Empirically and Theoretically Understanding Machine Learning Fairness through Face Recognition

by

Harrison Rosenberg

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Electrical And Computer Engineering)

at the

UNIVERSITY OF WISCONSIN–MADISON

2024

Date of final oral examination: January 9, 2024

The dissertation is approved by the following members of the Final Oral Committee:
Kassem Fawaz, Associate Professor, Electrical and Computer Engineering
Ramya Korlakai Vinayak, Assistant Professor, Electrical and Computer Engineering
Joseph Larry Austerweil, Associate Professor, Psychology
Barry Van Veen, Lynn H. Matthias Emeritus Professor

To my brother, father, mother, aunt, and uncle for
their unconditional love and unwavering support.

Acknowledgement

This PhD was made possible by the support and guidance of my advisor Professor Kassem Fawaz. Kassem and I both arrived in Madison the same year. Since our similarly-timed arrival, Kassem facilitated my growth as a researcher and as a person. Our routine constructive academic discussions have forged a strong professional relationship. I anticipate Kassem will continue to be a valued counsel. I am, and shall remain, eternally grateful.

My PhD committee members, Professors Barry Van Veen, Ramya Korlakai Vinayak, and Joseph "Joe" Austerweil have each contributed to my success as a graduate student: Throughout my PhD career, Barry has been a valuable mentor both formally and informally; Ramya has been a consequential collaborator; and Joe has provided valuable research guidance from an external perspective.

I must also acknowledge Professor Somesh Jha as an instrumental research collaborator, informal advisor, and insightful mentor. I also thank Professors Robert Nowak, Dimitris Papailiopoulos, and Sebastien Roch for their guidance in identifying a research niche.

My choice to attend University of Wisconsin-Madison was largely inspired by four mentors in my undergraduate career: Dr. Frank Ong, and Professors Jonathan "Jon" Tamir, Michael "Miki" Lustig, and Benjamin "Ben" Recht. Through my positive experience in Miki's lab, I decided to pursue a graduate career. Jon and Frank's mentorship led me to realize I should pursue a PhD. I applied to University of Wisconsin-Madison at Frank's suggestion. When Ben shared his Madison experience, I became convinced this university should be my choice.

The success of my PhD, is due, in large part, to my labmates. Several labmates require special recognition: Brian Tang has been a friend and close collaborator. I have known Shimaa Ahmed since the start of my graduate career. She too has been a good friend and wonderful co-author. Guruprasad Viswanathan Ramesh recently joined the lab. Since then, he has become a valued colleague and co-author. I thank the rest of the lab for creating a collaborative and productive research environment: Yue Gao, Ashish Hooda, Jingjie Li, Asmit Nayak, Rishabh Khandelwal, Samarth Mathur, Yucheng Yang, Yash Wani, and Shirley Zhang.

Arindam Paul, of American Family Insurance, has provided valuable co-authorship and technical motivation, predominantly from a business perspective. I am grateful for his willingness to regularly participate in academic discussion with myself and Kassem. I am also appreciative of Robi Bhattacharjee, a graduate student at University of California, San Diego, who provided useful research input and beneficial co-authorship.

I am thankful to friends in the program, yet outside the lab: Rahul Parhi and Danica Fliss. They have been incredibly helpful, like-minded sounding boards; truly supportive friends along the PhD journey.

I am also thankful to my many friends from undergrad who also pursued PhDs: Saavan Patel, Tavor Baharav, Brijen Thananjeyan, Beliz Gunel, Vikram Sreekanti, to name a few. By sharing their experiences, I gained perspective on my own. We were all in the trenches together.

I am appreciative of Paul Arnold, Shane Marquardt, William "Bill" Chappell, and Sheila Rose for, whether they realize it or not, motivating me to mentally refresh in a manner only

possible outside of Madison. They have also imparted practical skills and general wisdom learnable neither on my own nor in a university setting. They have made my whole grad school much more enjoyable.

Mendel and Henya Matusof along with Yosef and Basya Stevenson have created a welcoming home-away-from-home. Their Friday night dinners, and the accompanying spiritual innvigoration, helped get me through the final stretch of my PhD.

The most consequential acknowledgement goes to my family. The PhD would have been impossible without the enduring support of brother, father, mother, aunt, uncle, and cousins: my entire family really. When my schedule allowed, my parents and brother enabled restful travels home and participation in important family events. If my schedule did not allow travel, they were always willing to visit. Simon and my parents were always just a phone call away.

My father and my mother, in particular, helped mitigate limitations typically associated with graduate school. My father insisted I have a car. I was not sure I needed one, but he made it happen. After my first year, when I was looking for a roommate, my father urged me to find a living situation would not be in limbo every year. He made that happen too. My mother was also tremendously supportive. She maintains a dog pack, which is not easy, but it certainly made visits home particularly enjoyable. My mother was always willing to have a long phone call, no matter the time of the day, even if the subject matter was outside her wheelhouse. She also attended my dissertation defense, managing travel through difficult winter weather.

Though not technically immediate family, Aunt Frieda and Uncle Sterling have always been like a second set of parents to me. When circumstances prevented in-person visits, I could always rely on them for moral backing and personal guidance. My cousins Lewis and Elisa each made special effort to travel to Madison and buy me a meal. Martha Hillary and Kevin Hinz opened their home my first Thanksgiving outside California. My family made successful a move more than halfway across the country.

