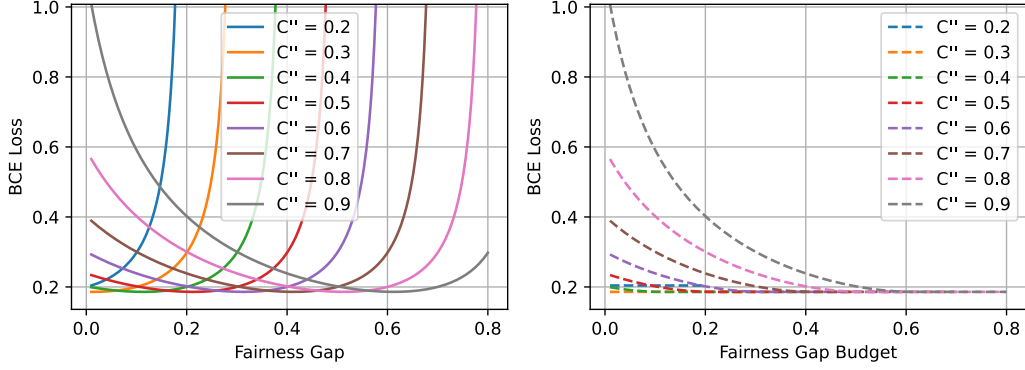


Figure 6: Simulations showing the shape of the closed-form expression for the Pareto frontier given in Theorem 4.2 for fixed values of $c = 0.2$ and $c' = 0.8$ while varying c'' . The left plot shows the precise closed-form, which is the lowest achievable loss among classifiers that have *exactly* the fairness gap value on the x-axis. The right plot shows lowest achievable loss among classifiers that satisfy the fairness gap *budget* on the x-axis.

We begin by visualizing the behavior of the (3). Recall that (3) expresses loss as a function of the fairness gap Δ , where there are four constants: B , c , c' , and c'' . In Figures 4 to 6, we show how the *shape* of (3) changes for different values of each constant, while keeping the other constants fixed and letting $B, \varepsilon = 0$. We provide further figures in Appendix J.

On the left, we plot the theoretically predicted Pareto frontier (PF) that is traced out by finding the lowest-loss classifier for an *exact* fairness gap Δ . Specifically, we evaluate the expression in Theorem 4.2, disregarding ε , over a range of Δ values (typically from 0 to 1). As one might expect, the loss decreases as Δ increases, then increases when a model is forced to have large fairness gap Δ .

To visualize the lowest loss achievable at or below a given fairness gap *budget* (rather than a fairness gap), we repeat the same simulation as in the left plot, then take the minimum loss value to the left of and including the current x-value (i.e., of the current fairness gap). Therefore, compared to the plot on the left, the plot on the right is, by definition, non-increasing. Intuitively, this gives the typical form of the PF by plotting the lowest achievable loss among classifiers with fairness gaps at or below the fairness gap budget given by the x-value.

We observe that increasing c shifts the PF upwards while also increasing its curvature; changing c' tilts the PF; and increasing c'' moves the PF to the right.

J.2 Synthetic experiments

In this section, we provide additional details on the setup for the synthetic experiments in Figure 3 as well as additional results below.

J.3 Model Architecture and Training

We used Pytorch to train our models. To test the scaling law, we trained models under 4 different MLP configurations, each with two hidden layers and ReLU activations, with a sigmoid output for binary classification:

- [80, 80] hidden units (8080 total parameters)
- [160, 160] hidden units (28960 total parameters)
- [320, 320] hidden units (109120 total parameters)
- [640, 640] hidden units (423040 total parameters)

For each architecture and λ value, the data is split into training (65%), validation (15%), and test (20%) sets. The models are trained for 30 epochs using the Adam optimizer with a learning rate of

