Does Privacy Always Harm Fairness? Data-Dependent Trade-offs via Chernoff Information Neural Estimation

Arjun Nichani [1] **Hsiang Hsu** [2] **Chun-Fu (Richard) Chen** [2] **Haewon Jeong** [1 3]

Abstract

Fairness and privacy are two central pillars of trustworthy machine learning, yet their interaction remains poorly understood. To better characterize this relationship, we introduce *Chernoff Difference* (CD), an information-theoretic notion of *data fairness* based on Chernoff Information, and its noisy variant for analyzing fairness, privacy, and accuracy jointly. To make this framework usable beyond closed-form settings, we develop **CINE** (Chernoff Information Neural Estimator), to our knowledge the first neural estimator of Chernoff Information from samples of unknown distributions. Using CD and CINE together, we identify three qualitatively distinct regimes of the privacy–fairness interaction: privacy can hurt fairness, have little effect, or improve it ("free fairness"). We prove the existence of all three regimes analytically in the Gaussian case and observe them empirically on mixtures of Gaussian and real datasets. Overall, our results provide the first principled, data-dependent characterization of when privacy and fairness may align or conflict.

1. Introduction

As machine learning (ML) systems are increasingly deployed in high-stakes settings, ensuring their trustworthiness has become critical. Among the many facets of trustworthy ML, fairness and privacy stand out as core concerns, yet both are frequently violated in practice. Unfairness has been documented across domains such as facial recognition (Garvie and Frankle, 2016), text-to-image generation (Friedrich et al., 2023; Bianchi et al., 2023), and predictive tasks like recidivism (Barocas and Selbst, 2016; Chouldechova, 2016) and loan approvals (Das et al., 2021), prompting a broad literature on fair learning (Hardt et al., 2016; Agarwal et al., 2018; Alghamdi et al., 2022). Privacy risks, highlighted by attacks that recover sensitive training data (Shokri et al., 2017; Carlini et al., 2021) and amplified by the rise of large-scale AI models (Gomstyn and Jonker, 2024), have likewise spurred extensive research on differentially private (DP) learning (Abadi et al., 2016; Papernot et al., 2018; Kang et al., 2020).

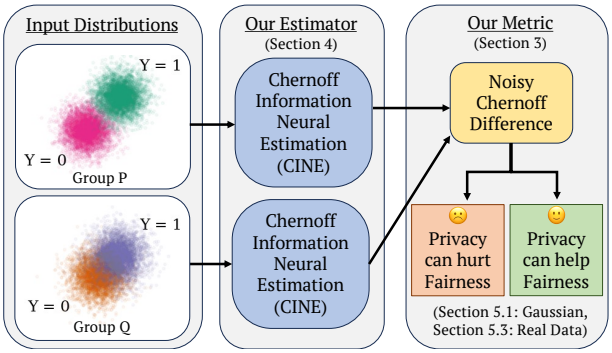


Figure 1. **Overview of our framework.** CINE estimates Chernoff Information from class-conditional distributions to analyze fairness–privacy trade-offs via the Chernoff Difference.

Yet while fairness and privacy have been deeply studied in isolation, their interaction remains less understood. Can models satisfy both simultaneously, or does enforcing one inevitably compromise the other? Recent work points to both incompatibility, exposing trade-offs between the two (Bagdasaryan et al., 2019; Chang and Shokri, 2021; Tran et al., 2021b; Sanyal et al., 2022), and potential compatibility under specific assumptions (Cummings et al., 2019; Mangold et al., 2023; Shaham et al., 2025). In this work, we show that the answer can lie in the structure of the data itself. We make the following contributions:

- We introduce **Chernoff Difference (CD)**, a metric grounded in Chernoff Information, which serves as a notion of *data fairness* by characterizing how groups differ in their classification separability. We then propose **Noisy Chernoff Difference**, an extension of CD that cap-

[1]University of California, Santa Barbara [2]JP Morgan Chase Global Technology Applied Research [3]Flatiron Institute. Correspondence to: Arjun Nichani <anichani@ucsb.edu>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

tures the fairness–privacy–accuracy triad by modeling how group separability evolves under privacy-preserving noise.

- We develop the **Chernoff Information Neural Estimator** (CINE, Section 4), to our knowledge the first algorithm to estimate Chernoff Information directly from samples of unknown distributions, leveraging advances in density ratio estimation via neural networks (Choi et al., 2022), and prove its consistency.
- Using CD and CINE, we identify **three qualitatively distinct regimes** of the fairness–privacy interaction: privacy can hurt fairness, have little effect, or improve it, effectively granting *"fairness for free"* when distributional conditions align. We prove existence of all three regimes analytically for isotropic Gaussians, and observe them empirically on Gaussian mixtures, MNIST (LeCun et al., 1998) representations, and the UCI Adult dataset (Becker and Kohavi, 1996).

To the best of our knowledge, our findings provide the first concrete evidence that whether fairness and privacy are in conflict or in harmony is not a universal law, but a property of the underlying data distribution, thereby building the core foundation of a data-dependent understanding of the complex relationships between fairness, privacy, and accuracy.

Related Work. The relationship between fairness and privacy has attracted increasing attention in recent years. Bagdasaryan et al. (2019) show that DP disproportionately harms the accuracy of underrepresented groups, while Chang and Shokri (2021) demonstrate that minority groups are more vulnerable to membership inference attacks. Tran et al. (2021a) demonstrate that privacy can amplify disparities for groups closer to the decision boundary. Cummings et al. (2019) prove the impossibility of achieving *exact* fairness and DP with nontrivial accuracy, but propose an algorithm that satisfies DP with approximate fairness under relaxed constraints. Sanyal et al. (2022) show that under long-tailed, imbalanced distributions and strict privacy, enforcing both fairness and privacy worsens accuracy, whereas Mangold et al. (2023) establish that fairness disparities between private and non-private models decay inversely with sample size. Several works also develop algorithms jointly ensuring fairness, privacy, and accuracy (Lyu et al., 2020; Lowy et al., 2023; Ghoukasian and Asoodeh, 2024; Jagielski et al., 2019). For a survey, see Fioretto et al. (2022) and Shaham et al. (2025). Despite these advances, there is still a limited understanding of how the input data distribution shapes this relationship, as highlighted in a recent review paper (Yao and Juarez, 2025). Existing work often treats the fairness–privacy dynamic as a universal trade-off, without explicitly characterizing how the underlying data distribution shapes when privacy may help or harm fairness. Our work addresses this gap by providing a distribution-level perspective on how privacy reshapes group separability.

2. Problem Setting and Background

Consider a binary classification setting in which our data is defined by continuous, non-sensitive features X, sensitive attributes $S = \{0, 1\}$, and labels $Y = \{0, 1\}$. Using these parameters, we can define our data distribution as a mixture of conditional distributions $P_0(x) = \Pr(X|S = 0, Y = 0)$, $P_1(x) = \Pr(X|S = 0, Y = 1)$, $Q_0(x) = \Pr(X|S = 1, Y = 0)$, and $Q_1(x) = \Pr(X|S = 1, Y = 1)$. From this definition, we can refer to our groups as P, to be the group where $S = 0$ and Q, to be the group where $S = 1$. Further, we make an equal prior assumption ($P(Y = 0|S = s) = P(Y = 1|S = s)$ for all s) for simplicity. However, we provide an extension to the unequal prior setting in Appendix B.2. For our model, we consider a split classifier setting where we train a different classifier for P and Q. Under the assumption of infinite model complexity (and sufficient group information), any model will converge towards the split classifier setting (Wang et al., 2021).

Goal. We aim to examine how the underlying data distributions (P_0, P_1, Q_0, and Q_1) shape the trade-off between fairness, privacy, and accuracy. Rather than positing a universal relationship between privacy and fairness, we provide a distribution-level characterization of how privacy affects fairness.

Privacy. We consider differential privacy in this paper as defined below:

Definition 2.1. A randomized learning algorithm $\mathcal{A}$ is (ε, δ)-DP if, for every pair of neighboring datasets D and D' that differ in at most one element, and for every measurable subset S of the output space of $\mathcal{A}$,

$$\Pr\big[\mathcal{A}(D) \in S\big] \;\leq\; e^{\varepsilon} \Pr\big[\mathcal{A}(D') \in S\big] + \delta. \qquad (1)$$

Differentially private machine learning methods are commonly divided into three categories (Jarin and Eshete, 2022): (1) input perturbation, (2) gradient/objective perturbation, and (3) output perturbation. Here, we focus on *input perturbation* (Fukuchi et al., 2017; Kang et al., 2020), as it aligns with our goal of examining how the input data distribution shapes fairness and privacy. This choice isolates the effect of distributional changes on group-wise separability. In contrast, mechanisms such as DP-SGD (Abadi et al., 2016) couple privacy with optimization, model architecture, and training dynamics, making it difficult to disentangle distributional effects from algorithmic ones. Input perturbation instead acts directly on the class-conditional distributions, which are the objects analyzed by CD. Empirical evidence also suggests that input perturbation performs well in high-privacy regimes (Zhao et al., 2020).

A standard approach to achieve (ε, δ)-DP for input pertur-

bation privacy is the Gaussian mechanism (Dwork et al., 2014), which adds independent Gaussian noise $\sim \mathcal{N}(0, \eta^2)$ to each coordinate. Here, the noise parameter η^2 scales inversely to ϵ and δ (Dwork et al., 2014). Prior work on input perturbation attempts to determine, under specific conditions, the minimal η^2 required to achieve (ε, δ)-DP, such as strong convexity of the loss function (Kang et al., 2020) or a quadratic loss (Fukuchi et al., 2017). However, the fundamental relationship between η^2 and (ε, δ) remains unchanged: larger η^2 corresponds to smaller ε and δ, indicating stronger privacy guarantees. Since our goal is to study *whether* adding privacy tends to help or hurt fairness rather than to quantify a precise tradeoff for a specific mechanism, we abstract away these assumptions and adopt η^2 as our measure of privacy for the rest of the paper. This abstraction also allows us to remain agnostic to the choice of learning algorithm, avoiding conflating distributional effects with algorithm-specific behavior. For completeness, we provide an overview of the input perturbation framework and characterize the relationship between η^2 and ε under different assumptions in Appendix A.

3. Chernoff Difference (CD) As A Data Fairness Measure

Chernoff Information. Chernoff Information (Chernoff, 1952) determines the optimal error exponent of Bayes-optimal classification. Although information-theoretic tools have been applied to fairness and privacy (Ghassami et al., 2018), Chernoff Information itself has rarely been used in these domains. Dutta et al. (2020) show that disparities in Chernoff Information imply a trade-off between equal opportunity fairness and accuracy, while Ünsal and Önen (2024) define Chernoff DP, where output distributions for neighboring inputs must have bounded Chernoff Information. To our knowledge, our work is the first to connect Chernoff Information with the joint analysis of accuracy, fairness, and privacy, and to provide an estimator beyond synthetic settings. We begin with the formal definition below.

Definition 3.1 (Chernoff Information (Chernoff, 1952))**.** For two distributions $P_0(x)$ and $P_1(x)$ the Chernoff Information is given by:

$$C(P_0, P_1) = -\inf_{u\in(0,1)} \log\Big(\int P_0(x)^{1-u} P_1(x)^u dx\Big). \quad (2)$$

Chernoff Information (CI) can be interpreted as a divergence that quantifies the *separability* between P_0 and P_1 (Dutta et al., 2020; Nielsen, 2011). For large values of CI, two hypotheses P_0 and P_1 are more separable, which indicates easier classification. Conversely, smaller values of CI imply less separability, and thus a more difficult classification setting. This is demonstrated by the role of CI in bounding the error exponent of the Bayes optimal classifier.

Lemma 3.2 ((Nielsen, 2011))**.** *For hypotheses $P_0(x)$ under $Y = 0$ and $P_1(x)$ under $Y = 1$, CI bounds the error of the Bayes Optimal Classifier H:*

$$\Pr[H(X) \neq Y] \leq e^{-C(P_0,P_1)}. \quad (3)$$

Asymptotically, (3) establishes an equivalence between Chernoff Information and the error of the Bayes-optimal classifier. Thus, CI provides an operational measure of class separability through Chernoff-style bounds on Bayes-optimal error behavior. Nielsen (2011; 2013; 2022) provides key properties we exploit throughout this work. For a deeper discussion please refer to Appendix B.

Chernoff Difference. We next define Chernoff Difference (CD).

Definition 3.3 (Chernoff Difference)**.** The Chernoff Difference between group P and group Q is defined as follows:

$$\text{CD} = \left|C(P_0, P_1) - C(Q_0, Q_1)\right|. \quad (4)$$

CD serves as a metric that compares the classification difficulty between two groups. A smaller CD indicates that the groups have similar separability between their positive and negative classes, whereas a larger CD implies a large disparity in separability. Unlike standard fairness metrics that evaluate a learned classifier, CD measures disparity intrinsic to the data, i.e., CD works as a *data fairness metric*. Rather than evaluating a specific learned predictor, CD characterizes the best achievable error behavior permitted by the data distributions (P_0, P_1) and (Q_0, Q_1).[1].

Connection to Existing Fairness Literature. A recent work (Dutta et al., 2020) used CD to show an impossibility result: if CD is not zero, it is impossible to achieve fairness without sacrificing accuracy. While Dutta et al. (2020) draws a connection between CD and fairness, their analysis was limited to whether CD equals zero or not. In this work, we go one step further by relating the value of CD to the steepness of the fairness–accuracy curves. Intuitively, a larger CD indicates a larger gap in Chernoff separability between groups. Therefore, equalizing an error metric, such as FPR or FNR, typically requires a larger group-dependent adjustment to the classifier, leading to a larger accuracy cost. An intuitive geometric description of this is given in (Dutta et al., 2020). To capture this relationship between CD and fairness-accuracy plots, we use equal opportunity (EO) as a fairness axis throughout this paper.

Noisy Chernoff Difference. Finally, to incorporate privacy, we define a noisy variant of CD.

Definition 3.4 (Noisy Chernoff Difference)**.** For $\eta^2 \geq 0$, we define the Noisy Chernoff Difference as:

$$\widetilde{\text{CD}}_{\eta^2} = \left|C(\widetilde{P}_0, \widetilde{P}_1) - C(\widetilde{Q}_0, \widetilde{Q}_1)\right|, \quad (5)$$

[1]This can be extended to multi-group settings, Appendix B.3.

Algorithm 1 CINE via DRE-∞ (Choi et al., 2022)

Input: Distributions $\boldsymbol{P_0}, \boldsymbol{P_1}$
$s_\theta \leftarrow$ Density Ratio Estimation(P_1, P_0) {Compute score function}
$\{\log r_i\}_{i=1}^n \leftarrow \int_1^0 s_\theta(x_i, t)[t]dt$ {Using $x_i \sim P_0$}
$f(u) \leftarrow \log\left(\frac{1}{n}\sum_{i=1}^n \exp(u \cdot \log r_i)\right)$ {Monte Carlo}
$CI \leftarrow -\inf_{u\in(0,1)} f(u)$ {Convex Optimization}
Return: CI

where $\widetilde{P}_i = P_i + \mathcal{N}(0, \eta^2\mathbf{I})$ and $\widetilde{Q}_i = Q_i + \mathcal{N}(0, \eta^2\mathbf{I})$.

While CD provides insight into the relationship between fairness and accuracy, $\widetilde{CD}_{\eta^2}$ provides a single value that can capture the relationship between all three: fairness, accuracy, *and privacy*. More specifically, it allows us to analyze how adding noise (stronger privacy) affects both fairness and accuracy. We illustrate this relationship using Gaussian distributions in Section 5.1, Mixture of Gaussians in Section 5.2, and real datasets in Section 5.3.