Contents

List of Tables

List of Figures

Abstract

The strong performance of today's Artificial Intelligence systems, combined with their proliferation, has brought forth both major privacy and social concerns. We explore these concerns through the medium of automatic face recognition. Our study leverages performant text-to-image generative models and an in-depth investigation of face recognition embedding space geometries. We systematically show that privacy and social concerns can be intertwined: Performance disparities in face recognition systems, and in systems intended to preserve privacy in the presence of face recognition, occur along demographic lines. Face recognition systems are shown to be more easily fooled by manipulating images which depict faces of particular demographic groups. Such performance disparities occur even when systems do not have explicit access to demographics. Further, the concerns we identify generalize to both online APIs and offline models. Through generalization bounds, we show that the training regime, including data composition, directly influences AI system performance disparities. In particular, training set representativeness over subsets of interest is tied to overall model performance. Synthetic data data is shown to partially useful in augmenting representativeness. These generalization bounds apply to general machine learning systems.

Chapter 1

Introduction

Modern machine learning systems have exhibited strong performance in many tasks, including tasks previously thought to be only accomplishable by humans. The cost and performance efficiencies of artificial intelligence (AI) systems have led to a broad usage scope: commercial and government sectors leverage AI, with purposes ranging from from security-critical applications in military intelligence to seemingly more benign background filter functions on social media. Broad adoption of AI systems has spurred study of their vulnerabilities.

Given its extensive history, and many potential use cases, perhaps the most tantalizing application for AI researchers, commercial operators, and government actors is face recognition. Since the 1880s, face recognition has been known to have utility in criminal identification [3, 4]. With modern communication networks and the widespread nature of internet-connected digital cameras, today's face recognition technologies have a greater application domain than century-old iterations. Today's AI-backed face recognition systems have additional applications including location tracking, general-purpose identity verification, and financial transactions. Well-performing face recognition systems can be used in-place of more physically invasive biometric identification such as fingerprinting or DNA testing.

Because modern digital cameras are inexpensive, and automatic face recognition software

is easily accessible, combined facial recognition systems have seen especially rapid proliferation. For example, the Transportation Security Administration (TSA) utilizes automatic face recognition as part of identification document verification[1]. Face recognition also augments existing security cameras to enhance retail loss prevention [5]. Some financial transactions even utilize face recognition instead of traditional cash or physical credit card[2].

Given the heightened cultural interest in face recognition systems, our study of face recognition is timely. In this dissertation, we identify and explore both social and privacy vulnerabilities in face recognition systems. We find many vulnerabilities stemming from unintended awareness of demographic and semantics. Our investigation is bolstered with generative AI model auditing techniques and a supplementing theoretical narrative. Much of our theory applies to AI systems broadly.

1.1 Face Recognition System Overview

At a fundamental level, face recognition systems convert face images, which are naively neither sortable nor indexable, into a format which is. A new, or previously unseen face, can then be compared to faces already enrolled in the system. If similar to an already enrolled face, the new face can be identified. The fundamental challenge in face recognition is design and construction of the system which converts faces into a sortable, indexable format. This has been an active area of study for more than a century.

1.1.1 Face Recognition of Centuries-old

Though often thought of as a human outside-the-loop system, earliest iterations of face recognition intertwined with manual index-card lookup procedure. The Bertillon System represents a widely adopted early face recognition scheme. Dating back to the 1880s, the

[1] `https://www.tsa.gov/news/press/factsheets/facial-recognition-technology`
[2] `https://paybyface.io/`

Bertillon System relied upon a combination of photographs and manual face and body measurements. A revolution in criminal identification, governments in the United States invested in the Bertillon System: New York State, for example, created a Bertillon Bureau that contained more than 20,000 records after merely two years of operation[3]. Though largely phased out in favor of more performant criminal identification strategies, insights from the Bertillon System translate into today's biometric identification schemes.

1.1.2 Modern Automatic Face Recognition

Modern Automatic Face Recognition schemes take leverage advances in computing to augment, or even replace, time-consuming manual lookup procedures of the Bertillon System. Furthermore, manual feature collection is now repeatable, computerized process. Perhaps the earliest well-known modern face recognition procedure is built on Eigenfaces [6]. Eigenfaces leverages the concept of eigendecomposition to compare and identify faces. Since the development of Eigenfaces, non-linear techniques, often built on neural networks, provide further performance advancement in automatic face recognition. Offline models such as FaceNet [7], and online application programming interfaces (APIs) such as AWS Rekognition, represent the forefront of modern face recognition technology. We thoroughly investigate both models.

1.1.3 Deployment and Ethical Considerations

Face recognition systems have seen rapid deployment in public and private settings. Many government-deployed cameras produce images with quality sufficient for face recognition. Internet-connected security cameras in both the business and home are capable of identifying individuals [4]. The rapid deployment of internet-connected cameras presents has significant security utility, but these same cameras could be maliciously spy on citizens, depriving us of freedom and liberty. In this dissertation, we do not focus on such major political and ethical

[3] https://www.criminaljustice.ny.gov/ojis/history/bert_ny.htm

[4] https://support.google.com/googlenest/answer/9268625?hl=en&co=GENIE.Platform%3DAndroid

concerns associated with widespread face recognition, but we would be remiss not to make note.

