CAN WE CATCH THE TWO BIRDS OF FAIRNESS AND PRIVACY?

Arjun Nichani[1], Hsiang Hsu[2], Haewon Jeong[1]
[1] University of California, Santa Barbara
[2] JP Morgan Chase Global Technology Applied Research
{anichani, haewon}@ucsb.edu
hsiang.hsu@jpmchase.com

ABSTRACT

Fairness and privacy are two vital pillars of trustworthy machine learning. Despite extensive research on these individual topics, the relationship between fairness and privacy has garnered significantly less attention. In this paper, we investigate the relationship between fairness, privacy, and accuracy using the information theoretic concept of Chernoff Information. We define Chernoff Difference, a tool that allows us to analyze the relationship between fairness, privacy, and accuracy. We then show that for Gaussian distributions, this value behaves in 3 distinct ways (depending on the distribution of the data). We highlight the distributions that these cases entail as well as their fairness and privacy implications. Additionally, we show that Chernoff difference acts as a proxy to the steepness of the fairness-accuracy curves. This work provides a foundation towards a more comprehensive understanding of the relationship between fairness, privacy and accuracy in machine learning and motivates the use of Chernoff information in this domain.

1 INTRODUCTION

Fairness and privacy are two fundamental components of trustworthy machine learning (ML) systems. The problem of ML models adopting detrimental social biases is prevalent across numerous fields, including facial recognition (Garvie & Frankle, 2016), text-to-image models (Friedrich et al., 2023; Bianchi et al., 2023), and various predictive tasks like recidivism (Barocas & Selbst, 2016; Chouldechova, 2016) and loan approval predictions (Das et al., 2021). Similarly, machine learning models have been shown to reveal information about their training data (Shokri et al., 2017; Carlini et al., 2021) creating significant privacy concerns (Gomstyn & Jonker, 2024).

Although these topics have been actively investigated in isolation (Caton & Haas, 2020; Mehrabi et al., 2022; Pessach & Shmueli, 2022; Mireshghallah et al., 2020), the interaction between these two concepts has received significantly less attention. This can create issues in real world systems where we want to apply ML. One such example of this is the infamous credit limit bias for the Apple Card. In this case, users reported that smaller lines of credit were provided to women compared to men, sparking large controversy surrounding the fairness of the credit limit decision process (Knight, 2019). The Apple Card, however, was designed to put "privacy and security first" (NYDFS, 2021). This poses a critical question: *When are privacy and fairness incompatible and when are they aligned?* In this paper, we aim to answer this question by demonstrating that the relationship between fairness, privacy, and utility is *data-dependent*. We make the following contributions:

1. We define Chernoff Difference (CD), a tool for evaluating the relationship between fairness, privacy, and accuracy that relies on the information-theoretic concept of Chernoff Information.
2. We prove that for Gaussian distributions the relationship between fairness, privacy, and accuracy falls into one of three cases, depending on the data distribution. We highlight the conditions for these cases and their fairness, privacy, and accuracy implications.
3. We demonstrate that CD acts as a proxy for the steepness of the fairness-accuracy curve.

Related Works. The relationship between fairness and privacy has been an area of interest over recent years. Cummings et al. (2019), prove that it is impossible to achieve *exact* fairness and privacy with non-trivial accuracy. Sanyal et al. (2022) show that under long-tail and imbalanced

distributions and strict privacy assumptions, private and accurate algorithms incur worse fairness, while enforcing privacy and fairness leads to reduced accuracy. Mangold et al. (2023) bound the difference in fairness between private and non-private models. Bagdasaryan & Shmatikov (2019) and Chang & Shokri (2021) empirically demonstrate the fairness violations in private algorithms. These strong privacy and fairness requirements can be relaxed, however, leading to works that have demonstrated success in designing algorithms that provide privacy, fairness, and accuracy across different domains (Lyu et al., 2020; Lowy et al., 2023; Ghoukasian & Asoodeh, 2024). Many works have aimed to study fairness and privacy through the lens of information theory (Ghassami et al., 2018; Ünsal & Önen, 2023). Despite this, very few works have explored their interaction or using Chernoff Information in this domain (Nielsen, 2013; Dutta et al., 2020). To our knoweledge, our work is the first to examine this interaction through the lens of Chernoff Information.