4. Chernoff Information Neural Estimator (CINE)

Computing CD requires estimating Chernoff Information (CI) between class-conditional distributions. Closed-form expressions are available only in limited cases (Nielsen, 2013) and bounds based on fixed divergences (e.g., KL or Bhattacharyya) are generally not tight (Nielsen, 2022), while real-world data generally requires sample-based estimation (Bishop and Nasrabadi, 2006; Vapnik, 2006). We therefore introduce the Chernoff Information Neural Estimator (CINE), a novel algorithm that estimates CI from samples of unknown distributions by recasting it as a density ratio estimation (DRE) problem. To the best of our knowledge, *this is the first method for estimating CI directly from real-world datasets.* In order to develop our CI estimation algorithm, we first make the following crucial observation:

Observation 4.1. *Chernoff Information can be written as an optimization of an expectation*

$$C(P_0, P_1) = -\inf_{u\in(0,1)} \log \mathbb{E}_{x\sim P_0}\left[\left(\frac{P_1(x)}{P_0(x)}\right)^u\right]. \quad (6)$$

This yields *a density ratio* within the expectation. Accurately estimating the density ratio $\frac{P_1(x)}{P_0(x)}$ is easier than directly estimating $P_1(x)$ and $P_0(x)$ (Sugiyama et al., 2012), as shown theoretically in Kanamori et al. (2010). We leverage this insight to develop our algorithm.

Background on neural density ratio estimation. To estimate the density ratio, we adopt a state-of-the-art method known as the *telescoping approach* (Choi et al., 2022; Rhodes et al., 2020), which expresses the ratio $r(x) = \frac{P_1(x)}{P_0(x)}$ as a product of ratios between intermediate bridge distributions $p_\lambda(x)$ [2]. These bridges, indexed by λ, are created via an interpolation scheme between the original distributions (e.g., $p_\lambda(x) = \lambda P_0(x) + \sqrt{1-\lambda^2}P_1(x)$). This approach improves estimation accuracy by decomposing the overall ratio into smaller steps between distributions that are closer to each other and thus easier to estimate. Choi et al. (2022) consider the case where the number of bridges approaches infinity, allowing the log-density ratio to be expressed as an integral: $\log r(x) = \int_1^0 \frac{\partial}{\partial\lambda}\log p_\lambda(x)\partial\lambda$. Observing the similarity between this telescoping process and diffusion models (Nichol and Dhariwal, 2021; Song et al., 2021), they propose *training a neural network to recover the time score function* $\frac{\partial}{\partial\lambda}\log p_\lambda(x)$, while the integral is computed using standard numerical methods. We refer the reader to the original work (Choi et al., 2022) for full details of the estimator and Appendix F.1 for details regarding our implementation.

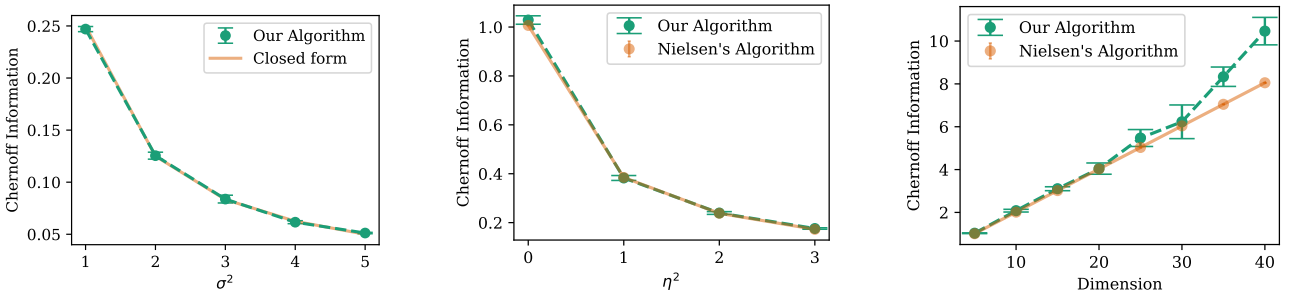
Once the neural density ratio estimator is trained, we compute the expectation $\mathbb{E}_{x\sim P_0}[(\frac{P_1(x)}{P_0(x)})^u]$ using a Monte Carlo method, and denote it as a function $g(u)$. Finally, the last step of estimating Chernoff Information is solving $-\inf_{u\in(0,1)} g(u)$ in (6). This optimization can be performed easily and efficiently due to the convex nature of g. Convexity follows from a result by Nielsen (2022):

Lemma 4.2 (Section 2.1 (Nielsen, 2022))**.** *The skewed Bhattacharyya distance,* $\log(\int P_0(x)^{1-u}P_1(x)^u dx)$ *is convex with respect to* u*.*

We provide a full overview of CINE in Algorithm 1. The key insight of CINE is that by reformulating the definition of CI, we can leverage both state-of-the-art methods for density ratio estimation (Choi et al., 2022) and recent theoretical advances on CI (Nielsen, 2022). Because CINE relies on sample-based density ratio estimation, its practical performance inherits many of the statistical challenges common to information-theoretic estimation. In particular, like other sample-based estimators of information-theoretic quantities, CINE is subject to the curse of dimensionality. Our goal in this work is not to address high-dimensional estimation in full generality, but to provide a practical estimator in settings where CD can be reliably estimated from samples, such as low- to moderate-dimensional data or learned representations. While finite-sample performance depends on dimensionality, we show that CINE is consistent as the number of samples grows (Appendix E):

Theorem 4.3. *Let* P_0 *and* P_1 *be distributions with common support and bounded density ratio. Then the Chernoff Information estimator, constructed using a consistent density ratio estimator, is itself consistent.*

[2] While we adopt a current state-of-the-art method for DRE, Algorithm 1 can plug in any DRE, and CINE can benefit from future developments in this area.



(a) CINE vs Closed form CI (2-D) *(b)* CINE vs Nielsen's (5-D) *(c)* CINE vs Nielsen's (n-D)

Figure 2. **Empirical evaluation of CINE.** (a) In 2D, using distributions $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and $\mathcal{N}(\mathbf{1}, \sigma^2\mathbf{I})$ Chernoff Information estimates are accurate for different σ^2. (b) In 5D, using distributions $\mathcal{N}(\mathbf{0}, \frac{1}{2}\mathbf{I})$ and $\mathcal{N}(\mathbf{1}, \mathbf{I})$ Chernoff Information estimation remains accurate, even as noise (η^2) is added. (c) Using distributions $\mathcal{N}(\mathbf{0}, \frac{1}{2}\mathbf{I})$ and $\mathcal{N}(\mathbf{1}, \mathbf{I})$, estimations are accurate in moderate dimensions before degrading.

4.1. Empirical Evaluation of CINE

Next, we validate CINE by comparing its results with both the closed-form solutions of CI and the algorithm proposed by Nielsen (2022). This evaluation demonstrates the accuracy of our method and its scalability with input dimension. In all experiments, we used 10,000 samples per class to train CINE. More details about the setting are in Appendix F.2.

We first test our algorithm on Gaussian distributions that have a closed form solution for Chernoff Information (see Appendix C). We construct 2D Gaussian distributions with the equal variance: $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and $\mathcal{N}(\mathbf{1}, \sigma^2\mathbf{I})$. As observed in Figure 2a, our estimator closely matches the closed form solution across a wide range of variances (σ^2).

Next, to evaluate on distributions with no closed form solution, we compare our estimator to the multidimensional Gaussian estimator from Nielsen (2022). This estimator relies on the fact that the CI between high dimensional Gaussians, can be written as a skewed KL-divergence where the optimal skewing parameter can be found with a binary search algorithm[3]. For experiments, we construct 5D Gaussian distributions $\mathcal{N}(\mathbf{0}, (\frac{1}{2}+\eta^2)\mathbf{I})$ and $\mathcal{N}(\mathbf{1}, (1+\eta^2)\mathbf{I})$. We then sweep across various noise values η^2 and show that our algorithm closely matches Nielsen's estimator (Figure 2b).

Finally, in Figure 2c, we examine how the performance of CINE scales with input dimension. We construct Gaussian distributions, $\mathcal{N}(\mathbf{0}, \frac{1}{2}\mathbf{I})$ and $\mathcal{N}(\mathbf{1}, \mathbf{I})$, with dimensions ranging from 5 to 40. With 10,000 samples per class, our estimator is nearly indistinguishable from Nielsen's up to 20 dimensions. CINE begins to show modest variance and small estimation errors between 20 and 30 dimensions, and beyond 30 dimensions the estimator starts to clearly overestimate CI compared to Nielsen's. This degradation in higher dimensions is a manifestation of the curse of dimensionality, a well-known challenge in estimating information-theoretic quantities from finite samples (Belghazi et al., 2018). We discuss these limitations further in Section 6.

[3]This relies on knowledge that the distributions are Gaussian.

5. Distribution-Dependent Analysis of Fairness–Privacy Trade-offs

Equipped with CD and CINE, we now ask: *how does the fairness–privacy interaction depend on the data distribution?* We answer this in three steps. First, we analyze the isotropic Gaussian case, where CD admits a closed form and we can prove that three qualitatively distinct regimes exist (Section 5.1). Second, we show CINE recovering different behaviors on Gaussian mixtures (Section 5.2). Third, we apply CINE to real datasets (Section 5.3), showing that the framework provides interpretable diagnostics in settings without closed-form analysis.

5.1. Isotropic Gaussians: Closed-Form Analysis

While CINE allows us to analyze CD beyond closed-form settings, the isotropic Gaussian setting admits a closed-form expression that lets us prove existence of all three regimes analytically. The empirical sections that follow show CINE recovering these same regimes in settings where no closed form is available.

We define the conditional data distributions as follows:

$$P_0 = \mathcal{N}(\mu_0, \sigma^2\mathbf{I}), P_1 = \mathcal{N}(\mu_1, \sigma^2\mathbf{I}) \tag{7}$$

$$Q_0 = \mathcal{N}(\zeta_0, \tau^2\mathbf{I}), Q_1 = \mathcal{N}(\zeta_1, \tau^2\mathbf{I}), \tag{8}$$

Under these distributions, CD and its noisy variant admit closed forms (see Appendix C for derivations and (Nielsen, 2022)):

$$\mathrm{CD} = \left| \frac{\|\mu_0 - \mu_1\|_2^2}{8\sigma^2} - \frac{\|\zeta_0 - \zeta_1\|_2^2}{8\tau^2} \right|, \tag{9}$$

$$\widetilde{\mathrm{CD}}_{\eta^2} = \left| \frac{\|\mu_0 - \mu_1\|_2^2}{8(\sigma^2+\eta^2)} - \frac{\|\zeta_0 - \zeta_1\|_2^2}{8(\tau^2+\eta^2)} \right|. \tag{10}$$

Analyzing $\widetilde{\mathrm{CD}}_{\eta^2}$ as a function of the noise η^2 yields three qualitatively distinct regimes.

Proposition 5.1. *Assume* $\|\mu_0-\mu_1\|_2 \geq \|\zeta_0-\zeta_1\|_2$ *(w.l.o.g.). The Noisy Chernoff Difference* $\widetilde{CD}_{\eta^2}$ *exhibits one of three behaviors:*

(i) $\widetilde{CD}_{\eta^2}$ *has a maximum point, when* $\frac{\|\zeta_0-\zeta_1\|_2^2}{\|\mu_0-\mu_1\|_2^2} < \frac{\tau^2}{\sigma^2} < \frac{\|\zeta_0-\zeta_1\|_2}{\|\mu_0-\mu_1\|_2} < 1$*;* ***(Case 1)***

(ii) $\widetilde{CD}_{\eta^2}$ *has a maximum point and a reflection point, when* $\frac{\tau^2}{\sigma^2} < \frac{\|\zeta_0-\zeta_1\|_2^2}{\|\mu_0-\mu_1\|_2^2} < 1$*;* ***(Case 2)***

(iii) $\widetilde{CD}_{\eta^2}$ *is non-increasing, otherwise.* ***(Case 3)***

The proof is in Appendix C. These three cases correspond to qualitatively different fairness–privacy relationships, which we construct and demonstrate below.

Three Cases of Gaussian Distributions. We now construct 1D Gaussian distributions that satisfy the conditions for each case in Proposition 5.1, and examine how different behaviors in noisy CD curves translate to different fairness-privacy relationships. We generate 100,000 samples from each distribution, and then train Naïve Bayes classifiers with different priors to obtain fairness-accuracy curves. See Appendix D.1 for more experimental details. The three cases are illustrated in Figure 3.

Case 1: Privacy Can Hurt Fairness. In Case 1, $\widetilde{\text{CD}}_{\eta^2}$ grows as we increase Gaussian noise until it reaches a maximum at around $\eta^2 = 1$. As $\widetilde{\text{CD}}_{\eta^2}$ increases, we observe the slopes of the fairness-accuracy curves becoming steeper (Figure 3b), indicating that adding privacy worsens the fairness-accuracy trade-off. We provide log fairness-accuracy plots in Appendix D, where this change in slope is more clearly visible. After the maximum point, additional noise leads to a gradual decay in $\widetilde{\text{CD}}_{\eta^2}$; however, the decay is very slow and its effect on the fairness-accuracy curves are nearly imperceptible (see Appendix D.3).

To provide intuition for Case 1, note that the group Q has closer group means than the group P. At the same time, the variance of group Q's distribution must fall within a specific range: it must be large enough (compared to P) so that Q would be the unprivileged group (i.e., less separable), but also small enough that modest noise degrades Q's separability more than P's. In this regime, because Q is already less separable, the gap in Chernoff Information between the groups increases as small noise is added. Once the noise becomes sufficiently large, it begins to affect P's separability slightly more, marking a transition where the Chernoff Difference starts to decrease.

Case 2 and Case 3: Privacy Can Help Fairness. In Case 2, the noisy CD decays very rapidly to 0 as we increase η^2 (Figure 3c). As $\widetilde{\text{CD}}_{\eta^2}$ decreases, the fairness-accuracy curves clearly flatten (Figure 3d). In fact, the curves flatten so quickly that the noisy curves begin to overlap with the clean fairness-accuracy curves. This shows that by adding privacy, we can achieve better fairness for the same accuracy—*privacy gives free fairness!*

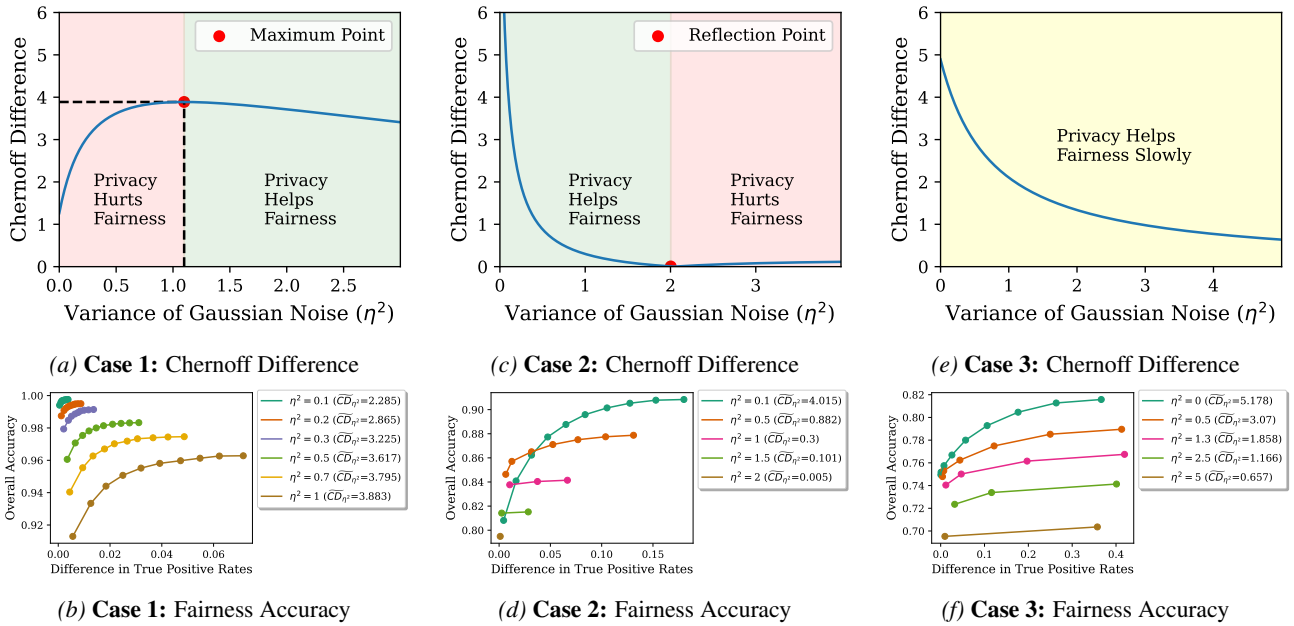
To explain the mechanism behind this "free fairness," note that the class means of group Q are still closer than those of group P's, but the variance of group Q is much smaller. As a result, Q becomes the privileged group (i.e., better separability). As we add noise, the separability of P and Q both decrease, but Q much faster than P, due to its smaller mean separation and variance. Because the privileged group loses separability faster, the Chernoff Difference decreases, effectively reducing unfairness. This trend continues until the reflection point is reached where $\widetilde{\text{CD}}_{\eta^2} = 0$. Beyond this point, $\widetilde{\text{CD}}_{\eta^2}$ starts to increase until the maximum and then decrease again, but the slope remains near zero after the reflection point (see Appendix D.4).