1.2 Face Recognition Vulnerabilities

While today's face recognition systems have promising performance, they have substantial vulnerabilities. Given their role in security-critical systems and economic activity, they also represent a channel through for malicious behavior. Further, malicious behavior may not be easily detectable. There is one major cause which renders malicious behavior detection challenging. Technology underlying face recognition systems, deep networks, is not human interpretable. Hence, understanding precisely *how* a given input to a face recognition system maps to a given output is nontrivial. Further, adversarial machine learning techniques are under active development, and those techniques transfer to face recognition systems. Generally speaking, exploiting and fixing AI vulnerabilities is a high-tech cat-and-mouse game. There is no free lunch. As we'll show in chapter 2, resolving one vulnerability gives rise to others.

1.2.1 Pixel-based Attacks

Recently discovered vulnerabilities in face recognition systems include sensitivity to small, algorithmically computed, pixel-level perturbations [8, 9]. Pixel-level vulnerabilities stem from general adversarial machine learning literature [10, 11]. Szegedy et al. showed it was possible to fool a vision system by adding algorithmically generated human-imperceptible perturbations to a natural image. Goodfellow et al. then put forth the Fast Gradient Sign Method (FGSM), an iterative method for which forms the basis of a large segment of adversarial machine learning literature. Adversarial machine learning algorithms such as FGSM, when applied to faces, generate human-imperceptibly perturbed images which fool

face recognition systems [8].

1.2.2 Wearable Accessories

Wearable accessories also confound face recognition systems [12]. Whereas pixel-based attacks are often applied in an offline manner, wearable accessories are intended to defeat an always-watching face recognition system. Consumers have also shown affinity for this technology. In particular, commercially available face recognition defeating clothing has gained media attention[5]. Face masks have become a somewhat common wearable accessory. Face masks have been shown to also induce failures among face recognition systems [13].

1.2.3 Conditional Failures: Training and Validation Data

The model's training data itself can also induce model failures. Pixel-based attacks can be utilized to poison training data [9]. To the human eye, poisoned training data is nearly indistinguishable from natural data, but model failures are still induced. The failures can be made to target selected identities.

Even when model training data is entirely natural, demographic representation within the training data can cause conditional model failures. Buolamwini and Gebru highlighted conditional failures often occur on demographic lines [14], hence the term "demographic disparities". These disparities have been shown to arise from training data composition: Albiero et al. demonstrate demographically balanced training data partially mitigates performance disparities [15]. Conditional failures in deep networks are not entirely understood.

[5] `https://www.capable.design/`

1.3 Dissertation Contributions

In this dissertation, we identify and explore vulnerabilities arising from the data distribution itself. Such distributional vulnerabilities have significant social and privacy implications on face recognition systems. Note that we place tremendous focus on the technology itself. We find vulnerabilities stem from architectural considerations underlying leading face recognition models. Our investigation is bolstered with generative AI techniques and supplementing theoretical narrative. Some of our theory applies to AI systems broadly. We make several contributions to existing domain knowledge:

1. Chapter 2 explores embedding space geometries of leading face recognition systems [16]. Despite being trained without explicit access to demographic information, we find face recognition systems are demographically aware. We also examine face obfuscation, a technique built on pixel-based evasion strategies, which claims to preserve privacy in the presence of face recognition. Face obfuscation itself is shown to have vulnerabilities tied to demographics. This chapter well-encapsulates the idea that no free lunch exists in the quest for privacy.

2. Chapter 3 shifts focus away from pixel-based limitations and towards the semantic limitations of leading commercially available face recognition systems. We utilize text-to-image diffusion models to construct our evaluation data. Despite the unwieldy nature of current generative models, we demonstrate how to finely control the desired face syntheses. The evaluation explores how leading face recognition systems parse human identity. Even the most feted face recognition systems fail to correctly identify realizable, allbeit uncommon, faces. This is attributed to poor representation in training data [17]. The usage of generated evaluation data allows for targeted, user-specified evaluation on semantic limitations of interest. Because evaluation data is generated, we can "zoom in" on to help better resolve model limitations. Face recognition systems

are again shown to be demographically aware.

3. Chapter 4 explores more generally applicable machine learning theory, though the results certainly apply to face recognition [18]. We examine sample complexities for learning across multiple data subsets. Sometimes, these data subsets are referred to as population categories or demographic groups. Categorywise generalization bounds are shown to be re-parametrizations of existing empirical risk minimization (ERM) bounds. Notably these bounds apply to synthetically generated data. The bounds provide insight into training dataset construction across AI applications. In the context of face recognition, these bounds could be utilized to capture how much training data is needed to correctly identify even the least frequent facial semantics which appear in nature.

We identify AI system failure modes. Our findings indicate that training data composition itself leads to observed performance characteristics of AI systems. Data subsets, including population demographics, with poor representation in training data often translate to an AI system with poor performance on that subset. For example, we observe commercially available face recognition systems struggle to recognize a female with facial hair, and the same female without facial hair, as the same person. Similarly, for males, deviation in hairstyle may confound commercially available face recognition systems. Both failures are attributable to poor representation in training data: women with facial hair and men with significant hair length deviation are infrequent in training data scraped from the internet. Observations are in line with previous literature, which identified disparate performance of face recognition performance with respect to demographics, especially if those demographics are infrequent in training data [15].