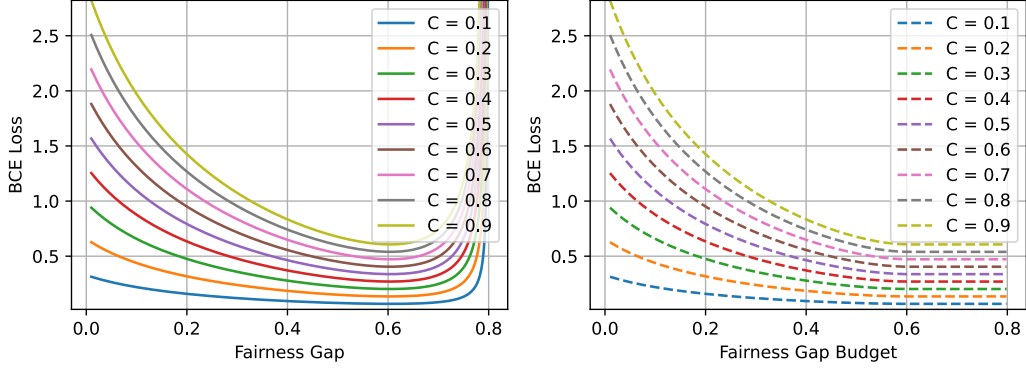


Figure 7: Simulations showing the shape of the closed-form expression for the Pareto frontier given in Theorem 4.2 for fixed values of $c'' = C_7 = 0.9$ and $c'' = 0.8$ while varying c . The left plot shows the precise closed-form, which is the lowest achievable loss among classifiers that have *exactly* the fairness gap value on the x-axis. The right plot shows lowest achievable loss among classifiers that satisfy the fairness gap *budget* on the x-axis.

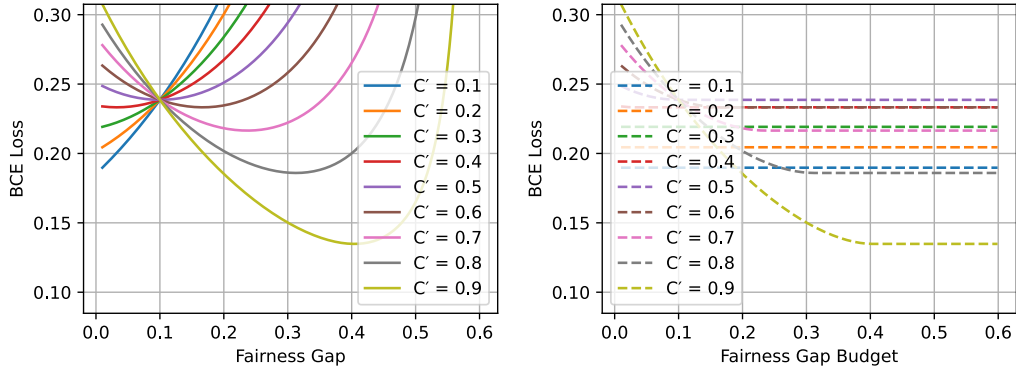


Figure 8: Simulations showing the shape of the closed-form expression for the Pareto frontier given in Theorem 4.2 for fixed values of $c = 0.2$ and $c'' = 0.6$ while varying c' . The left plot shows the precise closed-form, which is the lowest achievable loss among classifiers that have *exactly* the fairness gap value on the x-axis. The right plot shows lowest achievable loss among classifiers that satisfy the fairness gap *budget* on the x-axis.

0.001 and a batch size of 256, selecting the model with the lowest validation loss over the training epochs. We repeat each configuration over 3 random seeds and use all trials to find the Pareto frontier, resulting in 300 models per model size and 1200 models total.

J.3.1 Additional Results

We test on two combinations of ζ and π values.

We show results below. Interestingly, we found that the typical scaling law $N^{-0.5}$ did not necessarily fit our results perfectly, so we indicate the exponent used in each figure caption.