In Case 3, when neither condition (i) nor (ii) in Proposition 5.1 holds, $\widetilde{\text{CD}}_{\eta^2}$ smoothly decays as we increase η^2 (Figure 3e). Again, this is reflected in the flattening trend of the fairness-accuracy curve in Figure 3f. However, unlike Case 2, we do not observe any crossing of the curves, as the decay of CD is slower. Improvements in fairness cannot keep pace with the accuracy degradation induced by privacy. Here, group Q exhibits a larger variance in addition to its smaller mean separation, leading to significantly less separability compared to group P. As noise is added, the separability of P is reduced at a rate faster than Q, which closes the gap in Chernoff Information. However, the separability of Q is initially too small for CD to ever reach zero.

The CINE validation in Section 4 confirms that CINE matches closed-form Chernoff Information in Gaussian settings (Figure 2a), supporting its use in the non-closed-form experiments that follow.

5.2. Mixtures of Gaussians

We extend beyond the isotropic Gaussian setting by utilizing a more complex mixture of Gaussian setting. For each conditional distribution, we utilize two 2D Gaussian distributions to build the mixture. We consider 2 mixtures, which we name Mixture 1 and Mixture 2 and repeat the experiments from Section 5.1 (for more details on these mixtures and experiments, please refer to Appendix F). Using Mixture 1, we demonstrate that the CD spikes as we increase the variance of the noise (Figure 4a), similar to the trend we observed in Case 1 with isotropic Gaussians. We also observe a steepening in the fairness accuracy curve, as predicted by increasing CD (Figure 4b). Using Mixture 2, we show that CD can quickly decay towards 0 (Figure 4c), similar to Case 2 in the isotropic Gaussian setting. Further, we see that the fairness accuracy curves flatten as noise increases and we can obtain free fairness (Figure 4d).



(a) **Case 1:** Chernoff Difference

(b) **Case 1:** Fairness Accuracy

(c) **Case 2:** Chernoff Difference

(d) **Case 2:** Fairness Accuracy

(e) **Case 3:** Chernoff Difference

(f) **Case 3:** Fairness Accuracy

Figure 3. **Illustration of the three Gaussian cases from Proposition 5.1. (Case 1)** (a) $\widetilde{\mathrm{CD}}_{\eta^2}$ initially increases with noise, reaching a maximum corresponding to the worst fairness–accuracy trade-off. (b) The fairness–accuracy curves become steeper as noise increases up to this point. Parameters: $\mu_0 = 0, \mu_1 = 16.5, \sigma = 2.43, \zeta_0 = 0.5, \zeta_1 = 3.8, \tau = 0.55$. **(Case 2)** (c) $\widetilde{\mathrm{CD}}_{\eta^2}$ decreases toward 0 before slowly increasing again; the maximum occurs at a large η^2 value and is omitted (Appendix D.4). (d) Fairness–accuracy curves flatten as noise increases, and can intersect, indicating improved fairness at the same accuracy. Parameters: $\mu_0 = -4.2, \mu_1 = 1.3, \sigma = 3$, $\zeta_0 = 0.3, \zeta_1 = 2.7, \tau = 0.25$. **(Case 3)** (e) $\widetilde{\mathrm{CD}}_{\eta^2}$ decreases monotonically over the positive η^2 regime. (f) Fairness–accuracy curves flatten gradually with increasing noise, but without the clear crossover behavior observed in Case 2. Parameters: $\mu_0 = -4.2, \mu_1 = 1.3, \sigma = 0.85, \zeta_0 = 0.6, \zeta_1 = 1.6, \tau = 0.6$.

5.3. Real-world Datasets

Adult. We next apply CINE on the Adult dataset (Becker and Kohavi, 1996). We consider the standard income classification task and use sex as our protected attribute. We selected all the continuous features from the dataset, giving us 6 dimensions. For full details refer to Appendix F.3.

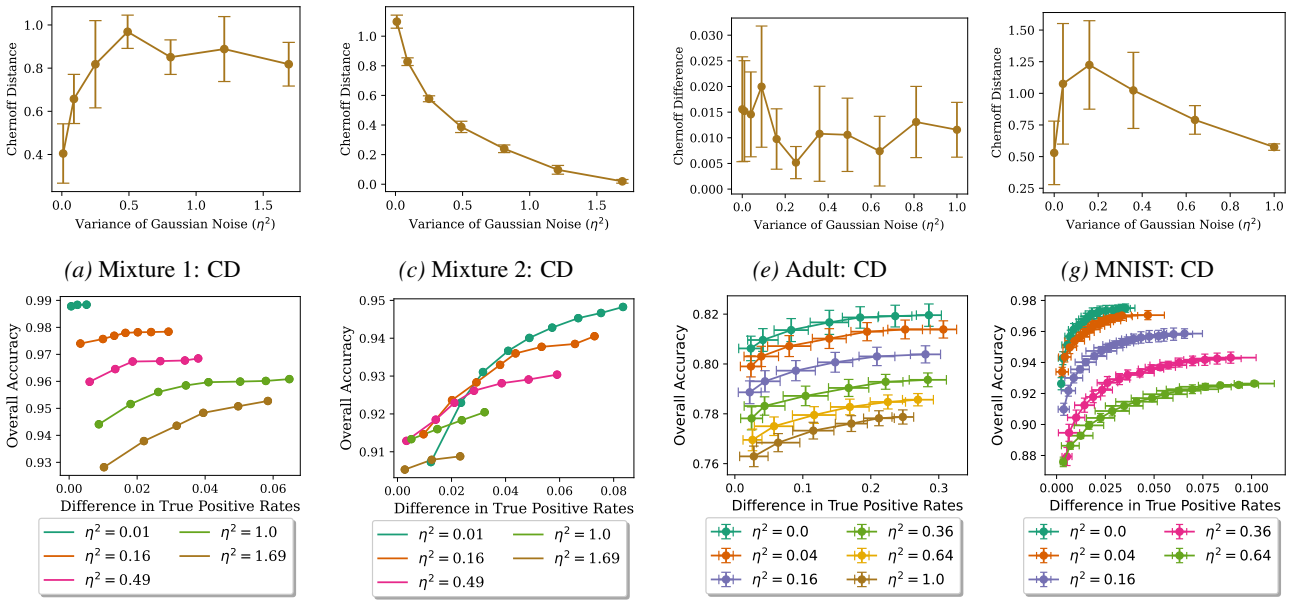
First, as the variance of the additive Gaussian noise increases, the CD remains relatively flat (Figure 4e). We observe that there may be slight trends in CD, however, they appear to be too small to make meaningful differences in the fairness, privacy dynamic. This is reflected by the fairness accuracy curves (Figure 4f). We train split logistic regression classifiers and perturb the class weights for the disadvantaged group to create our fairness-accuracy curves. For more details, please refer to Appendix F.3. As observed in Section 5.1, the CD acts as a proxy for the slope of the fairness accuracy curves. Here, as the input noise varies, the slopes of the fairness accuracy curves remain relatively stable. Further, as the CD is consistently low, we observe very flat fairness accuracy curves. This indicates that privacy does not have a strong effect on the relationship between fairness and accuracy for these features.

HSLS. We replicate this experiment on the HSLS dataset in Appendix G demonstrating CD continues to to capture the observed fairness–privacy behavior in this additional real-world setting.

MNIST. Next, we apply CINE to an image classification task on MNIST. Estimating information-theoretic quantities directly in high-dimensional spaces is fundamentally challenging. In practice, modern neural estimators such as MINE (Belghazi et al., 2018) are therefore typically evaluated in relatively low-dimensional settings, often below 20 dimensions, or rely on low intrinsic-dimensional structure. Motivated by these considerations, we apply CINE to 30D learned representations of MNIST digits rather than directly in pixel space. Additional discussion on representation quality and dimensionality is provided in Appendix I.

We construct a synthetic fair classification task on the MNIST dataset by assigning specific digits to represent each subgroup in our experiment. While MNIST does not involve human subjects, this setup allows us to study fairness–accuracy trade-offs in a controlled setting by synthetically defining majority and minority groups. Specifically, we choose the digit 3 for P_0, the digit 4 for P_1, the digit 7 for Q_0, and the digit 9 for Q_1.

For CD, we observe an increase in CD until $\eta^2 = 0.16$ before a steady decay as we increase the input perturbation noise (Figure 4g). This trend is reflected in the fairness-



(a) Mixture 1: CD *(c)* Mixture 2: CD *(e)* Adult: CD *(g)* MNIST: CD

(b) Mixture 1: Fairness-Acc *(d)* Mixture 2: Fairness-Acc *(f)* Adult: Fairness-Acc *(h)* MNIST: Fairness-Acc

Figure 4. **Illustrations on Gaussian mixtures, Adult, and MNIST. (Mixture 1)** (a) $\widetilde{\mathrm{CD}}_{\eta^2}$ increases with noise. (b) Fairness–accuracy curves become steeper, resembling Case 1 from isotropic setting. **(Mixture 2)** (c) $\widetilde{\mathrm{CD}}_{\eta^2}$ decreases toward 0. (d) Fairness–accuracy curves flatten, resembling Case 2 from isotropic setting. **(Adult)** (e) $\widetilde{\mathrm{CD}}_{\eta^2}$ remains nearly constant. (f) Fairness–accuracy slopes are stable. **(MNIST)** (g) $\widetilde{\mathrm{CD}}_{\eta^2}$ increases before decaying on MNIST representations. (h) Fairness–accuracy curves steepen and then flatten accordingly.

accuracy curves, which exhibit moderate steepening until $\eta^2 = 0.16$ before exhibiting a clear flattening effect (Figure 4h) demonstrating that CD effectively captures the fairness-privacy-accuracy trade-off on representations of real data.

6. Conclusion and Future Work

This work introduces *Chernoff Difference* (CD), an information-theoretic data fairness measure, together with *CINE*, a neural estimator that makes CD computable from samples beyond closed-form settings. Together, they provide a principled framework for analyzing the fundamental fairness–privacy–accuracy trade-off inherent in the input data distribution.

While this paper focuses on establishing the core concepts and tools, it naturally opens up several directions for future work. In particular, although we prove consistency of the proposed CINE algorithm and demonstrate strong empirical performance in moderate dimensions ($\sim$30), a deeper sample complexity analysis and further improvements to estimator performance in higher dimensions remain important directions for future study. Further, CINE enables the use of Chernoff Information in other domains such as data attribution, representation analysis, distribution shift detection, or other tasks involving comparisons between complex high-dimensional distributions.

Additionally, we study perturbation-based privacy, where privacy directly changes the input distribution. Other paradigms, such as DP-SGD, federated learning, secure aggregation, or institutional privacy constraints, alter the learning or information-sharing process rather than only the input distribution. In those settings, fairness–privacy interactions may also depend on optimization, communication, participation, and institutional structure. Extending CD-style distributional diagnostics to these mechanisms is a promising direction for future work. Finally, we believe that the proposed framework can inform fair and private mechanism design, e.g., Chernoff Difference analysis may be used as a preliminary diagnostic to assess whether privacy mechanisms are likely to adversely affect fairness for a given data distribution, helping practitioners decide when additional fairness constraints are necessary.

Impact Statement

Privacy and fairness are critical requirements in high-stakes domains such as finance, education, and law enforcement. Misconceptions about their relationship—for example, that privacy always harms fairness—can lead to suboptimal design choices that needlessly compromise one for the other. This work advances a data-dependent understanding of the fairness–privacy trade-off, enabling more informed algorithm design. This perspective can motivate future research on designing data-specific algorithms for fair and private machine learning.

Acknowledgments This work was supported by the National Science Foundation (NSF) under grant number 2341055.

Disclaimer. This paper was prepared for informational purposes by the Global Technology Applied Research center of JPMorgan Chase & Co. This paper is not a product of the Research Department of JPMorgan Chase & Co. or its affiliates. Neither JPMorgan Chase & Co. nor any of its affiliates makes any explicit or implied representation or warranty and none of them accept any liability in connection with this paper, including, without limitation, with respect to the completeness, accuracy, or reliability of the information contained herein and the potential legal, compliance, tax, or accounting effects thereof. This document is not intended as investment research or investment advice, or as a recommendation, offer, or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction.

References

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16. ACM, Oct. 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.

W. Alghamdi, H. Hsu, H. Jeong, H. Wang, P. Michalak, S. Asoodeh, and F. Calmon. Beyond adult and compas: Fair multi-class prediction via information projection. *Advances in Neural Information Processing Systems*, 35: 38747–38760, 2022.

E. Bagdasaryan, O. Poursaeed, and V. Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

S. Barocas and A. D. Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

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

M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1493–1504. ACM, June 2023. doi: 10.1145/3593013.3594095. URL http://dx.doi.org/10.1145/3593013.3594095.

C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006.

N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

H. Chang and R. Shokri. On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 292–303. IEEE, 2021.

H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.

K. Choi, C. Meng, Y. Song, and S. Ermon. Density ratio estimation via infinitesimal classification. In *International Conference on Artificial Intelligence and Statistics*, pages 2552–2573. PMLR, 2022.

A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2016. URL https://arxiv.org/abs/1610.07524.

T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 1999.

R. Cummings, V. Gupta, D. Kimpara, and J. Morgenstern. On the compatibility of privacy and fairness. In *Adjunct publication of the 27th conference on user modeling, adaptation and personalization*, pages 309–315, 2019.

S. Das, M. Donini, J. Gelman, K. Haas, M. Hardt, J. Katzman, K. Kenthapadi, P. Larroy, P. Yilmaz, and B. Zafar. Fairness measures for machine learning in finance. *The Journal of Financial Data Science*, 2021.

S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International conference on machine learning*, pages 2803–2813. PMLR, 2020.

C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

F. Fioretto, C. Tran, P. Van Hentenryck, and K. Zhu. Differential privacy and fairness in decisions and learning tasks: A survey. *arXiv preprint arXiv:2202.08187*, 2022.

F. Friedrich, M. Brack, L. Struppek, D. Hintersdorf, P. Schramowski, S. Luccioni, and K. Kersting. Fair diffusion: Instructing text-to-image generation models on fairness, 2023. URL https://arxiv.org/abs/2302.10893.

K. Fukuchi, Q. K. Tran, and J. Sakuma. Differentially private empirical risk minimization with input perturbation. In *International Conference on Discovery Science*, pages 82–90. Springer, 2017.

C. Garvie and J. Frankle. Facial-recognition software might have a racial bias problem. *The Atlantic*, 7(04):2017, 2016.

A. Ghassami, S. Khodadadian, and N. Kiyavash. Fairness in supervised learning: An information theoretic approach, 2018. URL https://arxiv.org/abs/1801.04378.

H. Ghoukasian and S. Asoodeh. Differentially private fair binary classifications. *arXiv preprint arXiv:2402.15603*, 2024.

A. Gomstyn and A. Jonker. Exploring privacy issues in the age of ai. *IBM Insights*, September 2024. URL https://www.ibm.com/think/insights/ai-privacy.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

S. J. Ingels, D. J. Pratt, D. R. Herget, L. J. Burns, J. A. Dever, R. Ottem, J. E. Rogers, Y. Jin, and S. Leinwand. High School Longitudinal Study of 2009 (HSLS:09): Base-Year Data File Documentation. Technical Report NCES 2011-328, National Center for Education Statistics, 2011.

M. Jagielski, M. Kearns, J. Mao, A. Oprea, A. Roth, S. Sharifi-Malvajerdi, and J. Ullman. Differentially private fair learning. In *International Conference on Machine Learning*, pages 3000–3008. PMLR, 2019.

I. Jarin and B. Eshete. Dp-util: comprehensive utility analysis of differential privacy in machine learning. In *Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy*, pages 41–52, 2022.