Interestingly, performance deficiencies we identify in today's face recognition systems parallel the deficiencies of the 1880s Bertillon face recognition system. J. Edgar Hoover noted the utility of the Bertillon system, but noted the system failed to generalize to minor

children and the elderly [4]. Fosdick noted similar deficiencies, and also highlighted the Bertillon system did not successfully apply to women [19]. Measurement instruments used by Bertillon system practitioners manually collect features were also prone to miscalibration. Failings of the Bertillon system led to an eventual replacement by fingerprinting.

Unlike facial recognition of a century ago, we now possess technology to identify model failure modes prior to large-scale deployment. Mitigation techniques are also discussed within the future work portion of chapter 5. We discuss how training and measurement techniques, paired with novel generative models, may improve face recognition system performance.

Chapter 2

Fairness Properties of Face Recognition and Obfuscation Systems

Automated face recognition has proliferated in various commercial and government sectors. Face recognition systems can identify users on social media, search for missing persons, aid law enforcement and surveillance, and verify identities of individuals [20, 21]. The widespread adoption of face recognition systems has been swift with the emergence of metric embedding networks such as FaceNet [7] and ArcFace [22] as well as the abundance of labeled face data [23, 24].

Recent coverage of data breaches, privacy law violations, and the adoption of face recognition by law enforcement entities have shed light on the significant privacy issues with face recognition systems. To mitigate growing privacy concerns, face obfuscation systems have been proposed to hide user identities. Several of these systems, such as Face-Off [8], LowKey [25], and Foggysight [26], leverage properties of evasion attacks against machine learning models [11, 27, 28]. By introducing small, structured, and imperceptible perturbations to their face, a user can evade identification by a face recognition system. Such systems are attractive for an end-user: perturbations are often visually acceptable to the user; features of social media applications, such as face-augmenting filters, do not suffer; and the

obfuscation mechanism may run locally, without access to target face recognition systems.

However, face obfuscation systems suffer major shortcomings. Researchers have identified the ability of face recognition systems to adaptively learn from perturbed faces [29] and re-identify them in the future. In this work, we uncover another shortcoming of such systems: *the presence of performance disparities with respect to demographics.* This disparity leads to the following research questions:

- Are the metric embedding networks underlying face obfuscation systems aware of demographic attributes in faces?
- How does the behavior of face obfuscation systems depend on the demographic attributes of faces?
- Do bias mitigation strategies for face recognition systems also mitigate bias in face obfuscation systems?

This chapter characterizes demographic disparities of face obfuscation systems and their underlying metric embedding networks[1]. Previous research has studied the fairness and robustness properties of face recognition [30, 31][2] in the classification setting. However, we study the fairness properties of face recognition and obfuscation in the context of the metric embedding networks — the real-world setting for such systems. Our work yields the following insights into fairness implications of face recognition and obfuscation.

Are the metric embedding networks underlying face obfuscation systems aware of the demographic attributes of faces? We observe face recognition systems are better at discerning individuals in different demographic groups than discerning individuals within the same demographic. Without explicit access to demographic information, face recognition systems still learn to differentiate demographic groups.

How does the behavior of face obfuscation systems depend on the demographic attributes of faces? We analyze two recent face obfuscation systems. Face-Off [8] and

[1]Code repo: github.com/wi-pi/fairness_face_obfuscation

[2]See section 2.5 for further discussion.

LowKey [25] serve as proxies for targeted face obfuscation and untargeted face obfuscation, respectively. Minority groups require stronger perturbations to successfully obfuscate a face. This is especially true in a black-box setting, such as the Face++, Azure, and AWS Rekognition face recognition APIs. We also show that faces perturbed by untargeted attacks often retain their original demographic attributes. We conclude that larger, more visible perturbations are necessary to successfully target identities in demographic groups different from the original image.

How does the training regime affect the behavior of face obfuscation? We compare the effects of a standard face recognition network training method with two alternative training regimes designed to mitigate bias. The first regime is the training procedure defined by Xu et al [32]. The second regime is the training of models on demographically balanced datasets. We show these techniques do not entirely eliminate demographic-wise performance disparities.

To aid our response to these three questions, we devise an analytical model based on a mixture of Gaussian distributions and Principal Component Analysis (PCA). This analytical model allows us to formalize the apparent behavior of face recognition and obfuscation when conditioning on the demographic group. Our model reveals that obfuscated faces are more likely to belong to their original demographic group.

2.1 Background

The terminology with respect to population demographics used in this chapter follows that of Buolamwini and Gebru [30] and Nanda et al. [31], two leading works on face recognition fairness. In the dataset annotations, there are only two sexes, hence we use "male" and "female." As for ethnicity, previous literature utilizes terms such as "White" "Black," "Asian," and "Indian" within their attribute annotations [33]. We find it more accurate to refer to these demographic labels as "race." For consistency, we use the same demographic

attribute labels in the VGGFace2 dataset in our face recognition and face obfuscation performance evaluations.