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, our data distribution can be defined as a mixture of the following 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 $S = 1$. In this paper, we operate under the assumption that these conditional distributions are Gaussian, defined as $P_0 = \mathcal{N}(\mu_0, \sigma^2\mathbf{I}), P_1 = \mathcal{N}(\mu_1, \sigma^2\mathbf{I}), Q_0 = \mathcal{N}(\zeta_0, \tau^2\mathbf{I}), Q_1 = \mathcal{N}(\zeta_1, \tau^2\mathbf{I}),)$, and our data is balanced ($\Pr(S = 0) = \Pr(S = 1)$). Extending beyond these assumptions is a future work. Our central goal will be to examine how the definition of these distributions affects the trade-off among fairness, privacy, and accuracy for a classifier. For privacy, we are considering differential privacy (Dwork et al., 2014), a mathematical notion of privacy which, using ε (privacy budget) and δ (error probability), bounds the impact that a single point can have on a randomized algorithm. This notion of privacy is common in machine learning, with many algorithms being developed to train models with differential privacy guarantees (Abadi et al., 2016). In this paper, we leverage the input perturbation privacy method, in which we can enforce differential privacy by adding Gaussian noise ($\mathcal{N}(0, \eta^2\mathbf{I})$) to all data points before training a model. The amount of noise, calibrated by η^2, has a direct relationship to ε and δ (Theorem A.1 Kang et al. (2020)). For our notion of fairness, we consider the equal opportunity notion (Hardt et al., 2016), in which fairness is evaluated as the difference between the true positive rates (or false negative rates) for each of our groups (P and Q). Thus, to evaluate the relationship between fairness, privacy and accuracy, we examine how fairness and accuracy change as we add Gaussian noise to our data. For our model, we consider a split classifier setting. In this setting, we train a different classifier for each of our groups 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). Recent work (Dutta et al., 2020) has demonstrated that Chernoff Information can characterize the relationship between fairness and accuracy. Building on this, we will leverage Chernoff Information as a tool to quantify the relationship between fairness, privacy, and accuracy. We begin with the formal definition below.

Definition 1. *(Chernoff Information (Chernoff, 1952)) For two distributions $P_0(x)$ and $P_1(x)$ the Chernoff Information is given by:*

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

Chernoff Information can be interpreted as a divergence term and thus acts as a metric that quantifies the *separability* between $P_0(x)$ and $P_1(x)$ (Nielsen, 2011; Dutta et al., 2020). For large values of Chernoff Information, our hypotheses $P_0(x)$ and $P_1(x)$ are more separable which indicates an easier classification setting. Conversely, smaller values Chernoff Information indicates less separability between our conditional distributions, reflected by a more difficult classification setting. This is demonstrated by the role of Chernoff Information in the bound of the Bayes optimal classifier.

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

Thus, in addition to quantifying the separability between hypotheses, the Chernoff Information acts as a proxy for the performance of the Bayes optimal classifier. This allows us to leverage this value to define Chernoff Difference and quantify the relationship between fairness and accuracy.

Definition 2. *(Chernoff Difference) The Chernoff Difference between group P and group Q is the absolute difference between the Chernoff Information between group P's conditional distributions and group Q's conditional distributions:* $CD = |C(P_0, P_1) - C(Q_0, Q_1)|$.

When groups have similar Chernoff Information (the Chernoff Difference is smaller), they have similar separability and similar error bounds for their Bayes optimal classifiers. Thus, classification results will "more fair." Conversely, when Chernoff Difference is large, there is a large disparity between separability and the error bounds of the Bayes optimal classifiers, leading to fairness violations. Additionally, Chernoff Difference acts as a direct proxy of the steepness of the fairness-accuracy curves (larger CD indicates steeper curves). This trend was theorized in Dutta et al. (2020), but our paper is the first to demonstrate it (Section 4).

3 Gaussian Noise Chernoff Difference

While the Chernoff Difference provides valuable insights into the relationship between fairness and accuracy, it is typically intractable to compute. In our setting, however, we operate under the assumption that $P_0(x)$, $P_1(x)$, $Q_0(x)$, and $Q_1(x)$ are Gaussian distributions. This, allows us to extend a result from Dutta et al. (2020) to derive a closed form expression for the Chernoff Difference.

Lemma 2. *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:*

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

The derivation of this closed-form expression is provided in Appendix A. This closed-form expression allows us to directly compute the Chernoff Difference from the definition of the distributions. Next, to incorporate our notion of privacy, we define a noisy variant of Chernoff Difference by leveraging the normal-sum theorem (Lemons & Langevin, 2002).

Definition 3. *(Gaussian Noise Chernoff Difference) For some $\eta^2 \geq 0$, we define the Gaussian noise Chernoff Difference as:*

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

While Chernoff Difference provides insight into the relationship between fairness and accuracy, $\widetilde{CD}_{\eta^2}$ provides a single value that can capture the relationship between fairness, privacy, and accuracy. More specifically, it allows us to analyze how adding noise (inducing stronger privacy protection) affects both fairness and accuracy. To analyze this relationship in more depth, we examine the behavior of $\widetilde{CD}_{\eta^2}$ as we vary the noise parameter η^2. This leads to the central result of this work.