H. Jeong, H. Wang, and F. P. Calmon. Fairness without imputation: A decision tree approach for fair prediction with missing values. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

T. Kanamori, T. Suzuki, and M. Sugiyama. Theoretical analysis of density ratio estimation. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 93(4):787–798, 2010.

Y. Kang, Y. Liu, B. Niu, X. Tong, L. Zhang, and W. Wang. Input perturbation: A new paradigm between central and local differential privacy, 2020. URL https://arxiv.org/abs/2002.08570.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

D. S. Lemons and P. Langevin. *An introduction to stochastic processes in physics*. JHU Press, 2002.

A. Lowy, S. Baharlouei, R. Pavan, M. Razaviyayn, and A. Beirami. A stochastic optimization framework for fair risk minimization, 2023. URL https://arxiv.org/abs/2102.12586.

L. Lyu, X. He, and Y. Li. Differentially private representation for nlp: Formal guarantee and an empirical study on privacy and fairness, 2020. URL https://arxiv.org/abs/2010.01285.

P. Mangold, M. Perrot, A. Bellet, and M. Tommasi. Differential privacy has bounded impact on fairness in classification. In *International Conference on Machine Learning*, pages 23681–23705. PMLR, 2023.

W. K. Newey. Uniform convergence in probability and stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, pages 1161–1167, 1991.

W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4: 2111–2245, 1994.

A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

F. Nielsen. Chernoff information of exponential families. *CoRR*, abs/1102.2684, 2011. URL http://arxiv.org/abs/1102.2684.

F. Nielsen. An information-geometric characterization of chernoff information. *IEEE Signal Processing Letters*, 20 (3):269–272, Mar. 2013. ISSN 1558-2361. doi: 10.1109/lsp.2013.2243726. URL http://dx.doi.org/10.1109/LSP.2013.2243726.

F. Nielsen. Revisiting chernoff information with likelihood ratio exponential families. *Entropy*, 24(10):1400, 2022.

N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

B. Rhodes, K. Xu, and M. U. Gutmann. Telescoping density-ratio estimation. *Advances in neural information processing systems*, 33:4905–4916, 2020.

A. Sanyal, Y. Hu, and F. Yang. How unfair is private learning? In *Uncertainty in Artificial Intelligence*, pages 1738–1748. PMLR, 2022.

S. Shaham, A. Hajisafi, M. K. Quan, D. C. Nguyen, B. Krishnamachari, C. Peris, G. Ghinita, C. Shahabi, and P. N. Pathirana. Privacy and fairness in machine learning: A survey. *IEEE Transactions on Artificial Intelligence*, 6 (7):1706–1726, 2025. doi: 10.1109/TAI.2025.3531326.

R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

S. Singh and B. Póczos. Finite-sample analysis of fixed-k nearest neighbor density functional estimators. *Advances in neural information processing systems*, 29, 2016.

Y. Song, C. Durkan, I. Murray, and S. Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34: 1415–1428, 2021.

M. Sugiyama, T. Suzuki, and T. Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

C. Tran, M. Dinh, and F. Fioretto. Differentially private empirical risk minimization under the fairness lens. *Advances in Neural Information Processing Systems*, 34: 27555–27565, 2021a.

C. Tran, F. Fioretto, P. Van Hentenryck, and Z. Yao. Decision making with differential privacy under a fairness lens. In *IJCAI*, pages 560–566, 2021b.

V. Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

H. Wang, H. Hsu, M. Diaz, and F. P. Calmon. To split or not to split: The impact of disparate treatment in classification. *IEEE Transactions on Information Theory*, 67 (10):6733–6757, 2021.

K. Yao and M. Juarez. Sok: What makes private learning unfair? *arXiv preprint arXiv:2501.14414*, 2025.

B. Z. H. Zhao, M. A. Kaafar, and N. Kourtellis. Not one but many tradeoffs: Privacy vs. utility in differentially private machine learning. In *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, pages 15–26, 2020.

A. Ünsal and M. Önen. Chernoff information as a privacy constraint for adversarial classification. In *2024 60th Annual Allerton Conference on Communication, Control, and Computing*, pages 1–8, 2024. doi: 10.1109/Allerton63246.2024.10735286.

Appendix

Appendix A provides background on privacy, the input perturbation notion, and how noise connects to differential privacy. Appendix B provides more background on Chernoff Information and its connection to fairness. Appendix C reiterates useful Theorems and Lemmas from the Isotropic Gaussian examples as well as presents their proofs. Appendix D provides a comparison between the fairness-accuracy curves and the log-fairness-accuracy curves for the closed-form Gaussian experiments in Section 5.1. Additionally it holds a more comprehensive overview of Case 1 and 2 in the Gaussian setting. Appendix E contains the proof of consistency for the Chernoff Information Neural Estimator. Appendix F holds experimental details. Appendix G contains additional experiments on the HSLS dataset. Appendix H provides a comparison of the fairness-accuracy curves and the log-fairness-accuracy curves for real world data. Appendix I provides additional discussion about representations. Finally, Appendix J provides a brief ablation over hyperparameter choices.

A. Background on Privacy

In this work, we utilize the common notion of differential privacy.

Definition A.1 (Differential Privacy)**.** A randomized learning algorithm $\mathcal{A}$ is (ε, δ)-differentially private if, for every pair of neighboring datasets D and D' that differ in at most one element, and for every measurable subset S of the output space of $\mathcal{A}$,

$$\Pr\big[\mathcal{A}(D) \in S\big] \;\leq\; e^{\varepsilon}\,\Pr\big[\mathcal{A}(D') \in S\big] + \delta. \tag{11}$$

Throughout this work, we utilize the input perturbation notion of privacy (Algorithm 2).

Algorithm 2 SGD with Gaussian Input Perturbation

Input: Dataset D, batch size m, learning rate α, noise variance η^2, initial parameters $\boldsymbol{\theta_0}$
Create noisy dataset $\widetilde{D} = \{\tilde{\mathbf{x}}_\mathbf{i}, y\}_{i=1}^n$ where $\tilde{\mathbf{x}}_\mathbf{i} = \mathbf{x_i} + \boldsymbol{\xi_i}$ with $\boldsymbol{\xi_i} \sim \mathcal{N}(\mathbf{0}, \eta^2\mathbf{I})$
for $t = 1, 2, \ldots, T$ **do**
    Sample a mini-batch $\{\mathbf{x_i}, y_i\}_{i=1}^m \sim \widetilde{D}$
    Compute stochastic gradient:
        $\mathbf{g_t} = \frac{1}{m}\sum_{i=1}^m \nabla_\theta \ell(\boldsymbol{\theta_{t-1}}; \tilde{\mathbf{x}}_\mathbf{i}, y_i)$
    Update parameters: $\boldsymbol{\theta_t} \leftarrow \boldsymbol{\theta_{t-1}} - \alpha\mathbf{g_t}$
end for
Return: θ_T

For input perturbation, the naive bound on privacy stems from the feature level notion of neighboring datasets.

Definition A.2 (Neighboring Datasets – Feature Level)**.** Two datasets $D = \{(x_i, y_i)\}_{i=1}^n$ and $D' = \{(x_i', y_i)\}_{i=1}^n$ are neighboring at the *feature-level* if they differ in the features of a single element only. That is,

$$\exists\, j \in [n] \text{ such that } x_j \neq x_j' \text{ but } y_j = y_j',$$

and for all $i \neq j$, $x_i = x_i'$ and $y_i = y_i'$.

The input perturbation notion of privacy leverages the Gaussian Mechanism to achieve differential privacy.

Definition A.3 (Gaussian Mechanism (Dwork et al., 2014))**.** Given a function $f : \mathcal{D} \to \mathbb{R}^d$ with ℓ_2-sensitivity $\Delta_2(f) = \sup_{D,D'} \|f(D) - f(D')\|_2$ over all neighboring datasets D, D', the Gaussian Mechanism with parameter $\eta > 0$ outputs

$$\mathcal{M}(D) = f(D) + \mathcal{N}(0, \eta^2 \boldsymbol{I}_d),$$

where $\mathcal{N}(0, \sigma^2 \boldsymbol{I}_d)$ is the d-dimensional Gaussian distribution with mean 0 and covariance $\eta^2 \boldsymbol{I}_d$.

It follows that the Gaussian Mechanism can provide differential privacy

Theorem A.4 (Theorem A.1 (Dwork et al., 2014))**.** *Let $\varepsilon \in (0, 1)$ be arbitrary. For $c^2 \geq 2\ln(1.25/\delta)$ the Gaussian Mechanism with parameter $\eta \geq c\Delta_2(f)/\varepsilon$ is (ε, δ)-differentially private.*

The input perturbation notion of privacy is equivalent to applying the Gaussian Mechanism to the identity function on our dataset and allows us to derive a relationship between privacy parameters (ε, δ) and the noise parameter η^2.

Lemma A.5 (Dwork et al. (2014)). *Assume $\|\mathbf{x}\|_2 \leq 1$ for all $\mathbf{x} \in \mathbb{R}^d$. When $\eta^2 \geq \frac{8\log(1.25/\delta)}{\varepsilon^2}$, Algorithm 1 is (ε, δ)-DP with respect to feature-level neighboring datasets (Definition A.2).*

Proof. Since $\|\mathbf{x}\|_2 \leq 1$, the ℓ_2-sensitivity of the identity function is bounded by 2. Thus, by Theorem A.5, we get (ε, δ)-differential privacy with noise $\eta^2 \geq \frac{8\log(1.25/\delta)}{\varepsilon^2}$. This (ε, δ)-differential privacy extends to Algorithm 1 by the postprocessing property of differential privacy. □

Recent works have explored tightening this bound and generalizing to the more common notion of neighboring datasets.

Definition A.6 (Neighboring Datasets). Two datasets $D = \{(x_i, y_i)\}_{i=1}^n$ and $D' = \{(x'_i, y'_i)\}_{i=1}^n$ are *neighboring* if they differ in the data of a single element. That is,

$$\exists\, j \in [n] \text{ such that } (x_j, y_j) \neq (x'_j, y'_j),$$

and for all $i \neq j$, $(x_i, y_i) = (x'_i, y'_i)$.

For T training steps and n samples, Kang et al. (2020) show that the noise bound becomes $O(\frac{G^2 T \log}{n(n-1)\sqrt{\Delta}\varepsilon^2})$ for a G-Lipschitz and Δ-strongly convex loss function . Fukuchi et al. (2017), show that for a quadratic loss function the noise bound becomes $O(\frac{G^2\log(16/\delta))+\varepsilon}{n\varepsilon^2})$. In all cases, $\eta^2 \propto \frac{\log(1/\delta)}{\varepsilon^2}$. Thus, throughout this paper we utilize η^2 as our privacy parameter, since it provides a simple and standard expression directly linked to (ε, δ) via the Gaussian mechanism, allowing us to avoid restricting our analysis with assumptions on the learning algorithm. While tighter bounds can be derived, they do not affect the qualitative behavior of our results.

B. Background on Chernoff Information

B.1. Bounding the Error of Bayes Optimal Classifier

Lemma B.1 ((Nielsen, 2011)). *For hypotheses $P_0(x)$ under $Y = 0$ and $P_1(x)$ under $Y = 1$, Chernoff Information bounds the error of the Bayes Optimal Classifier H:*

$$\Pr[H(X) \neq Y] \leq e^{-C(P_0, P_1)}.$$

Proof. Note: Most of this proof is from Nielsen (2011). We have included it for completeness.

Consider the binary classification setting. Let $\Pr[Y = 0] = w_0 > 0$ and let $\Pr[Y = 1] = w_1 = 1 - w_0 > 0$. The class probabilities can be defined as $p_0(x) = \Pr[x|Y = 0]$ and $p_1(x) = \Pr[x|Y = 1]$. Now, consider the Bayes decision rule in which $H(x)$ classifies as $\hat{Y} = 0$ if $Pr[Y = 0|x] > Pr[Y = 1|x]$ and $\hat{Y} = 1$ otherwise. Using Bayes rule, $\Pr[Y = i|x] = \frac{\Pr[Y=i]\Pr[x|Y=i]}{\Pr[x]} = \frac{w_i p_i(x)}{p(x)}$ for $i \in \{0, 1\}$. Now, we can obtain the probability of error:

$$\Pr(\text{Error} \mid x) = \begin{cases} \Pr(Y = 0 \mid x) & \text{if } H \text{ wrongly decided } \hat{Y} = 1, \\ \Pr(Y = 1 \mid x) & \text{if } H \text{ wrongly decided } \hat{Y} = 0. \end{cases}.$$

Thus, the Bayes decision rule minimizes by principle the average probability of error:

$$E^* = \int \Pr[\text{Error}|x]p(x)dx = \int \min\{\Pr[Y = 0|x], \Pr[Y = 1|x]\}p(x)dx.$$

Now, to upper bound this Bayes error, we can utilize knowledge that, for $a, b > 0$, $\min\{a, b\} \leq a^\alpha b^{1-\alpha}\ \ \forall \alpha \in (0, 1)$. This implies that:

$$E^* \leq w_0^\alpha w_1^{1-\alpha} \int p_0(x)^\alpha p_1(x)^{1-\alpha} dx.$$

Now, let $c_\alpha(p_0 : p_1) = \int p_0(x)^\alpha p_1(x)^{1-\alpha} dx$. This term is referred to as the Chernoff α-coefficient. Since the above inequality holds for all $\alpha \in (0, 1)$, the best exponent for upper bounding the Bayes error corresponds to the optimal Chernoff α-coefficient:

$$c^*(p_0 : p_1) = c_{\alpha^*}(p_0 : p_1) = \inf_{\alpha \in (0,1)} \int p_0(x)^\alpha p_1(x)^{1-\alpha} dx$$

The Chernoff Coefficient gives a measure of similarity which yields CI:

$$C(p_0, p_1) = -\log \inf_{\alpha \in (0,1)} \int p_0(x)^\alpha p_1(x)^{1-\alpha} dx.$$

CI yields the best achievable exponent for a Bayesian probability of error:

$$E^* \leq w_0^\alpha w_1^{1-\alpha} e^{-C(p_0, p_1)}.$$

Lemma 1 follows as $w_0, w_1 < 1$. □

B.1.1. TIGHTNESS OF BOUND

CI provides an asymptotically tight bound on the error exponent, i.e.,

$$\lim_{n \to \infty} -\frac{1}{n} \log P_e^{(n)} = C(P_0, P_1),$$

where n is the number of samples (Cover and Thomas, 1999). For instance, for Gaussian distributions, we can explicitly write:

$$c_1 e^{-nC(P_0, P_1)} \leq P_e^{(n)} \leq c_2 e^{-nC(P_0, P_1)},$$

which shows that CI provides an exponent-tight bound (c_1 and c_2 are constants with a closed-form expression).

B.2. Connection to Fairness Metrics

Next, we provide an overview of Chernoff Information and its connection to fairness following results from Dutta et al. (2020).

Consider the binary classification task ($Y \in \{0, 1\}$) with a binary protected attribute ($S \in \{0, 1\}$). Consider the standard Bayes optimal classification rule of split classifiers: $H_s(x) = \frac{\Pr[Y=1|S=s,x]}{\Pr[Y=0|S=s,x]} \geq \tau_s$ in which the classifier predicts positive if the ratio of the conditionals is greater than a threshold τ_s.

Lemma B.2 (Lemma 2 from Dutta et al. (2020))**.** *The Chernoff exponent of the probability of error of the Bayes optimal classifier is given by the Chernoff Information:*

$$E_e = C(P_0, P_1) = -\inf_{u \in (0,1)} \log\left(\int P_0(x)^{1-u} P_1(x)^u dx\right). \tag{12}$$

As highlighted in Lemma 3.2, this Chernoff exponent bounds the probability of error of the Bayes optimal classifier. We now show a connection between false positive and false negative error rates and Chernoff Information.

Definition B.3 (Chernoff Exponents—Definition 1 from Dutta et al. (2020))**.** The Chernoff exponents of the False positive error and False negative error are defined as

$$E_{\text{FP}}(\tau_s) = \sup_{u>0}(u\tau_s - \Lambda_0(u)) \tag{13}$$

$$E_{\text{FN}}(\tau_s) = \sup_{u<0}(u\tau_s - \Lambda_1(u)) \tag{14}$$

where Λ_0 and Λ_1 are the log-generating functions.

Similarly, these Chernoff exponents bound the probability of false positive and false negative errors.