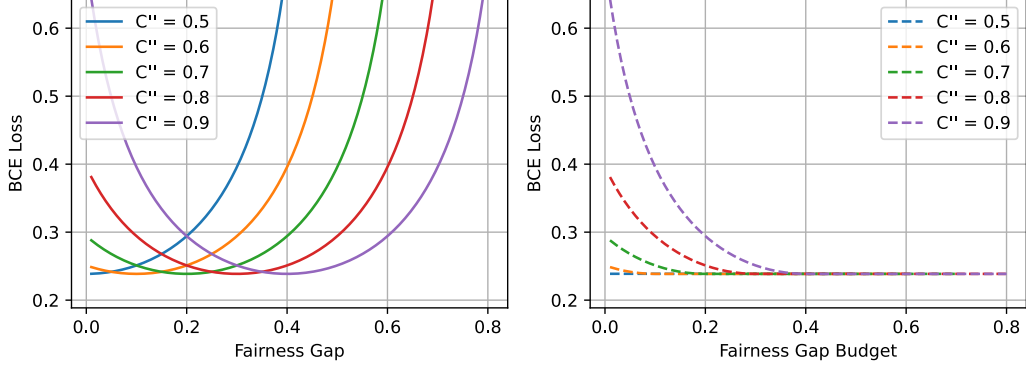


Figure 9: Simulations showing the shape of the closed-form expression for the Pareto frontier given in Theorem 4.2 for fixed values of $c = 0.2$ and $c'' = C_7 = 0.5$ while varying c'' . The left plot shows the precise closed-form, which is the lowest achievable loss among classifiers that have *exactly* the fairness gap value on the x-axis. The right plot shows lowest achievable loss among classifiers that satisfy the fairness gap *budget* on the x-axis.

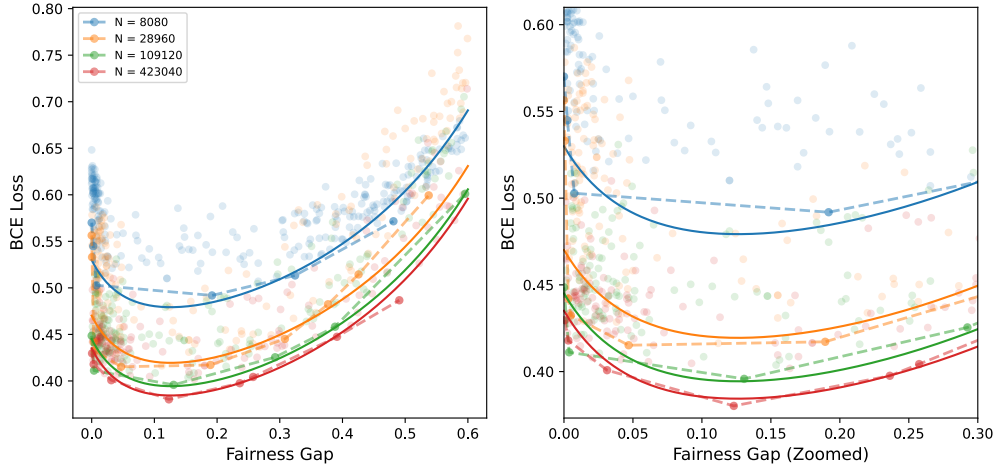


Figure 10: Pareto frontier for different model sizes under $(\pi = 0.2)$ and bias strength $\zeta = 0.5$. Each point corresponds to a trained model with a different fairness regularization weight λ . The dashed lines show the empirical Pareto frontier, created by finding the lower convex hull of all the points. The solid lines show fitted curves to the points on the Pareto frontier using Theorem 4.2 with $c = C_5 = 0.16$, $c' = C_6 = 0.11$, $c'' = C_7 = 0.92$, with bias $C_1 = -0.285$, $C_2 = 55$, $C_3 = 0.7$, and $C_4 = 0.5$. The left panel shows the frontier across the range of Δ , while the right panel zooms in on $\Delta \in [0, 0.3]$. The fitted curve mimics the empirical data well though it is imperfect. We found that there were many possible fits, depending on the precise choice. We show another possible fit in Figure 11 below.