Theorem 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 Gaussian Noise Chernoff Difference ($\widetilde{CD}_{\eta^2}$) over the $\eta^2 > 0$ regime: (i) The Chernoff Difference has a maximum point. (ii) The Chernoff Difference has a maximum point and a reflection point (where $\widetilde{CD}_{\eta^2} = 0$). (iii) the Chernoff Difference is non-increasing[1]. The conditions for these three cases are given as follows:*

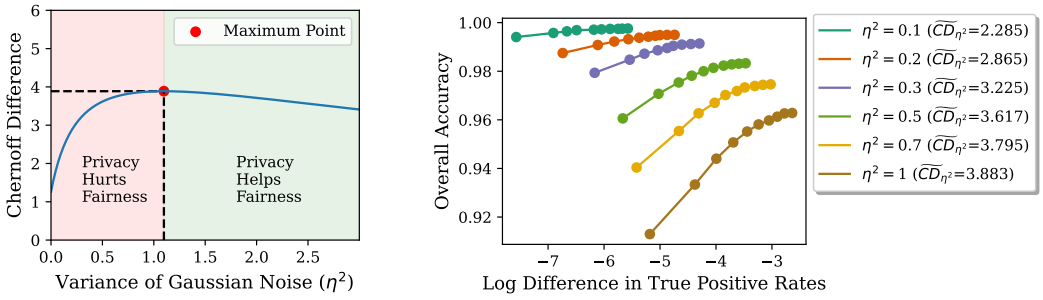
(i) $\frac{\|\zeta_0 - \zeta_1\|_2^2}{\|\mu_0 - \mu_1\|_2^2} < \frac{\tau^2}{\sigma^2} < \frac{\|\zeta_0 - \zeta_1\|_2}{\|\mu_0 - \mu_1\|_2} < 1,$ ***(Privacy Hurts Fairness)***

(ii) $\frac{\tau^2}{\sigma^2} < \frac{\|\zeta_0 - \zeta_1\|_2^2}{\|\mu_0 - \mu_1\|_2^2} < 1,$ ***(Privacy Can Give Free Fairness)***

(iii) Neither condition (i) or (ii) hold. ***(Triple Trade-off)***

We provide a proof for this theorem in Appendix A. These three cases correspond to different scenarios where we have distinct relationships between fairness, privacy, and accuracy. We discuss each of these cases and the distributions they describe in the following section.

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



(a) (b)

Figure 1: **(Case 1: Privacy Hurts Fairness)** (a) We observe that the $\widetilde{CD}_{\eta^2}$ increases until the maximum point, which represents the worst fairness-accuracy trade-off we encounter by adding noise. (b) The slopes of the fairness-accuracy plots increase as we add noise until the $\widetilde{CD}_{\eta^2}$ reaches the maximum. Parameters used here are: $\mu_0 = 0, \mu_1 = 16.5, \sigma = 2.43$ and $\zeta_0 = 0.5, \zeta_1 = 3.8, \tau = 0.55$.

4 Dynamics of Fairness, Privacy, and Accuracy

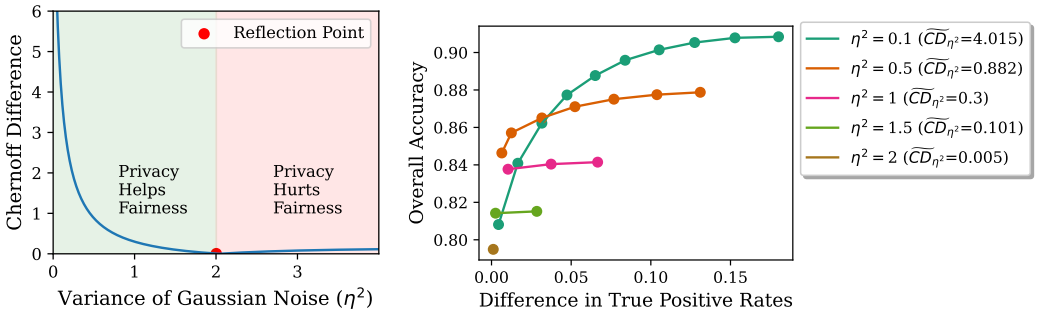
To validate the results obtained, we design a set of experiments in which $P_0(x), P_1(x), Q_0(x)$, and $Q_1(x)$ are 1-dimensional Gaussian distributions. We sample 100,000 samples from each distribution to create the balanced dataset. We then train a different Gaussian Naïve Bayes classifier for each of the groups (this provides the Bayes optimal classifier). We measure the overall accuracy of our classifier and quantify fairness by using the the true positive rate disparity between groups (equal opportunity). To clearly observe 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 are able to create fairness accuracy curves for each of our settings[2].

Case 1: Privacy Hurts Fairness. In the first case from Theorem 1, $\widetilde{CD}_{\eta^2}$ grows as we increase the variance of our Gaussian noise until it reaches a maximum. After this point, additional noise leads to a decay in $\widetilde{CD}_{\eta^2}$. This trend is reflected empirically in Figure 1(a). From Theorem 1, this scenario occurs when $\frac{\|\zeta_0-\zeta_1\|_2^2}{\|\mu_0-\mu_1\|_2^2} < \frac{\tau^2}{\sigma^2} < \frac{\|\zeta_0-\zeta_1\|_2}{\|\mu_0-\mu_1\|_2} < 1$. This condition is indicative of distributions where Q has means that are closer than that of P. Here, the variance of Q's conditional distributions is small enough (relative to the variance of group P's distributions) such that smaller amounts of noise affect group Q more than group P. However, it is not small enough (relative to the variance of group P's distributions) such that group P becomes the unprivileged group (less separable).