Lemma B.4 (Chernoff Bound - Lemma 1 from Dutta et al. (2020))**.** *The Chernoff exponents of false positive error and false negative error bound the probability of false positive and false negative error*

$$P_{FP} \leq e^{-E_{FP}(\tau_s)} \tag{15}$$

$$P_{FN} \leq e^{-E_{FN}(\tau_s)} \tag{16}$$

Under the equal prior condition, the overall Chernoff exponent can be written as simple function of the false positive and false negative exponents.

Definition B.5 (Definition 2 from Dutta et al. (2020))**.** Suppose $\Pr[S=0] = \Pr[S=1]$ and $\Pr[Y=1|S=s] = \Pr[Y=0|S=s]$ for all s. The Chernoff exponent of the overall probability of error P_e is defined as

$$E_e = \min\{E_{FP}(\tau_s), E_{FN}(\tau_s)\} \tag{17}$$

Thus, Chernoff Difference naturally measures disparities in error exponents across groups (False positive or False negative exponents).

$$\text{CD} = |\min\{E_{FP}(\tau_P), E_{FN}(\tau_P)\} - \min\{E_{FP}(\tau_Q), E_{FN}(\tau_Q)\}| \tag{18}$$

In the asymptotic regime, CD expresses disparities in terms of the maximum error rate on a log-scale, capturing a geometric difference rather than the arithmetic difference used in common fairness metrics such as Equal Opportunity. Next, we demonstrate how Chernoff Difference relates to the standard arithmetic difference in this asymptotic regime.

Theorem B.6. *Let* $M_p = \max\{P^P_{FP}, P^P_{FN}\}$ *and* $M_q = \max\{P^Q_{FP}, P^Q_{FN}\}$. *For all constants* $0 < c < C$ *with* $c \leq M_P, M_Q \leq C$*, we have*

$$c\,|\log M_p - \log M_Q| \leq |M_p - M_Q| \leq C\,|\log M_p - \log M_Q|\,. \tag{19}$$

Proof. Since $M_p, M_q \in [c, C]$ with $c > 0$, the Mean Value Theorem implies

$$\log M_p - \log M_q = \log'(\xi)(M_p - M_q) = \frac{1}{\xi}(M_p - M_q) \tag{20}$$

for some ξ between M_p and M_q. Thus

$$|M_p - M_q| = \xi\,|\log M_p - \log M_q|. \tag{21}$$

Since $\xi \in [c, C]$, we have

$$c\,|\log M_p - \log M_q| \;\leq\; |M_p - M_q| \;\leq\; C\,|\log M_p - \log M_q|. \tag{22}$$

□

When we break the equal prior assumption the Chernoff Information continues to capture error exponents, however skewed by the class priors.

Definition B.7 (Section E.1 (Dutta et al., 2020))**.** Let $\pi_0 = P[Y=0]$ and $\pi_1 = P[Y=1]$. The Chernoff exponent of the overall probability of error P_e is defined as

$$E_e = \min\{E_{FP}(\tau_s) - \log 2\pi_0, E_{FN}(\tau_s) - \log 2\pi_1\} \tag{23}$$

Thus, the Chernoff Difference continues to evaluate the difference of the error exponents across groups. In the case where where the exponents align as false negative exponents, we again achieve a similar value to equal opportunity (although skewed by the log-ratio of the priors).

$$\text{CD} = |\min\{E_{FP}(\tau_P) - \log 2\pi_0^{(P)}, E_{FN}(\tau_P) - \log 2\pi_1^{(P)}\} - \min\{E_{FP}(\tau_Q) - \log 2\pi_0^{(Q)}, E_{FN}(\tau_Q) - \log 2\pi_1^{(Q)}\}| \tag{24}$$

B.3. Multi-group Fairness

While we focus on the binary setting, we can generalize our framework to multi-group settings. In these settings, Chernoff Information can play the role of a group statistic (analogous to group accuracy or group false-positive rate), which can be compared across all groups. The most natural generalization is to compute the maximum Chernoff Difference across all groups:

$$\mathrm{CD} = \max_{i \neq j \in \mathcal{S}} \left| C\left(P_0^{(i)}, P_1^{(i)}\right) - C\left(P_0^{(j)}, P_1^{(j)}\right) \right|, \tag{25}$$

where the superscript denotes different sensitive groups.

This worst-case Chernoff Difference then captures the largest disparity in classification difficulty across all groups. This corresponds to the maximum notion of equal opportunity:

$$\mathrm{EO}_{\max} = \max_{i \neq j \in \mathcal{S}} |\mathrm{FNR}_i - \mathrm{FNR}_j|, \tag{26}$$

which is a common way to define equal opportunity for multi-group settings (Hardt et al., 2016; Alghamdi et al., 2022).

C. Isotropic Gaussian Proofs

Lemma C.1. *When $P_0(x) \sim \mathcal{N}(\mu_0, \sigma^2\mathbf{I})$, $P_1(x) \sim \mathcal{N}(\mu_1, \sigma^2\mathbf{I})$, $Q_0(x) \sim \mathcal{N}(\zeta_0, \tau^2\mathbf{I})$, and $Q_1(x) \sim \mathcal{N}(\zeta_1, \tau^2\mathbf{I})$, the Chernoff Difference is given as:*

$$CD = \left| \frac{\|\mu_0 - \mu_1\|_2^2}{8\sigma^2} - \frac{\|\zeta_0 - \zeta_1\|_2^2}{8\tau^2} \right|.$$

Proof. Recall the definition of Chernoff Difference and Chernoff Information

$$CD = |C(P_0, P_1) - C(Q_0, Q_1)| \tag{27}$$

$$= |\min_{u \in (0,1)} \log \int P_0(x)^u P_1(x)^{1-u} dx - \min_{v \in (0,1)} \log \int Q_0(x)^v Q_1(x)^{1-v} dx|. \tag{28}$$

Now, following a result from Dutta et al. (2020), we see that for $P_0 \sim \mathcal{N}(\mu_0, \sigma^2\mathbf{I})$, $P_1 \sim \mathcal{N}(\mu_1, \sigma^2\mathbf{I})$.

$$\log \int P_0(x)^u P_1(x)^{1-u} dx \tag{29}$$

$$= \log \int e^{-\frac{u}{2\sigma^2}\left((x-\mu_1)^T(x-\mu_1) - (x-\mu_0)^T(x-\mu_0)\right)} P_0(x)\, dx \tag{30}$$

$$= \log e^{-\frac{u}{2\sigma^2}\left(\mu_1^T\mu_1 - \mu_0^T\mu_0\right)} \int e^{-\frac{u}{2\sigma^2}\left(-2x^T(\mu_1-\mu_0)\right)} P_0(x)\, dx \tag{31}$$

$$= \log e^{-\frac{u}{2\sigma^2}\left(\mu_1^T\mu_1 - \mu_0^T\mu_0\right)} e^{-\frac{u}{2\sigma^2}\left(-2\mu_0^T(\mu_1-\mu_0)\right)} e^{\frac{u^2}{2\sigma^2}\left(\|\mu_1-\mu_0\|_2^2\right)} \tag{32}$$

$$= \log e^{-\frac{u}{2\sigma^2}\left(\|\mu_1-\mu_0\|_2^2\right)} e^{\frac{u^2}{2\sigma^2}\left(\|\mu_1-\mu_0\|_2^2\right)} \tag{33}$$

$$= \frac{u(u-1)}{2\sigma^2}\|\mu_1 - \mu_0\|_2^2. \tag{34}$$

Now, we can compute the derivative to minimize with respect to u:

$$\frac{d}{du}\frac{u(u-1)}{2\sigma^2}\|\mu_1 - \mu_0\|_2^2 = \frac{2u-1}{2\sigma^2}\|\mu_1 - \mu_0\|_2^2. \tag{35}$$

The derivative is 0, when $u = 0.5$. This critical point is a minimum following a first derivative test. Thus,

$$C(P_0, P_1) = \frac{\|\mu_0 - \mu_1\|_2^2}{8\sigma^2}. \tag{36}$$

When these distributions are 1-dimensional, they reduce to the closed form derived in Nielsen (2022). A similar approach can be done for $Q_0(x)$ and $Q_1(x)$. Thus, the CD for Gaussian distributions is defined as

$$CD = |\frac{\|\mu_0 - \mu_1\|_2^2}{8\sigma^2} - \frac{\|\zeta_0 - \zeta_1\|_2^2}{8\tau^2}|. \tag{37}$$

□

Noisy Chernoff Difference can be derived via a direct application of the normal-sum theorem (Lemons and Langevin, 2002).

Proposition C.2 (Detailed Restatement of Proposition 5.1). *Suppose $P_0(x) \sim \mathcal{N}(\mu_0, \sigma^2\mathbf{I})$, $P_1(x) \sim \mathcal{N}(\mu_1, \sigma^2\mathbf{I})$, $Q_0(x) \sim \mathcal{N}(\zeta_0, \tau^2\mathbf{I})$, and $Q_1(x) \sim \mathcal{N}(\zeta_1, \tau^2\mathbf{I})$. Without loss of generality, we assume that $\|\mu_0 - \mu_1\|_2 \geq \|\zeta_0 - \zeta_1\|_2$. There are three behaviors of the Noisy Chernoff Difference ($\widetilde{CD}_{\eta^2}$): (i) $\widetilde{CD}_{\eta^2}$ has a maximum point, (ii) $\widetilde{CD}_{\eta^2}$ has a maximum point and a reflection point (where $\widetilde{CD}_{\eta^2} = 0$), (iii) $\widetilde{CD}_{\eta^2}$ is non-increasing.[4] The respective conditions for these three cases are given as follows:*

(i) $\frac{\|\zeta_0-\zeta_1\|_2^2}{\|\mu_0-\mu_1\|_2^2} < \frac{\tau^2}{\sigma^2} < \frac{\|\zeta_0-\zeta_1\|_2}{\|\mu_0-\mu_1\|_2} < 1,$ ***(Case 1: Maximum Point)***

(ii) $\frac{\tau^2}{\sigma^2} < \frac{\|\zeta_0-\zeta_1\|_2^2}{\|\mu_0-\mu_1\|_2^2} < 1,$ ***(Case 2: Maximum and Reflection)***

(iii) Neither condition (i) or (ii) hold. ***(Case 3: Non-increasing)***

Proof. Recall the definition of noisy Chernoff Difference. We define a signed noisy Chernoff Difference $s\widetilde{\text{CD}}_{\eta^2}$ such that $|s\widetilde{\text{CD}}_{\eta^2}| = \widetilde{\text{CD}}_{\eta^2}$. Thus,

$$s\widetilde{\text{CD}}_{\eta^2} = \frac{1}{8(\tau^2 + \eta^2)}\|\zeta_0 - \zeta_1\|_2^2 - \frac{1}{8(\sigma^2 + \eta^2)}\|\mu_0 - \mu_1\|_2^2. \tag{38}$$

Case 1 and 2. Let $p = \|\mu_0 - \mu_1\|_2$ and let $q = \|\zeta_0 - \zeta_1\|_2$. First, we can analyze the positive η^2 regime for a critical point. To find potential critical points, consider the derivative of $s\widetilde{\text{CD}}_{\eta^2}$.

$$\frac{\partial}{\partial\eta^2} s\widetilde{\text{CD}}_{\eta^2} = \frac{p^2}{8(\sigma^2 + \eta^2)^2} - \frac{q^2}{8(\tau^2 + \eta^2)^2} = \frac{p^2(\tau^2 + \eta^2)^2 - q^2(\sigma^2 + \eta^2)^2}{8(\sigma^2 + \eta^2)^2(\tau^2 + \eta^2)^2} \tag{39}$$

Now, the potential critical points will be η^2 where $s\widetilde{\text{CD}}'_{\eta^2} = 0$. That is $0 = p^2(\tau^2 + \eta^2)^2 - q^2(\sigma^2 + \eta^2)^2$. Thus, the critical points are:

[4] When $\|\mu_0 - \mu_1\|_2 = \|\zeta_0 - \zeta_1\|_2$, $\widetilde{\text{CD}}_{\eta^2}$ will always fall into this case.

$$\eta_0^2 = \frac{\sigma^2 q - \tau^2 p}{p - q} \tag{40}$$

$$\eta_1^2 = \frac{-\sigma^2 q - \tau^2 p}{p + q} \tag{41}$$

However, we observe that η_1^2 is always negative, thus the only useful critical point in the postive η^2 region is η_0^2. By our assumption, we know that $p - q \geq 0$. First, we observe that there is no critical point when $p = q$ as η_0^2 does not exist. Thus, η_0^2 is positive when the following conditions hold.

$$\sigma^2 q - \tau^2 p > 0 \tag{42}$$

$$\frac{\tau^2}{\sigma^2} < \frac{\|\zeta_0 - \zeta_1\|_2}{\|\mu_0 - \mu_1\|_2} < 1 \tag{43}$$

Next, we will show that this critical point is always a maximum in the positive η^2 regime.

First, consider the second derivative of the signed noisy Chernoff difference.

$$\frac{\partial^2}{\partial(\eta^2)^2} s\widetilde{\mathrm{CD}}_{\eta^2} = \frac{q^2}{4(\tau^2 + \eta^2)^3} - \frac{p^2}{4(\sigma^2 + \eta^2)^3}. \tag{44}$$

Plugging in the relevant critical point we observe that

$$\frac{\partial^2}{\partial(\eta^2)^2} s\widetilde{\mathrm{CD}}_{\eta_0^2} = \frac{q^2}{4(\tau^2 + \eta_0^2)^3} - \frac{p^2}{4(\sigma^2 + \eta_0^2)^3} = \frac{q^2}{4(\frac{q(\sigma^2 - \tau^2)}{p-q})^3} - \frac{p^2}{4(\frac{p(\sigma^2 - \tau^2)}{p-q})^3} \tag{45}$$

$$= \frac{q^2(p-q)^3}{4q^3(\sigma^2 - \tau^2)^3} - \frac{p^2(p-q)^3}{4p^3(\sigma^2 - \tau^2)^3} = \frac{(p-q)^4}{4p^3q^3(\sigma^2 - \tau^2)^3}. \tag{46}$$

Now, we know that $\sigma^2 > \tau^2$ so the second derivative of this critical point of $s\widetilde{\mathrm{CD}}_{\eta^2}$ must be a minimum. However, the goal is to examine the behavior of $\widetilde{\mathrm{CD}}_{\eta^2}$. So, we can show that this is critical point is a maximum of $\widetilde{\mathrm{CD}}_{\eta^2}$, by showing that $s\widetilde{\mathrm{CD}}_{\eta_0^2}$ is negative. By plugging in the critical point, we can observe that it is a maximum of $\widetilde{\mathrm{CD}}_{\eta^2}$.

$$s\widetilde{\mathrm{CD}}_{\eta_0^2} = \frac{q^2}{8(\tau^2 + \eta_0^2)} - \frac{p^2}{8(\sigma^2 + \eta_0^2)} = \frac{q^2}{8(\tau^2 + \frac{\sigma^2 q - \tau^2 p}{p-q})} - \frac{p^2}{8(\sigma^2 + \frac{\sigma^2 q - \tau^2 p}{p-q})} \tag{47}$$

$$= \frac{(p-q)}{8}(\frac{q}{\sigma^2 - \tau^2} - \frac{p}{\sigma^2 - \tau^2}) = \frac{(p-q)(q-p)}{8(\sigma^2 - \tau^2)} \tag{48}$$

Now, this value is always negative as we know $p > q$ and $\sigma^2 > \tau^2$. Thus, the positive critical point is a maximum. Now, we can analyze the positive η^2 regime for a reflection point.

$$s\widetilde{\mathrm{CD}}_{\eta^2} = \frac{q^2}{8(\tau^2 + \eta^2)} - \frac{p^2}{8(\sigma^2 + \eta^2)} \tag{49}$$

$$8q^2(\sigma^2 + \eta^2) - 8p^2(\tau^2 + \eta^2) = 0 \tag{50}$$

$$\eta^2 = \frac{q^2\sigma^2 - p^2\tau^2}{p^2 - q^2} = \frac{\|\zeta_0 - \zeta_1\|_2^2\sigma^2 - \|\mu_0 - \mu_1\|_2^2\tau^2}{\|\mu_0 - \mu_1\|_2^2 - \|\zeta_0 - \zeta_1\|_2^2}. \tag{51}$$