2.1.1 Notation

We consider the setting in which there exists an input space $\mathcal{X} \subseteq \mathbb{R}^d$ and a discrete label set $\mathcal{Y}$. A subset of examples, also referred to as a dataset, is denoted as $S \subseteq \mathcal{X} \times \mathcal{Y}$. Sometimes we abuse notation and let S contain only unlabeled examples $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots\}$. A sample $\boldsymbol{x}$ is a d-dimensional real-valued vector. Often, $\boldsymbol{x}$ refers to a cropped face, d is the number of pixels in the cropped face multiplied by three (the RGB channels), and $\mathcal{Y}$ to the set of identities. Given a vector $\boldsymbol{z}$, z_j denotes its j^{th} entry. Calligraphic capital letters denote probability distributions. $\mathcal{D}$ represents the distribution from which training data S are drawn. $\mathbb{1}$ denotes the indicator function. Let z be a Boolean expression. $\mathbb{1}[z]$ evaluates to 1 if z is true, otherwise $\mathbb{1}[z]$ evaluates to 0. A metric is denoted by $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$.

2.1.2 Face Recognition System

The core component of a face recognition system is a *metric embedding network*. The metric embedding network, denoted by $\mathrm{f}_k : \mathbb{R}^d \to \mathbb{R}^k$, is a neural network which takes an RGB face image $\boldsymbol{x}$ as input, and returns a k dimensional embedding. Thus, the term *embedding* refers to the k-dimensional face representation output by the metric embedding network. We sometimes omit the subscript k when referring to a generic embedding function. The goal of a metric embedding network is to map high dimensional images into an embedding space such that any two images belonging to the same identity have lower pairwise distance than any two images with different identities. Metric embedding networks are typically trained with one of two classes of loss functions: contrastive loss [34] and triplet loss [7]. The functionality of a face recognition system is depicted in fig. 2.1.

Upon obtaining an embedding, tasks such as clustering, matching, or classification may

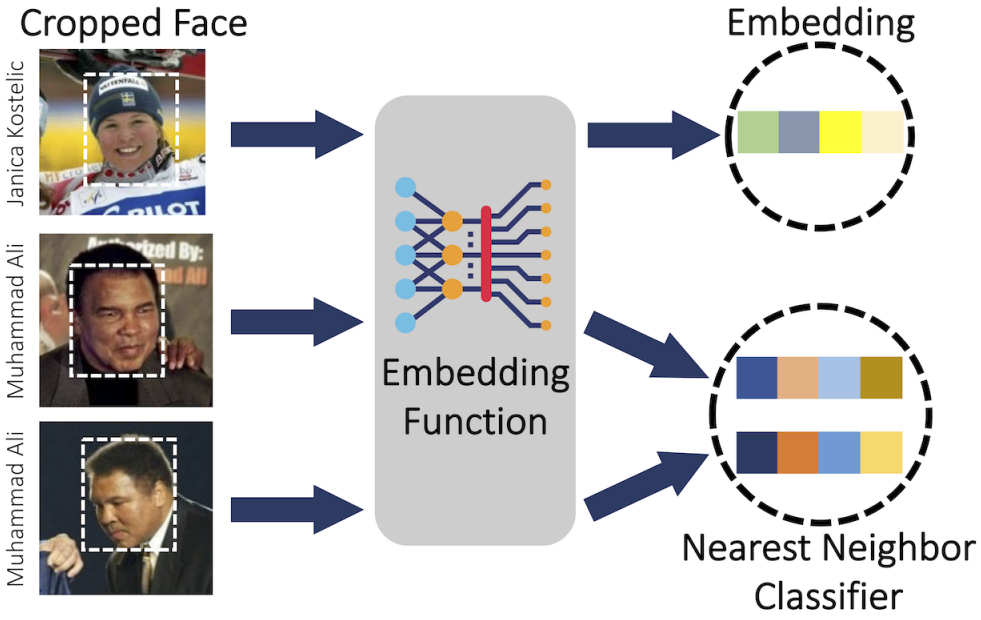


Figure 2.1: For well-trained embedding functions, embeddings of images belonging to the same identity will have smaller pairwise distances than the pairwise distances between embeddings of different identities.

be performed. A more performant metric embedding network is one which only matches embeddings of faces belonging to the same identity. A match occurs when two embeddings are sufficiently close: Given a non-negative, real-valued threshold τ and two examples $\boldsymbol{x}$, $\boldsymbol{x}'$, a match occurs when $\rho(\boldsymbol{x}, \boldsymbol{x}') \leq \tau$. Often, ρ is the ℓ_2 norm.

Matching performance is measured by TPR_z which is a parametrized notion of true positive rate. For choice of metric ρ, the match threshold threshold τ is chosen such that it satisfies a false acceptance rate upper-bounded by z.

$$\text{TPR}_z(S, S') \triangleq \sum_{(\boldsymbol{x}_i, y_i, \boldsymbol{x}'_i, y'_i) \in \left\{S, S' | y_i = y'_i\right\}}^{m} \frac{\mathbb{1}\left[\rho\left(\text{f}(\boldsymbol{x}), \text{f}(\boldsymbol{x}')\right) < \tau\right]}{|\{S, S' \mid y_i = y'_i\}|} \tag{2.1}$$

where threshold τ is defined as

$$\max \tau \quad \text{s.t.} \quad z > \sum_{(\boldsymbol{x}, y, \boldsymbol{x}', y') \in \left\{S, S' | y_i \neq y'_i\right\}} \frac{\mathbb{1}\left[\rho\left(\text{f}(\boldsymbol{x}), \text{f}(\boldsymbol{x}')\right) < \tau\right]}{|\{S, S' \mid y_i \neq y'_i\}|} \tag{2.2}$$

and the right side of inequality (2.2) is false acceptance rate.