Next, we validate the relationship between the $\widetilde{CD}_{\eta^2}$ and the fairness-accuracy curves by examining the behavior of the curves as we add noise (increase Chernoff Difference). Since fairness-accuracy curves tend to have logarithmic shapes, we examine the slope of the logarithm fairness-accuracy curves (constructed by taking the log of the fairness values). This slope is then indicative of the steepness of our fairness-accuracy curves. We highlight the relationship between fairness-accuracy curves and log (fairness)-accuracy curves in Appendix B.1. As we observe in Figure 1(b), as we approach our maximum value of the Chernoff Difference the slope of our log (fairness)-accuracy plots increases, indicating an increase in steepness of our fairness-accuracy curves. This highlights a scenario in which privacy can hurt both accuracy and fairness.

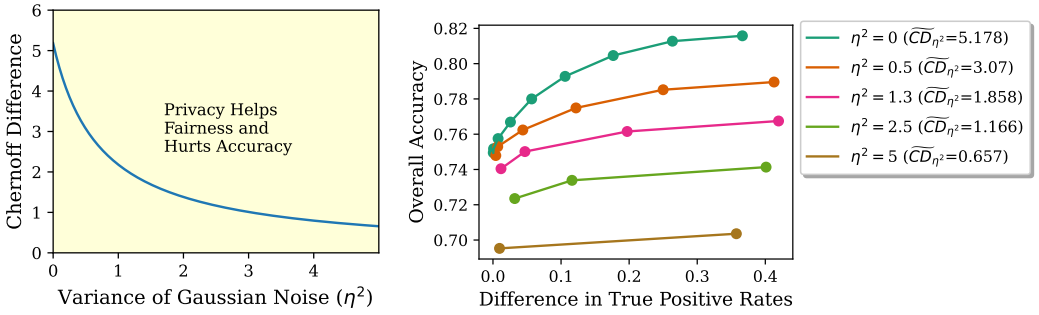
Case 2: Privacy Can Give Free Fairness. In this case, we observe that the Chernoff Difference decays to 0 before exhibiting behavior like Case 1 (Figure 2(a)). From Theorem 1, we encounter this scenario when $\frac{\tau^2}{\sigma^2} < \frac{\|\zeta_0-\zeta_1\|_2^2}{\|\mu_0-\mu_1\|_2^2} < 1$. This occurs when we have distributions in which the group with conditional distributions that have means further apart (P) also has conditional distributions with significantly larger variance. This implies that group Q has better separability (larger Chernoff Information) than group P. As we add noise, the separability of P and Q both decrease (Q much faster than P), but their values become more similar, and our classification setting becomes more fair. This is observed until the reflection point is reached ($\widetilde{CD}_{\eta^2}$ values are equal) and the setting

[2]Changing the prior probabilities is a proxy for adjusting the threshold of the Bayes optimal classifier (Dutta et al., 2020)



(a) (b)

Figure 2: **(Case 2: Privacy Can Give Free Fairness)** (a) In this case, $\widetilde{CD}_{\eta^2}$ decays until it reaches 0. It then reflects back and follows a similar pattern to Case 1. The maximum occurs at a large η^2 (6.86) value and thus is not plotted (Appendix B.2). (b) The steepness of the fairness-accuracy plots decrease as we add noise and CD decreases. The intersecting lines show that in some cases, we can achieve better fairness for the same accuracy when we add noise. Parameters used here are: $\mu_0 = -4.2, \mu_1 = 1.3, \sigma = 3$ and $\zeta_0 = 0.3, \zeta_1 = 2.7, \tau = 0.25$.



(a) (b)

Figure 3: **(Case 3: Triple Trade-off)** (a) We observe that $\widetilde{CD}_{\eta^2}$ decays steadily over the entire positive η^2 regime. (b) The fairness-accuracy curves become flatter as $\widetilde{CD}_{\eta^2}$ decreases with the increasing noise. Parameters used here are: $\mu_0 = -4.2, \mu_1 = 1.3, \sigma = 0.85$ and $\zeta_0 = 0.6, \zeta_1 = 1.6, \tau = 0.6$.

begins to become unfair. After this, adding more noise affects the separability of group Q more than that of group P until the maximum point is reached and the $\widetilde{CD}_{\eta^2}$ decays as in Case 1.

Now, to analyze the fairness, privacy, accuracy relationship, we define our Gaussian distributions so that the above condition holds. As we observe in Figure 2(b), as we increase the noise and reduce $\widetilde{CD}_{\eta^2}$, the fairness accuracy curves become less steep. Additionally, we observe that in these scenarios, the noisy curves can overlap with the clean fairness-accuracy curves, providing better accuracy for some levels of fairness. This phenomenon is indicative of scenarios in which privacy can appear to give "free fairness".