Now, the denominator of this is always positive, thus, η^2 is positive when the following holds:

$$\|\zeta_0 - \zeta_1\|_2^2\sigma^2 - |\mu_0 - \mu_1\|_2^2\tau^2 > 0 \tag{52}$$

$$\frac{\tau^2}{\sigma^2} < \frac{\|\zeta_0 - \zeta_1\|_2^2}{\|\mu_0 - \mu_1\|_2^2} < 1. \tag{53}$$

Case 3. Finally, we can examine the behavior when none of these conditions hold. Suppose $\frac{\tau^2}{\sigma^2} \geq \frac{\|\zeta_0-\zeta_1\|_2}{\|\mu_0-\mu_1\|_2} = \frac{q}{p}$. We can analyze the sign of the numerator of $s\widetilde{\text{CD}}'_{\eta^2}$ by writing it as

$$p^2(\tau^2 + \eta^2)^2 - q^2(\sigma^2 + \eta^2)^2 \tag{54}$$

$$= (p^2(\tau^2 + \eta^2) - q^2(\sigma^2 + \eta^2))(p^2(\tau^2 + \eta^2) + q^2(\sigma^2 + \eta^2)). \tag{55}$$

Thus, to analyze the sign, we can analyze $p^2(\tau^2 + \eta^2) - q^2(\sigma^2 + \eta^2)$. We observe

$$p^2(\tau^2 + \eta^2) - q^2(\sigma^2 + \eta^2) = p^2\tau^2 - q^2\sigma^2 + \eta^2(p^2 - q^2). \tag{56}$$

Now, from our initial assumption, $p^2 - q^2 \geq 0$. From the assumption that $\frac{\tau^2}{\sigma^2} \geq \frac{q}{p}$, $p^2\tau^2 - q^2\sigma^2 \geq 0$. Thus, the sign is always positive. Now, to show that $\widetilde{\text{CD}}_{\eta^2}$ is non-increasing, we will show $s\widetilde{\text{CD}}_{\eta^2}$ is not positive.

$$s\widetilde{\text{CD}}_{\eta^2} = \frac{q^2}{8(\tau^2 + \eta^2)} - \frac{p^2}{8(\sigma^2 + \eta^2)} = \frac{q^2(\sigma^2 + \eta^2) - p^2(\tau^2 + \eta^2)}{8((\tau^2 + \eta^2))(\sigma^2 + \eta^2)} \tag{57}$$

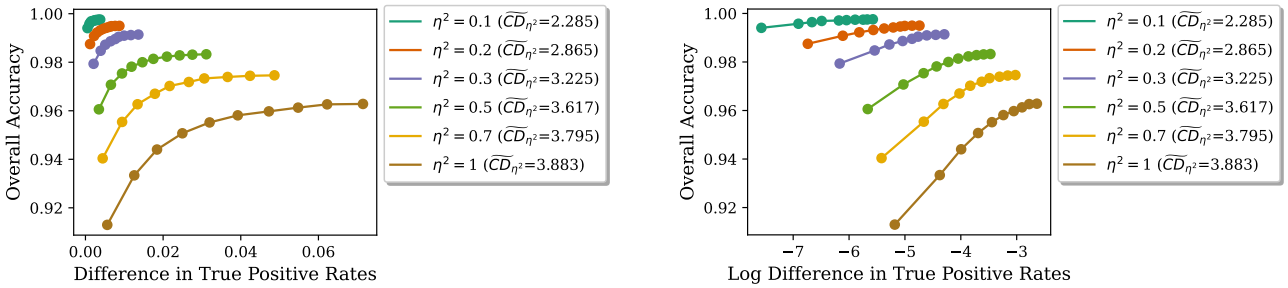
Now, we analyze $q^2(\sigma^2 + \eta^2) - p^2(\tau^2 + \eta^2)$. We can see that this is equivalent to $q^2\sigma^2 - p^2\tau^2 + \eta^2(q^2 - p^2)$. From our initial assumption, $q^2 - p^2 \leq 0$. From the assumption that $\frac{\tau^2}{\sigma^2} \geq \frac{q}{p}$, $q^2\sigma^2 - p^2\tau^2 \leq 0$. Thus, $s\widetilde{\text{CD}}_{\eta^2}$ is always negative and $\widetilde{\text{CD}}_{\eta^2}$ is non-increasing over the positive η^2 regime. □

D. Supplemental Gaussian Figures

D.1. Experimental Details

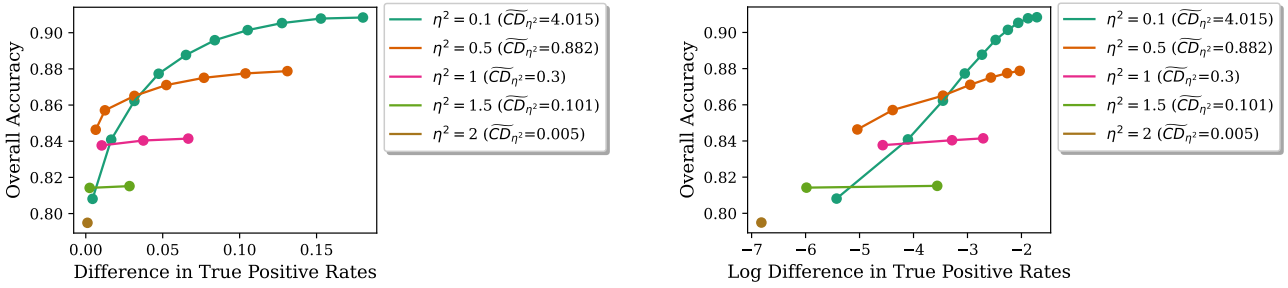
For the synthetic Gaussian experiments, we use 1-D Gaussian distributions for $P_0(x), P_1(x), Q_0(x)$, and $Q_1(x)$ and get 100,000 i.i.d. samples from each distribution to create a balanced dataset. We train a Gaussian Naïve Bayes classifier for each of the groups (this provides the Bayes optimal classifier). We measure the overall accuracy of our classifier and quantify fairness by using the true positive rate disparity between groups (as this corresponds to the dominating exponent). To achieve fairness-accuracy trade-offs, we choose different prior beliefs of our labels for the Gaussian Naïve Bayes classifier. By perturbing the prior probabilities for the unprivileged group, we create fairness-accuracy curves for each of our settings[5].

D.2. Fairness-Accuracy vs Log Fairness-Accuracy



(a) *(b)*

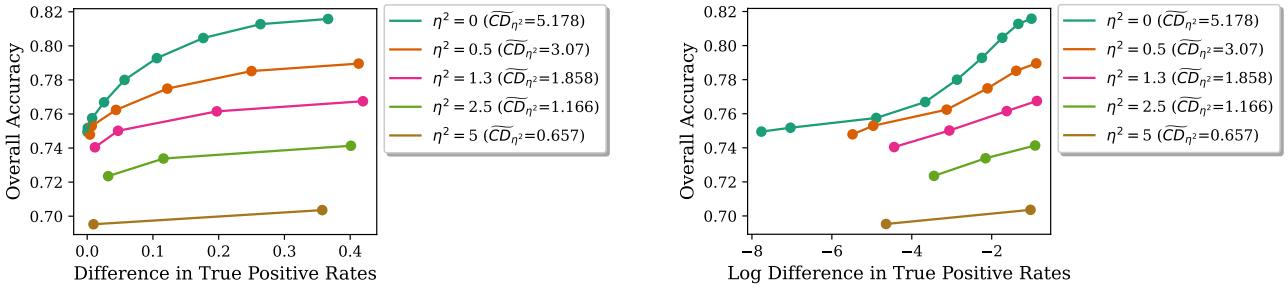
Figure D.1. **(Case 1: Privacy Hurts Fairness)** $\mu_0 = 0, \mu_1 = 16.5, \sigma = 2.43, \zeta_0 = 0.5, \zeta_1 = 3.8, \tau = 0.55$. (a) Fairness-Accuracy Curve. (b) Log Fairness-Accuracy Curve. We observe a steepening effect.



(a) *(b)*

Figure D.2. **(Case 2: Privacy Can Give Free Fairness)** $\mu_0 = -4.2, \mu_1 = 1.3, \sigma = 3, \zeta_0 = 0.3, \zeta_1 = 2.7, \tau = 0.25$. (a) Fairness-Accuracy Curve. (b) Log Fairness-Accuracy Curve. We observe a flattening effect.

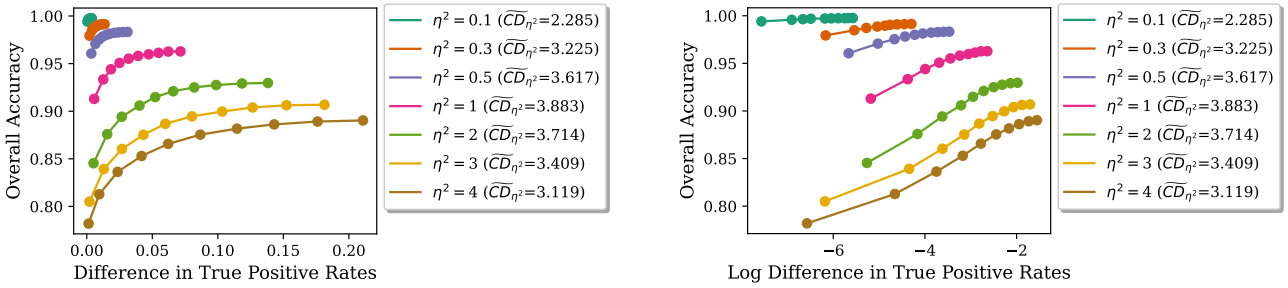
[5] A proxy for adjusting the threshold of the Bayes optimal classifier (Dutta et al., 2020).



(a) *(b)*

Figure D.3. **(Case 3: Triple Trade-off)** $\mu_0 = -4.2, \mu_1 = 1.3, \sigma = 0.85, \zeta_0 = 0.6, \zeta_1 = 1.6, \tau = 0.6$. (a) Fairness-Accuracy Curve. (b) Log Fairness-Accuracy Curve. We observe a much slower flattening effect.

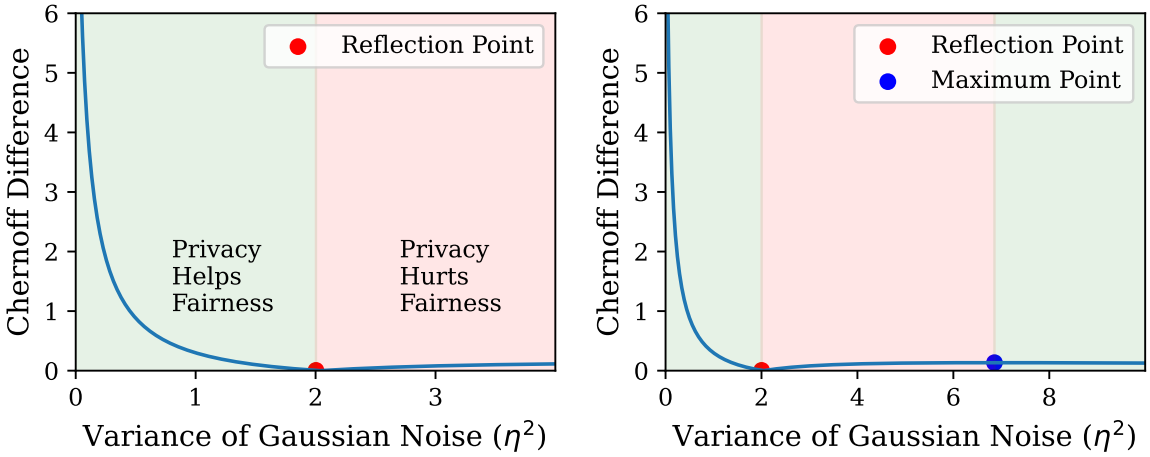
D.3. Case 1: Extended plots



(a) *(b)*

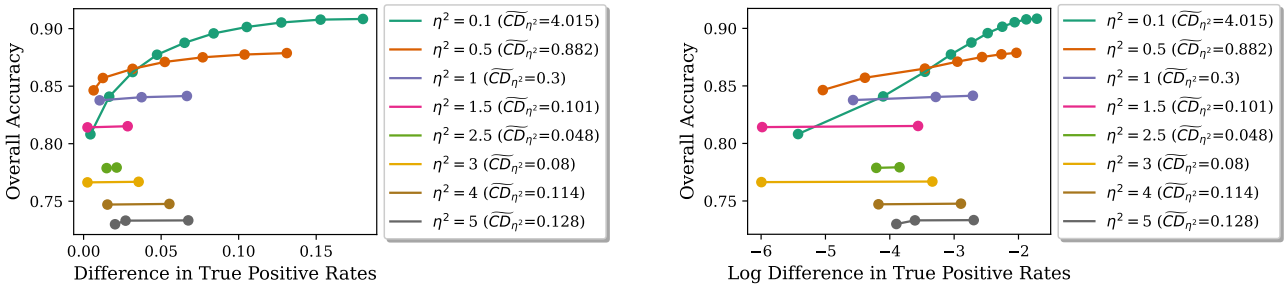
Figure D.4. **(Case 1: Extension)** $\mu_0 = 0, \mu_1 = 16.5, \sigma = 2.43, \zeta_0 = 0.5, \zeta_1 = 3.8, \tau = 0.55$ (a) Fairness-Accuracy Curve. (b) Log Fairness-Accuracy Curve. After reaching the maximum CD, the fairness-accuracy curves begin to slowly flatten (nearly imperceptible), reflecting the slow decay in CD.

D.4. Case 2: More detail



(a) *(b)*

Figure D.5. **(Case 2: Privacy Can Give Free Fairness)** $\mu_0 = -4.2, \mu_1 = 1.3, \sigma = 3, \zeta_0 = 0.3, \zeta_1 = 2.7, \tau = 0.25$. (a) Initial plot of $\widehat{\text{CD}}_{\eta^2}$. (b) Full plot showing presence of maximum point.



(a) *(b)*

Figure D.6. **(Case 2: Extension)** (a) Fairness-Accuracy Curve. (b) Log Fairness-Accuracy Curve. After reaching the reflection point, CD begins to increase. However this spike is so small any change in slope in the fairness-accuracy curve is imperceptible.

E. Consistency

E.1. Setting

Let $r(x) = \frac{P_1(x)}{P_0(x)}$ be the density ratio and let $f(u) = -\log \mathbb{E}_{x\sim P_0}[r(x)^u]$. We also define the corresponding empirical estimators $r_n(x)$ and $f_n(u) = -\log \frac{1}{n}\sum r_n(x_i)^u$. We operate under the assumption that P_0 and P_1 share a support, $r(x)$ is bounded i.e., $r(x) \in [c, C]$ for some constants $0 < c < C < \infty$.

We further assume that $r_n(x)$ is a consistent estimator of $r(x)$ and is bounded via a clipping mechanism i.e., $r_n(x) \in [c, C]$. Clipping is typically applied after estimation to prevent numerical instability and prevent estimates from growing too large or small. This clipping does not affect consistency under our assumptions if the clipping bound is wider than the true bounds on the density ratio. Finally, we assume that the estimator $r_n(x)$ is trained on its own samples, independent of the evaluation points.

E.2. Proofs

We will use Newey-McFadden's main consistency theorem to show consistency of the Chernoff Information estimator.

Theorem E.1 (**Newey-McFadden Main Consistency Theorem**, Theorem 2.1 (Newey and McFadden, 1994))**.** *If there is a function f such that (i) $f(u)$ is uniquely maximized at u^*; (ii) U is compact; (iii) f is continuous; (iv) $f_n(u)$ converges uniformly in probability to $f(u)$, then $\hat{u} \to u^*$.*

We begin with a series of lemmas before completing the proof of our theorem. First, we are able to show compactness via:

Lemma E.2. *We can write Chernoff Information as* $-\inf_{[0,1]} \log(\int P_0(x)^{1-u} P_1(x)^u dx)$.