2.1.3 Face Obfuscation Systems

Face recognition systems are vulnerable to adversarial examples. Such vulnerabilities were famously discovered by Szegedy et al [35]. Small, structured perturbations, imperceptible to humans, may cause networks to misclassify given inputs. This branch of machine learning research has led to the design of many attack algorithms, of which the most prominent are so-called evasion attacks. Examples of evasion attacks include the Projected Gradient Descent (PGD) [27] and Carlini-Wagner (CW) [28] attacks. The bulk of study focuses on algorithmic generation of ℓ_p norm-bounded perturbations via noisy gradient-based optimization.

Instead of seeing adversarial examples as a challenge, privacy researchers have demonstrated that systems which leverage such structured perturbations can provide users with privacy utility against face recognition. These privacy benefits are encapsulated in systems known as face obfuscation systems. In this section, we present notation and discuss work relevant to such face obfuscation systems.

Threat Model

In the face obfuscation setting, an end user considers the machine learning provider to be the main threat. Such threats include data breaches [36], cyber-stalkers [37], web scrapers, big government entities [38], and more. Face obfuscation systems counter such threats: Prior to upload, the user applies (imperceptible) perturbations to their face images. Perturbations are designed so the face recognition system is aware of the presence of the face, but the predicted identity of the face is incorrect. In other words, perturbations are constructed so that, from the perspective of the face recognition system, the user impersonates another identity. This is the *white-box* setting: examples are generated on the target model directly.

We also consider the black-box setting. Perturbations applied in the black-box setting leverage an important property of adversarial examples: their ability to transfer across mod-

els [39, 40]. Transferability allows users to impersonate identities without accessing a target model. In the *black-box* setting, face obfuscation systems leverage transferability by querying a surrogate model to generate perturbed faces.

Overview of Face Obfuscation Systems

The earliest works in face obfuscation explore physically perturbed examples. For example, Sharif et al. [12] present physically realizable glasses that allow a human user to impersonate a target individual. With the recent societal focus of online privacy, researchers have shifted their focus to digital face obfuscation systems. Digital face obfuscation systems are better suited for social media and internet applications. Examples of such systems include FAWKES [9], Face-Off [8], Low-Key [25], and FoggySight [26]. With the exception of FAWKES, which leverages data poisoning attacks, these face obfuscation systems utilize evasion attacks in an attempt to hide the user's identity.

We now describe face obfuscation systems more formally. Let Δ denote an evasion attack function (e.g. PGD, CW), and let $\boldsymbol{\delta}$ denote a generic perturbation output by the evasion attack function, i.e. $\Delta(\boldsymbol{x})$. A face obfuscation system feeds $\boldsymbol{x} + \boldsymbol{\delta}$ into the face recognition system. Note that Δ may include dependence on the metric embedding network, the underlying dataset, or some other surrogate model.

Evasion Attacks on Metric Embedding Networks

There are two types of evasion attacks: targeted and untargeted. We first describe the embedding centroid before defining the attacks: Given dataset S, denote by $\boldsymbol{c}_{\mathrm{f},y}$ the embedding centroid of identity y as computed on embedding function f:

$$\boldsymbol{c}_{\mathrm{f},y} \triangleq \frac{1}{|\{(\boldsymbol{x}', y') \in S \mid y' = y\}|} \sum_{(\boldsymbol{x}', y') \in S} \mathrm{f}(\boldsymbol{x}') \cdot \mathbb{1}\left[y' = y\right] \tag{2.3}$$

Untargeted Attacks: Given a labeled example $(\boldsymbol{x}, y)$, untargeted attacks find a perturba-

tion $\boldsymbol{\delta}$ for which obfuscated face $\boldsymbol{x}+\boldsymbol{\delta}$ impersonates some identity y_0, where $y_0 \neq y$. Given a metric embedding network $\mathrm{f} : \mathcal{X} \to \mathcal{Y}$, and metrics ρ_1, ρ_2 an *untargeted attack* is formulated as:

$$\min_{\boldsymbol{\delta} | \{\boldsymbol{x}+\boldsymbol{\delta} \in \mathcal{X}\}} \rho_1(\boldsymbol{\delta}, \mathbf{0})$$

$$\text{s.t.} \quad \arg\min_{y' \in \mathcal{Y}} \rho_2(\boldsymbol{c}_{\mathrm{f},y'}, \mathrm{f}(\boldsymbol{x})) \neq \arg\min_{y' \in \mathcal{Y}} \rho_2(\boldsymbol{c}_{\mathrm{f},y'}, \mathrm{f}(\boldsymbol{x}+\boldsymbol{\delta}))$$

The problem with the formulation above is its intractable constraint. To make the attack implementable in practice, the constraints must be relaxed. For a labeled example $(\boldsymbol{x}, y)$, the optimization objective which yields an untargeted perturbation can be written as:

$$\arg\max_{\boldsymbol{\delta} | \{\boldsymbol{x}'+\boldsymbol{\delta} \in \mathcal{X}\}} \rho_2(\mathrm{f}(\boldsymbol{x}'+\boldsymbol{\delta}), \boldsymbol{c}_{\mathrm{f},y}) \quad \text{s.t.} \quad \rho_1(\boldsymbol{\delta}, \mathbf{0}) \leq \epsilon \tag{2.4}$$

In this chapter, we use the LowKey attack [25] to instantiate the attack described in eq. (2.4). Within LowKey, PGD is used to solve eq. (2.4), ρ_1 is the Learned Perceptual Image Patch Similarity (LPIPS) metric [41], and ρ_2 is the distance between the original face and perturbed face in the embedding space. Note that the latter distance is averaged through an ensemble of models and after applying Gaussian smoothing.