Case 3: Triple Trade-off. In this final case, $\widetilde{CD}_{\eta^2}$ naturally decays over the positive η^2 region (Figure 3(a)). This case occurs when neither condition (i) or (ii) from Theorem 1 hold. That is $\frac{\tau^2}{\sigma^2} \geq \frac{\|\zeta_0 - \zeta_1\|_2}{\|\mu_0 - \mu_1\|_2}$. In this case, group Q exhibits a larger variance in addition to their smaller mean separation (compared to group P), leading to significantly less separability. As noise is added, the separability of P is reduced at a rate faster than Q, however, the separability of Q is initially too small for P to catch up in the positive η^2 regime, leading to the non-increasing behavior.

As we observe in Figure 3(b), the fairness accuracy curve becomes less steep as the Chernoff Difference decays (reflecting a similar trend to that of which we observed in Case 2). In this scenario, however, we do not observe any overlap between the different fairness-accuracy curves as the decay of the CD occurs too slowly. This reflects the traditional "triple trade-off" that has been observed in prior works (Sanyal et al., 2022).

5 CONCLUSION

Using Chernoff Information, we reveal interesting relationships between data distributions and the fairness-privacy-accuracy trade-off. While our work is the first to expose such relationships, our analysis relies on simple Gaussian assumptions. An important next step would be extending our analysis to real-world datasets. Furthermore, designing a noise mechanism that can achieve a more favorable fairness-privacy-accuracy trade-off would be an exciting future direction.

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

Disclaimer. This paper was prepared by Hsiang Hsu prior to his employment at JPMorgan Chase & Co.. Therefore, 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

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16. ACM, October 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

Eugene Bagdasaryan and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy, 2019. URL https://arxiv.org/abs/1905.12101.

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

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 1493–1504. ACM, June 2023. doi: 10.1145/3593013.3594095. URL http://dx.doi.org/10.1145/3593013.3594095.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *CoRR*, abs/2010.04053, 2020. URL https://arxiv.org/abs/2010.04053.

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

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

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

Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct publication of the 27th conference on user modeling, adaptation and personalization*, pp. 309–315, 2019.

Sanjiv Das, Michele Donini, Jason Gelman, Kevin Haas, Mila Hardt, Jared Katzman, Krishnaram Kenthapadi, Pedro Larroy, Pinar Yilmaz, and Bilal Zafar. Fairness measures for machine learning in finance. 2021.

Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International conference on machine learning*, pp. 2803–2813. PMLR, 2020.

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

Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness, 2023. URL https://arxiv.org/abs/2302.10893.

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

AmirEmad Ghassami, Sajad Khodadadian, and Negar Kiyavash. Fairness in supervised learning: An information theoretic approach, 2018. URL https://arxiv.org/abs/1801.04378.

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

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

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Yilin Kang, Yong Liu, Ben Niu, Xinyi Tong, Likun Zhang, and Weiping Wang. Input perturbation: A new paradigm between central and local differential privacy, 2020. URL https://arxiv.org/abs/2002.08570.

Will Knight. The apple card didn't'see'gender—and that's the problem. *Wired. November*, 19, 2019.

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

Andrew Lowy, Sina Baharlouei, Rakesh Pavan, Meisam Razaviyayn, and Ahmad Beirami. A stochastic optimization framework for fair risk minimization, 2023. URL https://arxiv.org/abs/2102.12586.

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

Paul Mangold, Michaël Perrot, Aurélien Bellet, and Marc Tommasi. Differential privacy has bounded impact on fairness in classification. In *International Conference on Machine Learning*, pp. 23681–23705. PMLR, 2023.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2022. URL https://arxiv.org/abs/1908.09635.

Fatemehsadat Mireshghallah, Mohammadkazem Taram, Praneeth Vepakomma, Abhishek Singh, Ramesh Raskar, and Hadi Esmaeilzadeh. Privacy in deep learning: A survey, 2020. URL https://arxiv.org/abs/2004.12254.

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

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

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

NYDFS. Report on apple card investigation, March 2021. URL https://www.dfs.ny.gov/system/files/documents/2021/03/rpt_202103_apple_card_investigation.pdf.

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

Amartya Sanyal, Yaxi Hu, and Fanny Yang. How unfair is private learning? In *Uncertainty in Artificial Intelligence*, pp. 1738–1748. PMLR, 2022.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Ayşe Ünsal and Melek Önen. Information-theoretic approaches to differential privacy. *ACM Comput. Surv.*, 56(3), October 2023. ISSN 0360-0300. doi: 10.1145/3604904. URL https://doi.org/10.1145/3604904.

Hao Wang, Hsiang Hsu, Mario Diaz, and Flavio 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.

A APPENDIX

Theorem A.1. *(Theorem 1 from Kang et al. (2020)) In Differentially Private Empirical Risk Minimization with Input Perturbation with T iteration rounds and n samples, for $\varepsilon, \delta > 0$, if the loss function is G-Lipschitz and Δ-strongly convex over the parameters and the additive noise is parameterized by*

$$\sigma^2 = c\frac{G^2 T \log(1/\delta)}{n(n-1)\sqrt{\Delta}\varepsilon^2}, \tag{1}$$

it is (ε, δ)-differentially private for some constant c.