Proof. Computing the infimum over the open interval (0,1) and closed interval [0,1] yield the same result following the strict convexity of f. We point to Remark 2 from Nielsen (2022) for a discussion of the boundary conditions. In fact, in some references, Chernoff Information is presented using this closed interval (Cover and Thomas, 1999). This provides a compactness property that can be utilized in the following lemmas. □

Lemma E.3. $\{f_n\}$ *converges pointwise in probability for* $u \in [0, 1]$

Proof. First, we see that we can decompose the following difference into two terms,

$$\frac{1}{n}\sum r_n(x_i)^u - \mathbb{E}_{x\sim P_0}[r(x)^u] = (\frac{1}{n}\sum r_n(x_i)^u - \mathbb{E}_{x\sim P_0}[r_n(x)^u]) + (\mathbb{E}_{x\sim P_0}[r_n(x)^u] - \mathbb{E}_{x\sim P_0}[r(x)^u]) \tag{58}$$

Next, we can see that the first term $\frac{1}{n}\sum r_n(x_i)^u - \mathbb{E}_{x\sim P_0}[r_n(x)^u]$ converges to 0 almost surely (a.s.) by the law of large numbers.

Since r_n is consistent and uniformly bounded in $[c, C]$, there exists $\widetilde{r}_n(x)^u$ and $\widetilde{r}(x)^u$ defined on another space such that $\widetilde{r}_n(x)^u$ has the same distribution as $r_n(x)^u$, $\widetilde{r}(x)^u$ has the same distribution as $r(x)^u$, and $\widetilde{r}_n(x)^u$ converges to $\widetilde{r}(x)^u$ a.s.

(via the Continuous Mapping Theorem and Skorohod's Representation Theorem). The Dominated Convergence Theorem then implies $\mathbb{E}[\widetilde{r}_n(x)^u] - \mathbb{E}[\widetilde{r}(x)^u]$ converges to 0. Since the two sequences share the same distributions, convergence in the Skorokhod space implies convergence in the original space. It follows that $\mathbb{E}[r_n(x)^u] - \mathbb{E}[r(x)^u]$ converges to 0.

Since both terms converge to 0, their sum converges a.s. to 0 and hence in probability to 0. Thus, we have $\frac{1}{n}\sum r_n(x_i)^u - \mathbb{E}[r(x)^u] \to 0$ in probability. By applying the continuous mapping theorem, we have $-\log(\frac{1}{n}\sum r_n(x_i)^u) + \log(\mathbb{E}[r(x)^u]) \to 0$. Thus, we have $\{f_n\}$ converges pointwise in probability. □

Lemma E.4. *$\{f_n\}$ is stochastically equicontinuous on $[0,1]$*

Proof. First, recall the density ratio $r(x)$ is bounded between $[c, C]$. Via clipping we assume the estimator $r_n(x)$ follows the same bound. Now, for any x, the function $u \to r_n(x)^u$ has a bounded derivative, $r_n(x)^u \log r_n(x)$, that only depends on constants c and C, thus is Lipschitz in u. It follows that the average

$$\frac{1}{n}\sum_{i=1}^{n} r_n(x_i)^u$$

is Lipschitz in u again with the constant dependent only on c and C. Now, since $-\log(\cdot)$ is Lipschitz on the relevant range (away from 0), it follows from the compositional property of Lipschitz functions that

$$f_n(u) = -\log\left(\tfrac{1}{n}\sum_{i=1}^{n} r_n(x_i)^u\right)$$

is Lipschitz in u, with constant only dependent on c and C. Thus, we have a uniform Lipschitz bound across all elements of the sequenece $\{f_n\}$. It follows that $\{f_n\}$ is stochastically equicontinuous on $[0,1]$. □

Lemma E.5. *$\{f_n\}$ converges uniformly in probability for $u \in [0,1]$*

Proof. This follows from Theorem 1 from Newey (1991) as well as pointwise convergence in Lemma E.3 and stochastic equicontinuity in Lemma E.4. □

Finally, using these lemmas, we can complete the proof of our theorem.

Theorem E.6 (Restatement of Theorem 2)**.** *Let P_0 and P_1 be distributions with common support and bounded density ratio. Then the Chernoff Information estimator, constructed using a consistent density ratio estimator, is itself consistent.*

Proof. This result follows from Theorem E.1. Continuity and the uniqueness of the maximizer are given by Nielsen (2022). Compactness is given by Lemma E.2. Uniform Convergence in probability is given by Lemma E.5. Hence,

$$\widehat{\text{CI}} \to \text{CI}. \tag{59}$$

Therefore, we have consistency of the Chernoff Information estimator. □

F. Experiment Details

F.1. Estimation

To implement the density ratio estimation from Choi et al. (2022), we utilize the codebase provided by the authors[6]. Following the success in their paper, we choose to implement the "Joint" model where the model, in addition to time scores, aims to recover data scores. For our interpolation technique, we utilize $p_\lambda(x) = \lambda P_0(x) + \sqrt{1-\lambda^2}P_1(x)$. For the model architecture, we utilize the Joint architecture proposed in Appendix F of Choi et al. (2022) which leverages multilayer perceptrons with Exponential Linear Unit (ELU) activations as the modules:

1. **Joint (Shared)**:

$$\text{Linear}(3, 256) \;\to\; \text{ELU} \;\to\; \text{Linear}(256, 512) \;\to\; \text{chunk}(2)$$

[6] https://github.com/ermongroup/dre-infinity

(a) **Time Module**:

$$\text{Linear}(256, 256) \rightarrow \text{ELU} \rightarrow \text{Linear}(256, 256) \rightarrow \text{ELU} \rightarrow \text{Linear}(256, 1)$$

(b) **Data Module**:

$$\text{Linear}(256, 256) \rightarrow \text{ELU} \rightarrow \text{Linear}(256, 256) \rightarrow \text{ELU} \rightarrow \text{Linear}(256, 2).$$

We utilize a Cosine Annealing learning rate scheduler with an initial learning rate of 1e-5 and a batch size of 128. We train all models for 30,000 steps with no importance weighting. All other parameters are standard to the original implementation. All models are trained using NVIDIA L40S GPUs. We highlight an ablation study over hyperparameters in Section J. For the Monte Carlo method, we utilize all available data; we find that this yields more accurate estimation. To solve the infimum, we utilize Brent's algorithm, although we highlight that this is not required following Lemma 4.2 (any convex optimization method would suffice).

F.1.1. Time Complexity

The overall time complexity of our Chernoff Information estimator can be given as:

$$T_{\text{total}} = T_{\text{DRE}} + n \cdot C_{\text{int}} + n \cdot C_{\text{MC}} + T_{\text{cvx}} \tag{60}$$

where n is the number of samples used to compute Chernoff Information, T_{DRE} is time spent on density ratio estimation (DRE) in Line 1, C_{int} is the time required for one integral operations given in Line 2, C_{MC} is the time required for simple addition/floating point operations required in Line 3, and T_{cvx} is the time required to solve the convex optimization in Line 4, which is a constant time operation. Hence, for other than T_{DRE}, the rest of the operation is done in $O(n)$ and training a neural network for DRE is often the most time-consuming part, which took approximately 30 minutes on a single NVIDIA L40S GPU.

F.2. Gaussian Estimation

For all Gaussian experiments, we estimate Chernoff information using the parameters and algorithm specified in F.1. We work with 10,000 samples to perform our estimation. First, to compare to closed form Chernoff Information, we create 2D Gaussians $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and $\mathcal{N}(\mathbf{1}, \sigma^2\mathbf{I})$ and vary the parameter σ^2. For each σ^2, we perform 5 estimates to create error bars.

For more complex 5D Gaussian distributions, we compare to the algorithm provided by Nielsen (2022). We implement the algorithm directly from the paper and use a stopping criteria of 1e-7. We compute Chernoff Information across different levels of noise added to $\mathcal{N}(\mathbf{0}, \frac{1}{2}\mathbf{I})$ and $\mathcal{N}(\mathbf{1}, \mathbf{I})$. We repeat each trial 5 times to obtain error bars.

For the final experiment, we repeat the previous example; however instead of adding noise, we scale up the dimension. We again compare to Gaussian estimation algorithm for distributions $\mathcal{N}(\mathbf{0}, \frac{1}{2}\mathbf{I})$ and $\mathcal{N}(\mathbf{1}, \mathbf{I})$ and repeat each trial 5 times to obtain error bars.

F.3. Datasets

Adult. For the Adult dataset, we select the continuous features, "age", "fnlwgt", "education-num", "capital-gain", "capital-loss", "hours-per-week", for our experiments. We apply a standard scaler, transforming the distributions so that they are centered at 0 with standard deviation of 1. We consider the standard income classification task and use Sex as the sensitive attribute. To obtain Chernoff Information estimates, we train our model as specified in Section F.1. We train each estimate 5 times to create error bars. We compute Chernoff Difference from these Chernoff Information values.

For fairness-accuracy curves, we train split logistic regression models using an 80/20 train-test split for each model. We fix the model for the dominating group (Male) and sweep class weights for the other group, dropping non-Pareto optimal points. We find that the False Negative Rate disparity dominates and thus utilize that in our plot.

HSLS. Following Jeong et al. (2022), we select a subset of features from the HSLS dataset. We further restrict to continuous features, giving: "X1MTHID","X1MTHUTI", "X1MTHEFF", "X1FAMINCOME", "X1SCHOOLBEL". We apply a standard scaler, giving distributions centered at 0 with standard deviation of 1. For our sensitive attribute create

groups using the Race attribute, giving Asian/White and URM (Other). We utilize the "X1TXMSCR" column to create a binary classification task (top 50th percentile or not). For Chernoff Information and fairness-accuracy curves, we replicate the experiments we conduct on the Adult dataset. We find that the False Positive Rate disparity dominates and thus utilize that in our plot.

MNIST We replicate the previous experiments using 2D embeddings of MNIST data which were obtained via a trained autoencoder with latent dimension $d_z = 30$.

Table 1. Convolutional autoencoder architecture used for MNIST experiments.

Layer	Kernel / Units	Stride	Output Shape
Input	–	–	$1 \times 28 \times 28$
Conv2D + ReLU	3×3, 32 ch.	2	$32 \times 14 \times 14$
Conv2D + ReLU	3×3, 64 ch.	2	$64 \times 7 \times 7$
Flatten	–	–	$64 \cdot 7 \cdot 7$
Linear	$d_z = 30$	–	30
Linear	$64 \cdot 7 \cdot 7$	–	$64 \times 7 \times 7$
ConvTranspose2D + ReLU	4×4, 32 ch.	2	$32 \times 14 \times 14$
ConvTranspose2D + Sigmoid	4×4, 1 ch.	2	$1 \times 28 \times 28$

We train the autoencoder for 15 epochs, with a batch size of 256, using the Adam optimizer and mean squared error loss function. We choose digit 3 for P_0, digit 4 for P_1, digit 7 for Q_0, and digit 9 for Q_1.

Dataset Licenses All datasets used in this work (Adult, HSLS, MNIST) are publicly available and approved for non-commercial research use under their respective data use agreements..

F.4. Mixture of Gaussian

For mixture of Gaussian experiments, we design the distributions as follows. We work with 10,000 samples from each group for our experiments. We utilize the Chernoff Information estimator with the same parameters from Section F.1 and obtain the fairness accuracy curves by replicating the experimental procedure from Section 5.1.

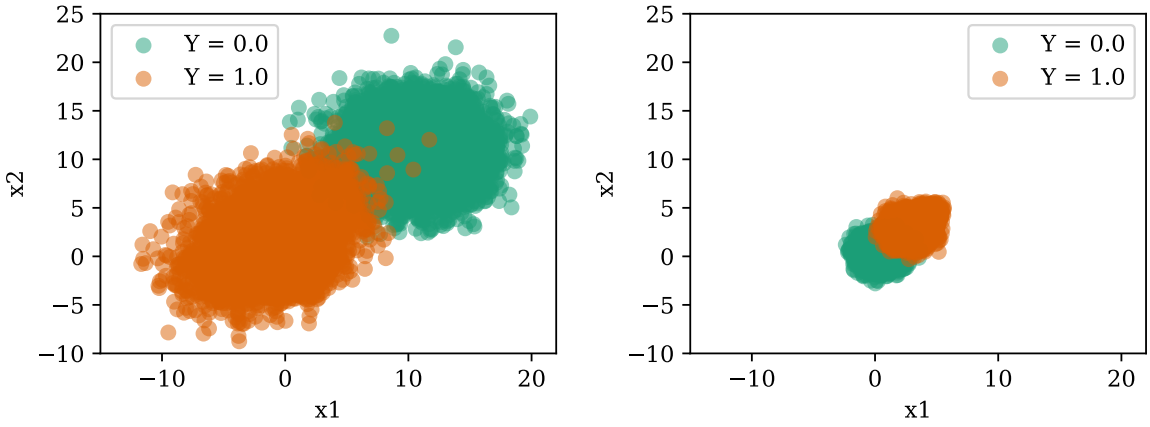
Mixture 1

$$P_0 = \tfrac{1}{2}\mathcal{N}\left(\begin{bmatrix}10.5\\10.5\end{bmatrix}, 6.56\,\mathbf{I}\right) + \tfrac{1}{2}\mathcal{N}\left(\begin{bmatrix}10.8\\11.2\end{bmatrix}, 6.43\,\mathbf{I}\right), \tag{61}$$

$$P_1 = \tfrac{1}{2}\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, 1.89\,\mathbf{I}\right) + \tfrac{1}{2}\mathcal{N}\left(\begin{bmatrix}-1\\2\end{bmatrix}, \begin{bmatrix}9.43 & 3.3\\3.3 & 9.43\end{bmatrix}\right), \tag{62}$$

$$Q_0 = \tfrac{1}{2}\mathcal{N}\left(\begin{bmatrix}0.5\\0.5\end{bmatrix}, 0.55\,\mathbf{I}\right) + \tfrac{1}{2}\mathcal{N}\left(\begin{bmatrix}0.5\\0.5\end{bmatrix}, 0.45\,\mathbf{I}\right), \tag{63}$$

$$Q_1 = \tfrac{1}{2}\mathcal{N}\left(\begin{bmatrix}3.8\\3.8\end{bmatrix}, \begin{bmatrix}0.2 & 0.1\\0.1 & 0.2\end{bmatrix}\right) + \tfrac{1}{2}\mathcal{N}\left(\begin{bmatrix}2.8\\2.8\end{bmatrix}, 0.55\,\mathbf{I}\right). \tag{64}$$



(a) Group P *(b)* Group Q

Figure F.1. **(Mixture 1)**

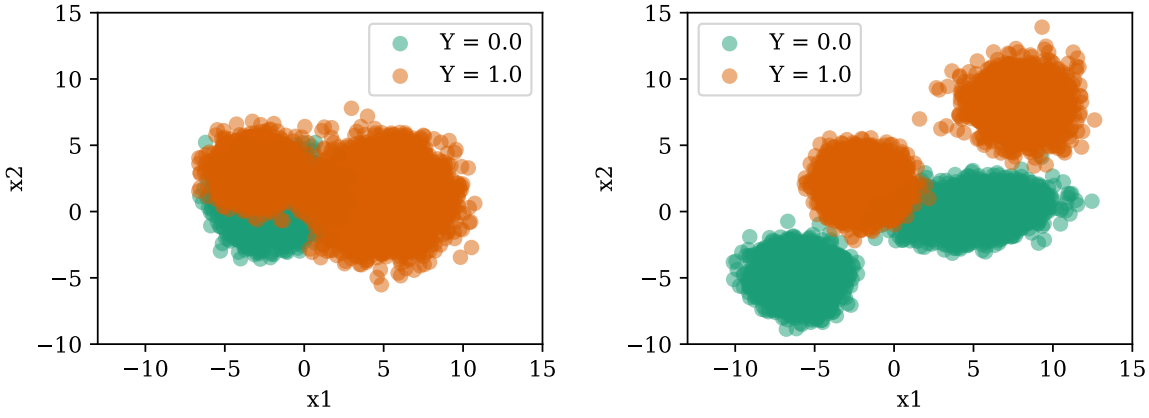
Mixture 2

$$P_0 = \tfrac{1}{2}\mathcal{N}\left(\begin{bmatrix}0.2\\0.1\end{bmatrix}, 0.2\,\mathbf{I}\right) + \tfrac{1}{2}\mathcal{N}\left(\begin{bmatrix}0.5\\0.2\end{bmatrix}, \begin{bmatrix}0.25 & 0.1\\0.1 & 0.25\end{bmatrix}\right), \tag{65}$$

$$P_1 = \tfrac{1}{2}\mathcal{N}\left(\begin{bmatrix}3.1\\2.4\end{bmatrix}, \begin{bmatrix}0.22 & 0.19\\0.19 & 0.22\end{bmatrix}\right) + \tfrac{1}{2}\mathcal{N}\left(\begin{bmatrix}2.9\\3.2\end{bmatrix}, 0.24\,\mathbf{I}\right), \tag{66}$$

$$Q_0 = \tfrac{1}{2}\mathcal{N}\left(\begin{bmatrix}-4.2\\-5.2\end{bmatrix}, 9\,\mathbf{I}\right) + \tfrac{1}{2}\mathcal{N}\left(\begin{bmatrix}-6.2\\-3.2\end{bmatrix}, \begin{bmatrix}9 & 5\\5 & 9\end{bmatrix}\right), \tag{67}$$