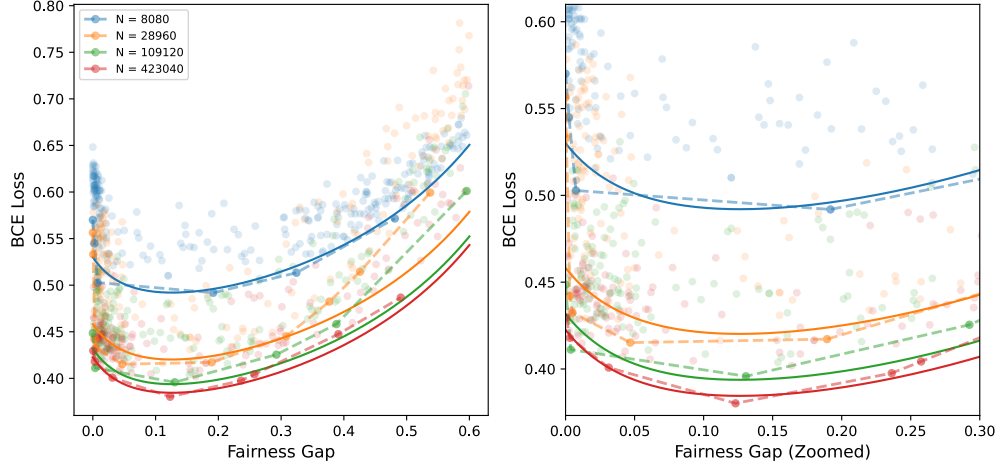


Figure 11: Pareto frontier for different model sizes under ($\pi = 0.2$) and bias strength $\zeta = 0.5$. Each point corresponds to a trained model with a different fairness regularization weight λ . The dashed lines show the empirical Pareto frontier, created by finding the lower convex hull of all the points. The solid lines show fitted curves to the points on the Pareto frontier using Theorem 4.2 with $c = C_5 = 0.12$, $c' = C_6 = 0.11$, $c'' = C_7 = 0.92$, with bias $C_1 = -1.205$, $C_2 = 150$, $C_3 = 0.8$, and $C_4 = 0.5$. The left panel shows the frontier across the range of Δ , while the right panel zooms in on $\Delta \in [0, 0.3]$. The fitted curve is intentionally chosen to lower bound the larger models. That is, it fits to the small model curve, then uses it to extrapolate the larger model curve; we do so to exhibit one possible way to fit the curves since our training procedure was not optimized for the larger models and, as such, our empirical Pareto frontier is likely not optimal. We show another possible fit in Figure 13 below.