Lemma A.1. ***(Restatement of Lemma 2)*** *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

$$\begin{aligned}
CD &= |C(P_0, P_1) - C(Q_0, Q_1)| && (2)\\
&= |\min_{u\in(0,1)} \log \int P_0(x)^u P_1(x)^{1-u} dx - \min_{v\in(0,1)} \log \int Q_0(x)^v Q_1(x)^{1-v} dx|. && (3)
\end{aligned}$$

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

$$\begin{aligned}
&\log \int P_0(x)^u P_1(x)^{1-u} dx && (4)\\
&= \log \int e^{-\frac{u}{2\sigma^2}\left((x-\mu_1)^T(x-\mu_1) - (x-\mu_0)^T(x-\mu_0)\right)} P_0(x)\, dx && (5)\\
&= \log e^{-\frac{u}{2\sigma^2}\left(\mu_1^T\mu_1 - \mu_0^T\mu_0\right)} \int e^{-\frac{u}{2\sigma^2}\left(-2x^T(\mu_1-\mu_0)\right)} P_0(x)\, dx && (6)\\
&= \log e^{-\frac{u}{2\sigma^2}\left(\mu_1^T\mu_1 - \mu_0^T\mu_0\right)} e^{-\frac{u}{2\sigma^2}\left(-2\mu_0^T(\mu_1-\mu_0)\right)} e^{\frac{u^2}{2\sigma^2}\left(\|\mu_1-\mu_0\|_2^2\right)} && (7)\\
&= \log e^{-\frac{u}{2\sigma^2}\left(\|\mu_1-\mu_0\|_2^2\right)} e^{\frac{u^2}{2\sigma^2}\left(\|\mu_1-\mu_0\|_2^2\right)} && (8)\\
&= \frac{u(u-1)}{2\sigma^2}\|\mu_1 - \mu_0\|_2^2. && (9)
\end{aligned}$$

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

$$\frac{d}{du}\frac{u(u-1)}{2\sigma^2}\|\mu_1 - \mu_0\|_2^2 = \frac{2u-1}{2\sigma^2}\|\mu_1 - \mu_0\|_2^2. \tag{10}$$

The derivative is 0, when $u = 0.5$. This critical point is a minimum as the derivitative is negative when $u = 0$ and positive when $u = 1$. Thus,

$$C(P_0, P_1) = \frac{\|\mu_0 - \mu_1\|_2^2}{8\sigma^2}. \tag{11}$$

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 to be

$$CD = |\frac{\|\mu_0 - \mu_1\|_2^2}{8\sigma^2} - \frac{\|\zeta_0 - \zeta_1\|_2^2}{8\tau^2}|. \tag{12}$$

□

Theorem A.2. ***(Restatement of Theorem 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 Gaussian Noise Chernoff Difference ($\widetilde{CD}_{\eta^2}$) over the $\eta^2 > 0$ regime: (i) The Chernoff Difference has a maximum point. (ii) The Chernoff Difference has a maximum point and a reflection point (where $\widetilde{CD}_{\eta^2} = 0$). (iii) the Chernoff Difference is non-increasing[3]. The conditions for these three cases are given as follows:*

(i) $\frac{\|\zeta_0-\zeta_1\|_2^2}{\|\mu_0-\mu_1\|_2^2} < \frac{\tau^2}{\sigma^2} < \frac{\|\zeta_0-\zeta_1\|_2}{\|\mu_0-\mu_1\|_2} < 1,$ ***(Privacy Hurts Fairness)***

(ii) $\frac{\tau^2}{\sigma^2} < \frac{\|\zeta_0-\zeta_1\|_2^2}{\|\mu_0-\mu_1\|_2^2} < 1,$ ***(Privacy Can Give Free Fairness)***

(iii) Neither condition (i) or (ii) hold. ***(Triple Trade-off)***

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

$$s\widetilde{CD}_{\eta^2} = \frac{1}{8(\tau^2+\eta^2)}\|\zeta_0 - \zeta_1\|_2^2 - \frac{1}{8(\sigma^2+\eta^2)}\|\mu_0 - \mu_1\|_2^2. \tag{13}$$

Case (i) and (ii). 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 $s\widetilde{CD}_{\eta}^2$.

$$s\widetilde{CD}'_{\eta^2} = \frac{p^2}{8(\sigma^2+\eta^2)^2} - \frac{q^2}{8(\tau^2+\eta^2)^2} = \frac{p^2(\tau^2+\eta^2)^2 - q^2(\sigma^2+\eta^2)^2}{8(\sigma^2+\eta^2)^2(\tau^2+\eta^2)^2} \tag{14}$$

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

$$\eta_0^2 = \frac{\sigma^2 q - \tau^2 p}{p - q} \tag{15}$$

$$\eta_1^2 = \frac{-\sigma^2 q - \tau^2 p}{p + q} \tag{16}$$

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.