Targeted Attacks: Given a labeled example $(\boldsymbol{x}, y)$, and target identity y_0, targeted attacks find perturbation $\boldsymbol{\delta}$ so that obfuscated face $\boldsymbol{x}+\boldsymbol{\delta}$ impersonates y_0. Given a metric embedding network $\mathrm{f} : \mathcal{X} \to \mathbb{R}^k$, and metrics ρ_1, ρ_2 a *targeted attack* may be formulated as:

$$\min_{\boldsymbol{\delta} | \{\boldsymbol{x}+\boldsymbol{\delta} \in \mathcal{X}\}} \rho_1(\boldsymbol{\delta}, \mathbf{0}) \quad \text{s.t.} \quad y_0 = \arg\min_{y \in \mathcal{Y}} \rho_2(\boldsymbol{c}_{\mathrm{f},y}, \mathrm{f}(\boldsymbol{x}+\boldsymbol{\delta})) \tag{2.5}$$

Similar to the untargeted case, targeted attacks on face recognition systems relax the above constraints to arrive at a more tractable optimization formulation. Multi-class hinge loss, as used by FaceOff [8], provides the desired relaxation of the constraints in the targeted attack eq. (2.5). Given a perturbed example $\boldsymbol{x}+\boldsymbol{\delta}$, target label y_0, and positive real number κ, the

multi-class hinge loss is denoted by $\mathrm{G}_\kappa(\boldsymbol{x}+\boldsymbol{\delta}, y_0)$ where:

$$\mathrm{G}_\kappa(\boldsymbol{x}+\boldsymbol{\delta}, y_0) \triangleq \max\Big\{0, \kappa + \rho_2(\boldsymbol{x}+\boldsymbol{\delta}, \boldsymbol{c}_{\mathrm{f},y_0}) - \max_{y' \neq y_0} \rho_2(\boldsymbol{x}+\boldsymbol{\delta}, \boldsymbol{c}_{\mathrm{f},y'})\Big\}$$

For a labeled example $(\boldsymbol{x}, y)$, and target label $y_0 \neq y$, the CW attack can minimize the following optimization objective to yield a targeted perturbation:

$$\underset{\boldsymbol{\delta}|\{\boldsymbol{x}+\boldsymbol{\delta}\in\mathcal{X}\}}{\arg\min}\ \rho_1(\boldsymbol{\delta}, \mathbf{0}) \quad \text{s.t.} \quad \rho_1(\boldsymbol{\delta}, \mathbf{0}) \leq \epsilon \text{ and } \mathrm{G}_\kappa(\boldsymbol{x}+\boldsymbol{\delta}, y_0) \leq 0 \tag{2.6}$$

2.1.4 Analytical Techniques

We utilize a combination of established data analysis techniques to assess the fairness of both face recognition and face obfuscation systems.

PCA: Neural networks with nonlinearities are notoriously difficult to analyze. To gain intuition for face obfuscation, we analyze PCA. PCA is a theoretically tractable embedding function which can be represented as a neural network. Hence, we use PCA as a proxy for non-linear embedding functions.

t-SNE: The t-Distributed Stochastic Neighbor Embedding (t-SNE) [42] is a dimensionality reduction technique useful for visualizing high dimensional embedding spaces and image datasets. Visualizations rendered by t-SNE are two or three dimensional. t-SNE is a variant of Stochastic Neighbor Embeddings [43] which avoids crowding data points and can capture the implicit structure of data. t-SNE plots aid our discussion of the embedding space geometry.

TCAV: Testing with Concept Activation Vectors (TCAV) [44] is a tool used to interpret deep neural networks. Given user-specified high-level concepts, such as patterns or colors, linear classifiers are trained on the neural network's activations for those concepts. Concept Activation Vectors (CAVs) are extracted from the vector orthogonal to the linear classifier's decision boundary, and a statistical significance test is performed on the TCAV score

generated from the dot product of each CAV and the model's gradients.

Fairness Definitions: There is no single definition of fairness, mathematical or otherwise [45]. Hence, we study several quantities which have an intuitive connection to both face recognition and face obfuscation systems. These quantities include a comparison, by demographic, of face obfuscation success. When success rates are equal across demographics, the fairness constraint known as statistical parity [46] is satisfied. We also compare, by demographic, the strength of perturbations necessary to yield a successful face obfuscation. For targeted obfuscation, we study the strength of such perturbations necessary to impersonate identities within both the inter-demographic group and intra-demographic group settings. We also study how likely untargeted perturbations are to change the perceived demographic of an image.