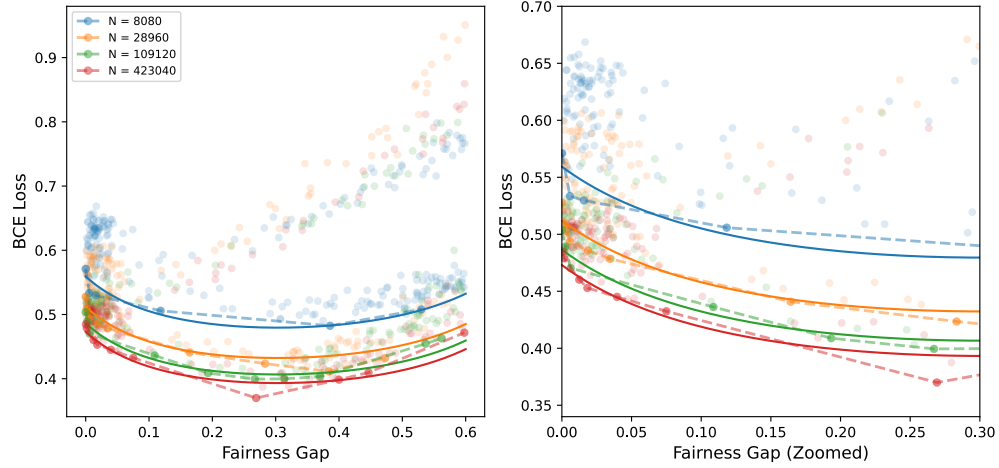


Figure 12: Pareto frontier for different model sizes under ($\pi = 0.4$) and bias strength $\zeta = 2$. Each point corresponds to a trained model with a different fairness regularization weight λ . The dashed lines show the empirical Pareto frontier, created by finding the lower convex hull of all the points. The solid lines show fitted curves to the points on the Pareto frontier using Theorem 4.2 with $c = C_5 = 0.08$, $c' = C_6 = 0.43$, $c'' = C_7 = 0.85$, with bias $C_1 = 0.195$, $C_2 = 9$, $C_3 = 0.5$, and $C_4 = 0.5$. The left panel shows the frontier across the range of Δ , while the right panel zooms in on $\Delta \in [0, 0.3]$. The fitted curve mimics the empirical data well though it is imperfect. We found that there were many possible fits, depending on the precise choice. We show another possible fit in Figure 13 below.

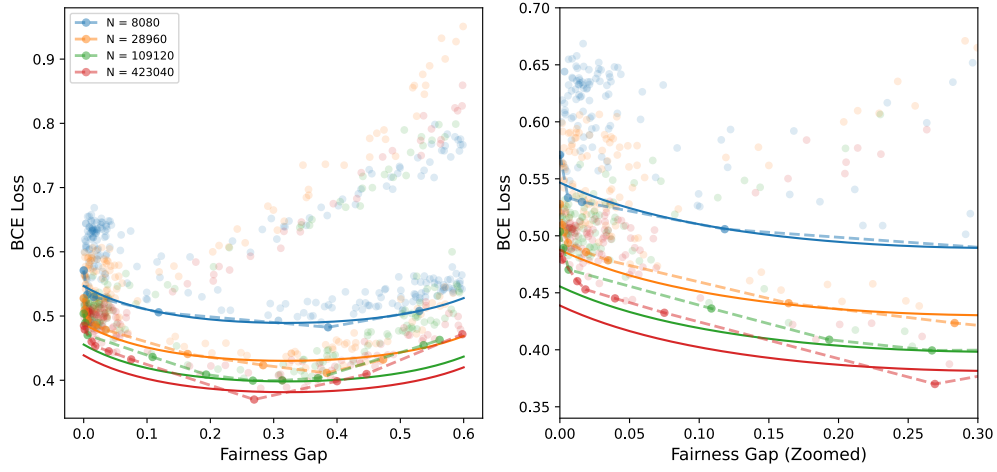


Figure 13: Pareto frontier for different model sizes under ($\pi = 0.4$) and bias strength $\zeta = 2$. Each point corresponds to a trained model with a different fairness regularization weight λ . The dashed lines show the empirical Pareto frontier, created by finding the lower convex hull of all the points. The solid lines show fitted curves to the points on the Pareto frontier using Theorem 4.2 with $c = C_5 = 0.06$, $c' = C_6 = C_7 = 0.5$, $c'' = C_7 = 0.82$, with bias $C_1 = 0.18$, $C_2 = 11.25$, $C_3 = 0.8$, and $C_4 = 0.5$. The left panel shows the frontier across the range of Δ , while the right panel zooms in on $\Delta \in [0, 0.3]$. The fitted curve is intentionally chosen to lower bound the larger models. That is, it fits to the small model curve, then uses it to extrapolate the larger model curve; we do so to exhibit one possible way to fit the curves since our training procedure was not optimized for the larger models and, as such, our empirical Pareto frontier is likely not optimal. We show another possible fit in Figure 12