$$\sigma^2 q - \tau^2 p > 0 \tag{17}$$

$$\frac{\tau^2}{\sigma^2} < \frac{\|\zeta_0 - \zeta_1\|_2}{\|\mu_0 - \mu_1\|_2} < 1 \tag{18}$$

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

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

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

$$\widetilde{sCD}''_{\eta^2} = \frac{q^2}{4(\tau^2+\eta^2)^3} - \frac{p^2}{4(\sigma^2+\eta^2)^3}. \tag{19}$$

Plugging in the relevant critical point we observe that

$$\widetilde{sCD}''_{\eta_0^2} = \frac{q^2}{4(\tau^2+\eta_0^2)^3} - \frac{p^2}{4(\sigma^2+\eta_0^2)^3} = \frac{q^2}{4(\frac{q(\sigma^2-\tau^2)}{p-q})^3} - \frac{p^2}{4(\frac{p(\sigma^2-\tau^2)}{p-q})^3} \tag{20}$$

$$= \frac{q^2(p-q)^3}{4q^3(\sigma^2-\tau^2)^3} - \frac{p^2(p-q)^3}{4p^3(\sigma^2-\tau^2)^3} = \frac{(p-q)^4}{4p^3q^3(\sigma^2-\tau^2)^3}. \tag{21}$$

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

$$\widetilde{sCD}_{\eta_0^2} = \frac{q^2}{8(\tau^2+\eta_0^2)} - \frac{p^2}{8(\sigma^2+\eta_0^2)} = \frac{q^2}{8(\tau^2+\frac{\sigma^2q-\tau^2p}{p-q})} - \frac{p^2}{8(\sigma^2+\frac{\sigma^2q-\tau^2p}{p-q})} \tag{22}$$

$$= \frac{(p-q)}{8}(\frac{q}{\sigma^2-\tau^2} - \frac{p}{\sigma^2-\tau^2}) = \frac{(p-q)(q-p)}{8(\sigma^2-\tau^2)} \tag{23}$$

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.

$$\widetilde{sCD}_{\eta^2} = \frac{q^2}{8(\tau^2+\eta^2)} - \frac{p^2}{8(\sigma^2+\eta^2)} \tag{24}$$

$$8q^2(\sigma^2+\eta^2) - 8p^2(\tau^2+\eta^2) = 0 \tag{25}$$

$$\eta^2 = \frac{q^2\sigma^2 - p^2\tau^2}{p^2-q^2} = \frac{\|\zeta_0-\zeta_1\|_2^2\sigma^2 - |\mu_0-\mu_1\|_2^2\tau^2}{|\mu_0-\mu_1\|_2^2 - \|\zeta_0-\zeta_1\|_2^2}. \tag{26}$$

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

$$\|\zeta_0-\zeta_1\|_2^2\sigma^2 - |\mu_0-\mu_1\|_2^2\tau^2 > 0 \tag{27}$$

$$\frac{\tau^2}{\sigma^2} < \frac{\|\zeta_0-\zeta_1\|_2^2}{|\mu_0-\mu_1\|_2^2} < 1. \tag{28}$$

Case (iii). 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 $\widetilde{sCD}_{\eta^2}$ by writing it as

$$p^2(\tau^2+\eta^2)^2 - q^2(\sigma^2+\eta^2)^2 \tag{29}$$

$$= (p^2(\tau^2+\eta^2) - q^2(\sigma^2+\eta^2))(p^2(\tau^2+\eta^2) + q^2(\sigma^2+\eta^2)). \tag{30}$$

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

$$p^2(\tau^2+\eta^2) - q^2(\sigma^2+\eta^2) = p^2\tau^2 - q^2\sigma^2 + \eta^2(p^2-q^2). \tag{31}$$

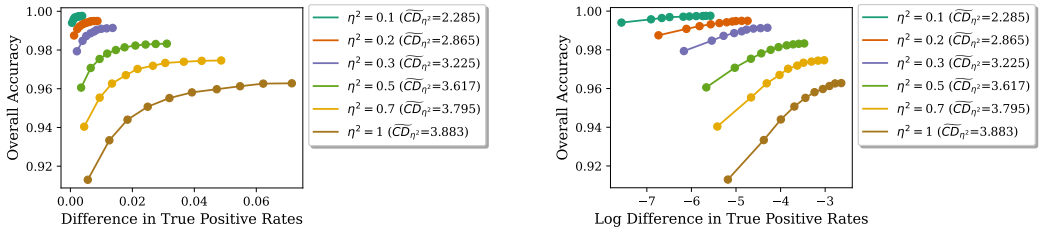
Now, from our initial assumption, $p^2 - q^2 \geq 0$. From the assumption that $\frac{\tau^2}{\sigma^2} \geq \frac{q}{p}$, $p^2\tau^2 - q^2\sigma^2 \geq 0$. Thus, the sign is always positive. Now, to show that $\widetilde{CD}_{\eta^2}$ is non-increasing, we will show $s\widetilde{CD}_{\eta^2}$ is always negative.