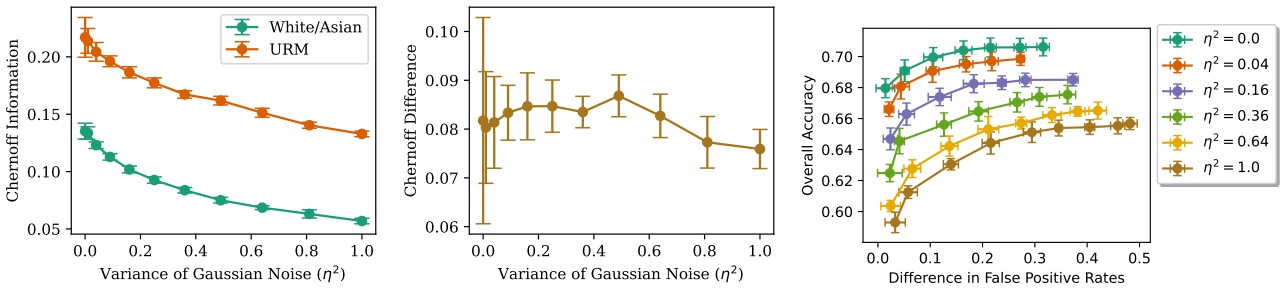
$$Q_1 = \tfrac{1}{2}\mathcal{N}\left(\begin{bmatrix}1.3\\1.3\end{bmatrix}, \begin{bmatrix}9 & 4\\4 & 9\end{bmatrix}\right) + \tfrac{1}{2}\mathcal{N}\left(\begin{bmatrix}1.3\\1.3\end{bmatrix}, 8\,\mathbf{I}\right). \tag{68}$$



(a) Group P *(b)* Group Q

Figure F.2. **(Mixture 2)**

G. HSLS



(a) *(b)* *(c)*

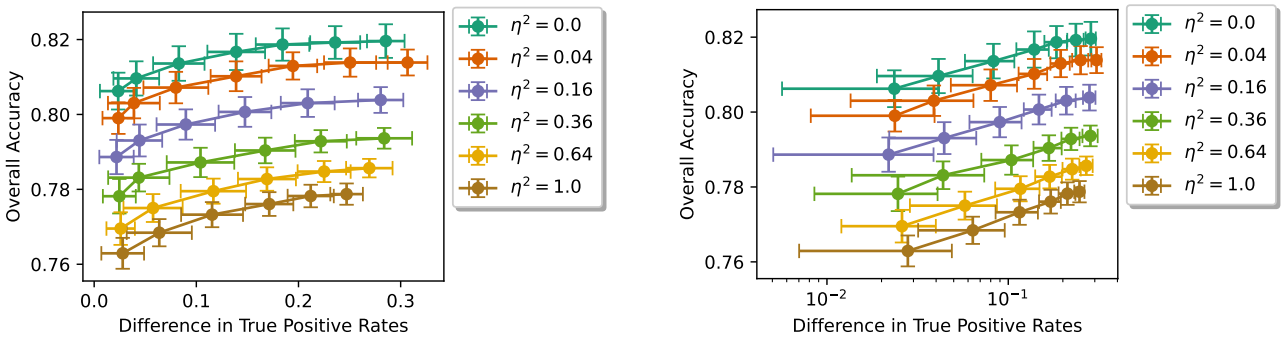
Figure G.1. **(HSLS Experiments)** (a) Chernoff Information for White/Asian and URM groups decrease as noise (η^2) increases. (b) Chernoff Difference values are larger with a small increase but still remain flat (Case 1). (c) Following the low Chernoff Difference values, are steeper and grow slightly more steep as noise is added.

HSLS. We examine the behavior of Chernoff Information on the HSLS dataset (Ingels et al., 2011). Following Jeong et al. (2022), we select a subset of the features (further filtering to include only continuous features) to perform a binary classification task (top 50th percentile mathematics test score). This gives a 5 dimensional dataset on which we perform our analysis. For the binary groups, we split the data into two groups based on Race, Asian/White and Under Represented Minorities (URM). We follow a similar approach to the Adult dataset to create the Chernoff Difference plots. Additionally, we follow a similar approach to generate fairness-accuracy plots, however, we find that the false positive rate dominates the error and the Chernoff Exponent, using that as our fairness metric.

Again, the Chernoff Information decays as noise is added, however as opposed to the Adult dataset, there is a much larger disparity between the value across groups. This is reflected as we see much larger values of Chernoff Difference in Figure G.1b and steeper fairness accuracy curves (Figure G.1c).

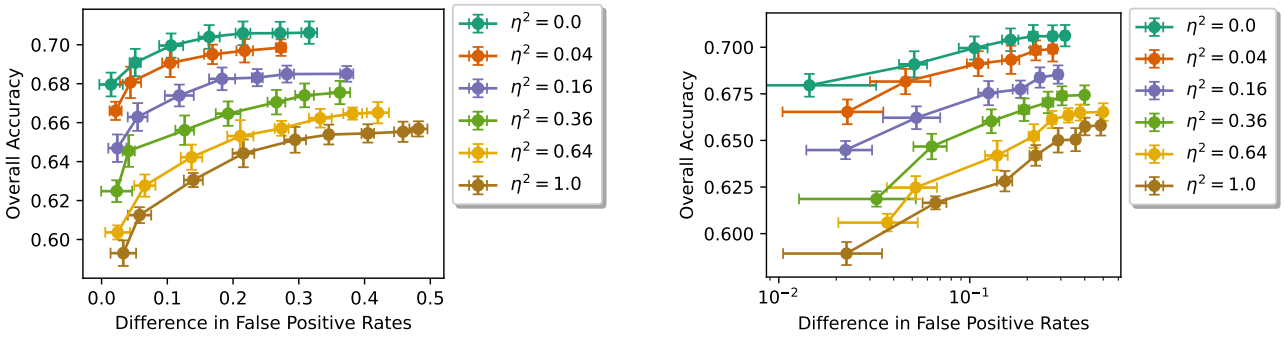
Similar to the Adult dataset, we observe the Chernoff Difference remains relatively stable as the variance of noise (η^2) increases, however, we do note a small increase before the Chernoff Difference begins to slightly decrease (indicative of Case 1). We see that this trend is somewhat reflected, by a very subtle steepening of the fairness accuracy curve (we point the reader to Appendix H for the log-fairness accuracy curve where this very subtle trend can be more easily examined). These observations could resemble Case 1, however, we point out that this trend is very subtle and for the most part, the slopes remain flat. Overall, these experiments suggest that in this dataset, privacy and fairness may be slightly less compatible, although the effect is minor.

H. Supplemental Real Data Figures



(a) *(b)*

Figure H.1. **(Adult Dataset)** (a) Fairness-Accuracy Curve. (b) Log Fairness-Accuracy Curve. We find that the curves remain very stable, reflected by the small changes in Chernoff Difference.



(a) *(b)*

Figure H.2. **(HSLS Dataset)** (a) Fairness-Accuracy Curve. (b) Log Fairness-Accuracy Curve. We observe a very slight steepening of the fairness accuracy curves, further emphasized by the log-fairness accuracy curves.

I. MNIST Representations

Remark I.1 (Curse of Dimensionality). Under standard Hölder smoothness assumptions with parameter s in d dimensions, the minimax sample complexity scales as $n \asymp \varepsilon^{-(2s+d)/s}$ (Singh and Póczos, 2016), reflecting the curse of dimensionality.

Here we discuss the ability of reasoning about the input via Chernoff Information in the representation space. We begin with a remark centered around the preservation of Chernoff Information in the latent space.

Remark I.2 (CI under representation learning). For any bijection $\phi : \mathcal{X} \to \mathcal{Z}$, $C(P_0, P_1) = C(\phi_\# P_0, \phi_\# P_1)$, where $\phi_\# P$ denotes the pushforward of P under ϕ. Chernoff Information is invariant under reparameterization: compressing into a latent space therefore preserves CI as long as the encoder is near-bijective on the data manifold.

The manifold hypothesis (Pope et al., 2021) reports MNIST's intrinsic dimensionality at ≈ 13, well below our 30D latent, suggesting our autoencoder can learn a near-bijective transformation. We can further analyze this by examining autoencoder reconstructions.

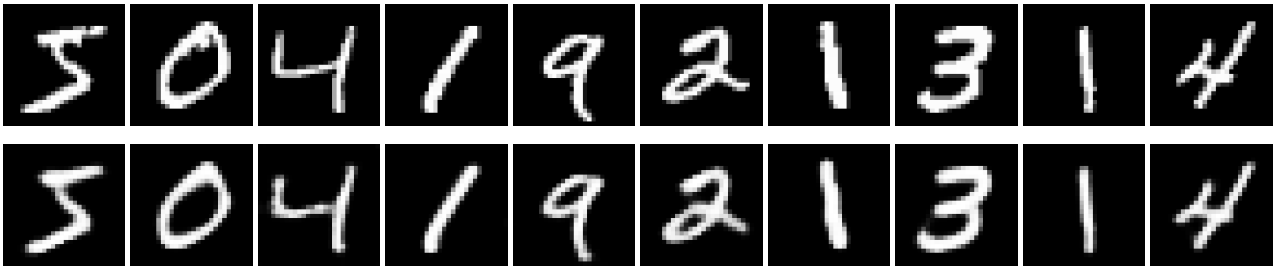


Figure I.1. **Autoencoder reconstructions on MNIST.** Top row shows original inputs, while the bottom row shows reconstructions from the learned 30-dimensional latent space. The high visual fidelity indicates that the learned representation preserves most of the input structure.

Here we see nearly identical MNIST reconstructions, providing further evidence that the transformation learned by the autoencoder is nearly bijective.

J. Ablation Study

All experiments are performed on 5D Gaussians $\mathcal{N}(\mathbf{0}, \frac{1}{2}\mathbf{I})$ and $\mathcal{N}(\mathbf{1}, \mathbf{I})$, with the standard setup mentioned in Section F.1 unless otherwise stated. All plots are Chernoff Information estimates across steps of the training procedure for the density ratio estimation.

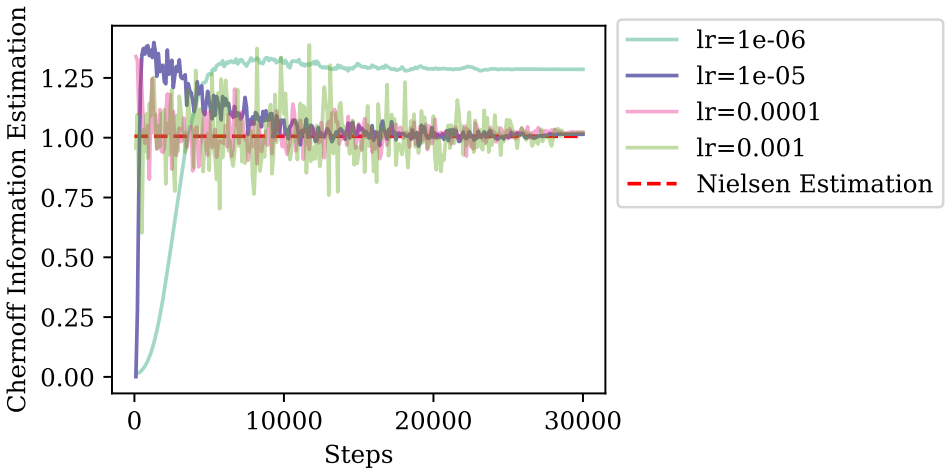


Figure J.1. **(Learning Rate)**: Comparison of Chernoff Information Estimation for different learning rates. We find learning rate is the most sensitive hyperparameter and may converge to an incorrect value if too small or be unstable if too large. We explore a selection of learning rates and find that 1e-5 works well under Gaussian and tabular settings.

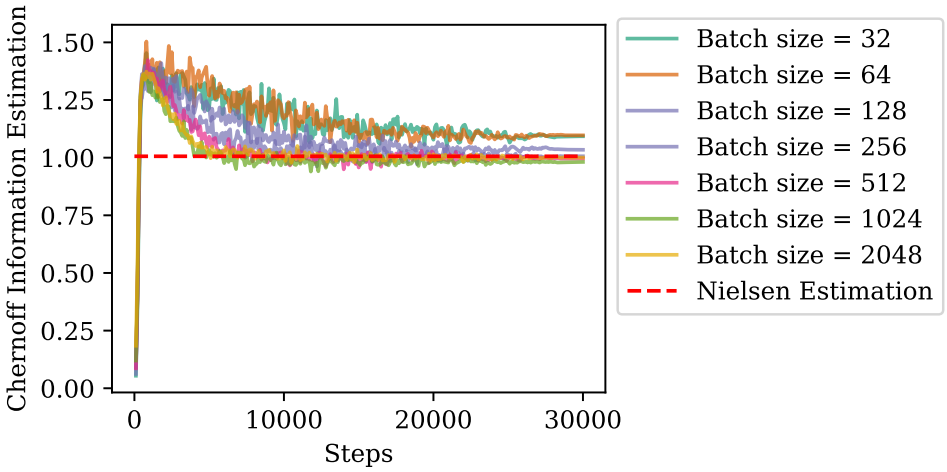


Figure J.2. **(Batch Size)** Comparison of Chernoff Information Estimation for different batch sizes. We find that batch size has a more subtle effect on the estimation outcome. Larger batch sizes tend to be slightly more beneficial for estimation. However, the sample sizes of the tabular datasets we use are smaller than the synthetic Gaussian experiments, causing us to scale the batch size down to 128 for our experiments. These larger batch sizes lead to explosion of the density ratio estimates in tabular settings. We find that this slightly smaller value performs well on both the Gaussian and Tabular settings and prevents the density ratio from exploding in the later case.

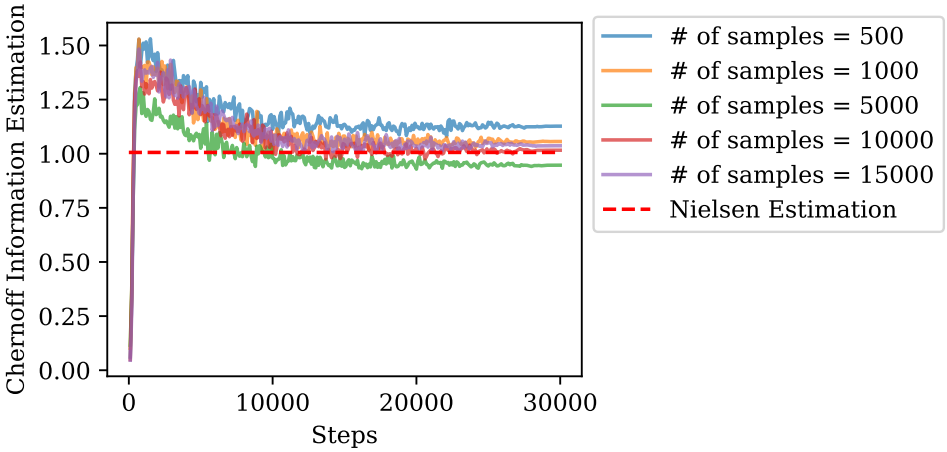


Figure J.3. **(Sample Size)** Comparison of Chernoff Information Estimation for sample sizes. We find that even down to 1000 samples (sample sizes similar to the smallest group in the tabular datasets), the estimation remains fairly accurate. When there are not enough samples, the estimate begins to degrade, as we can begin observing with 500 samples. This sample size and dimension relationship is further highlighted in Figure 2c.