$$s\widetilde{CD}_{\eta^2} = \frac{q^2}{8(\tau^2+\eta^2)} - \frac{p^2}{8(\sigma^2+\eta^2)} = \frac{q^2(\sigma^2+\eta^2) - p^2(\tau^2+\eta^2)}{8((\tau^2+\eta^2))(\sigma^2+\eta^2)} \tag{32}$$

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, $s\widetilde{CD}_{\eta^2}$ is always negative and $\widetilde{CD}_{\eta^2}$ is non-increasing over the positive η^2 regime. □

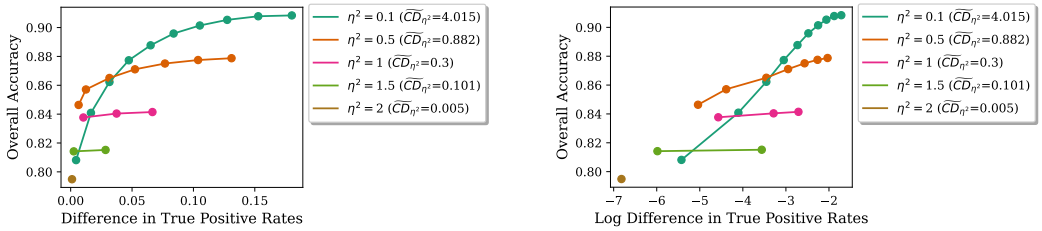
B Suplemental Figures

B.1 Fairness-Accuracy vs Log Fairness-Accuracy



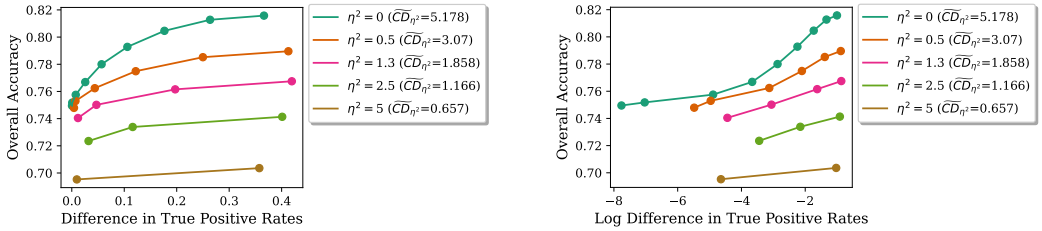
(a) (b)

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



(a) (b)

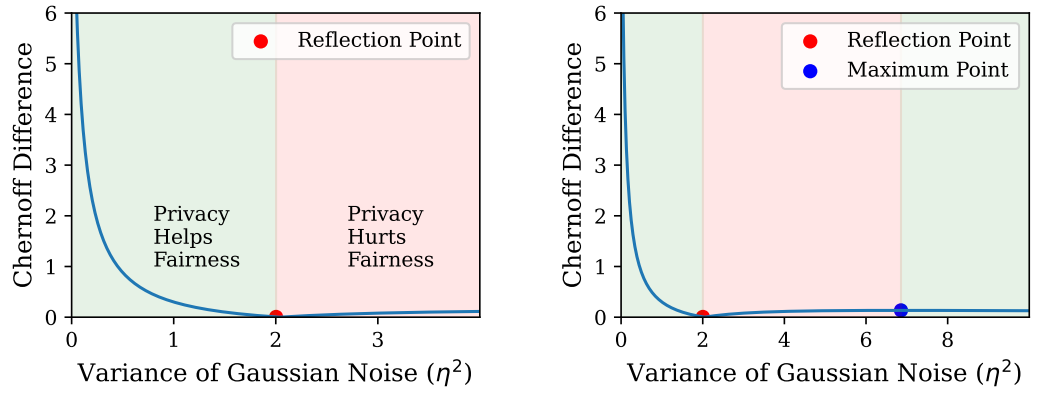
Figure B.2: **(Case 2: Privacy Can Give Free Fairness)** $\mu_0 = -4.2, \mu_1 = 1.3, \sigma = 3$ and $\zeta_0 = 0.3, \zeta_1 = 2.7, \tau = 0.25$ (a) Fairness-Accuracy Curve (b) Log Fairness-accuracy Curve.



(a) (b)

Figure B.3: **(Case 3: Triple Trade-off)** $\mu_0 = -4.2, \mu_1 = 1.3, \sigma = 0.85$ and $\zeta_0 = 0.6, \zeta_1 = 1.6, \tau = 0.6$ (a) Fairness-Accuracy Curve (b) Log Fairness-accuracy Curve.

B.2 Case 2: More Detail



(a) (b)

Figure B.4: **(Case 2: Privacy Can Give Free Fairness)** $\mu_0 = -4.2, \mu_1 = 1.3, \sigma = 3$ and $\zeta_0 = 0.3, \zeta_1 = 2.7, \tau = 0.25$ (a) Initial plot of $\widetilde{CD}_{\eta^2}$. (b) Full plot that shows presence of maximum point.