True positive rate balancing is another notion of fairness which appears in machine learning literature. In section 2.3.3, we see that a specialized training procedure by Xu et al. [32] yields a metric embedding network for which demographic groups are approximately equal in their matching performance TPR_z. An optimization constraint explicitly enforcing equal true positive rates between groups is the fairness metric known as Equalized Odds [47].

Statistical Significance: We use statistical significance testing to draw conclusions from experiments. The most general two-sample t-test, used to determine if two distributions have unequal means, is Welch's t-test [48]. Welch's t-test is applicable when population variances are nonequal and/or the number of elements in each sample differs. For statistical tests on more than two samples, we apply the Alexander-Govern test [49], a multi-sample generalization of Welch's t-Test. Each statistical test is designed to determine if any two or more samples are drawn from the same distribution. In all our statistical tests, we consider a p-value less than 0.05 to be significant. For the remainder of the chapter, we will omit the specific t-test used to obtain p-values. p-values are implicitly assumed to have been obtained by either Welch's t-test or the Alexander-Govern test. Because we have only applied one null

hypothesis per statistical test, our statistical tests do not suffer from the multiple testing problem.

2.2 Experiment Motivation and Overview

From our experiments and analysis, we wish to understand how biases inherent in both face recognition datasets and metric embedding networks impact the performance of face obfuscation systems. We answer three questions:

1. ***Bias in Face Recognition*: What is the baseline bias present in face recognition systems?** Our experiments concur with existing literature: when conditioning by demographic, there is a disparity in face recognition system performance. Further, we identify that networks learn to identify skin-tone in early layers of the network.
2. ***Effectiveness of Obfuscation*: How does the strength of perturbation necessary to obfuscate a face depend on a face's demographic?** Our findings indicate face recognition systems are less robust to perturbations applied to faces from minority demographic groups. For minority demographic groups, the perturbation strength necessary to impersonate an identity is smaller compared to majority demographics. Consequently, obfuscated faces tend to remain classified as a member of the same demographic group. Performance disparities are evident between demographics in both targeted and untargeted obfuscation.
3. ***Bias Mitigation*: How do bias mitigation strategies applied during training affect the utility of face obfuscation?** We apply two bias mitigation strategies to the training of metric embedding networks. The first strategy is a training procedure defined by Xu et al. [32]. The second strategy is training on demographically balanced datasets. With such training, we see the resulting embeddings are less clustered. Performance disparities of the face obfuscation system with respect to demographics are also attenuated.

The benefits of bias mitigation are not free: overall model accuracy is reduced.

2.2.1 Experimental Setup

Datasets and Models: We utilize the LFW and VGGFace2 datasets in our evaluation. In the white-box setting, we generate perturbed faces on the FaceNet model. In the black-box setting, we test perturbed faces for transferability by evaluating them on seven different models. Three of the seven models are commercial face recognition APIs: Face++ [50], Azure Face [51], and AWS Rekognition [52]. The remaining four are pre-trained open source models: OpenFace [53], DeepFace [54], ArcFace [22], and VGGFace [24]. Each model uses convolutional neural network architecture similar to FaceNet.

Fundamental Comparison: To determine the relationship between demographics and face obfuscation performance, we consider the performance of obfuscation targeting identities in the same demographic group as the source identity, and obfuscation targeting identities in demographic groups different from that of the source identity. In particular, we consider six demographic attributes, four for race (Black, White, Indian, Asian) and two for sex (female, male). 50 identities per attribute from LFW are sampled[3]. These images are referred to as source images. Source images are inputs to the untargeted attack. Two targeted attacks scenarios are considered:

- **Same demographic:** We choose 49 pairwise combinations of target identities from the same race/sex for each source identity. This sampling leads to 2450 source-target pairs of the same demographic.
- **Different demographic:** We subsample, uniformly at random, 15 target identities from each race group of which the source identity is not a member for a total of 45 target identities. For the sex demographic, we assemble 50 target identities from the opposite sex. This sampling leads to 2250 source-target pairs of the different races and 2500 source-

[3]The LFW demographic attributes were annotated using attribute classifiers described by Kumar et al. [33].

target pairs of the different sex.

We generate untargeted adversarial examples for each of the 5,749 identities in the LFW dataset and their associated images. We generate targeted adversarial examples for earlier scenario's 300 identities and 80,000+ targeted examples corresponding to the 28,700 pairs of identities. If unstated within a caption, the chosen dataset is LFW and the chosen metric embedding network is FaceNet.

Obfuscation Techniques: We perform our evaluation using untargeted and targeted variants of face obfsucation systems, as described in section 2.1.3. Utilizing the Face-Off face obfuscation system [8] and the FaceNet model [7], we generate adversarial examples on a subset of LFW [23]. For sections 2.3.1 and 2.3.2 we use 0, 5, and 10 as our margin values (κ in eq. (2.6)). We also generate untargeted perturbations with the LowKey [25] procedure described in section 2.1.3. The ensemble used by LowKey includes two pre-trained ArcFace models [22] and two pre-trained Cosface models [55]. We report the obfuscation success rate as an indicator of perturbation effectiveness. In the targeted case, obfuscation success rate measures the proportion of perturbed faces which match their intended targets, so we expect targeted obfuscation success rate to decrease as a function of the threshold τ. Furthermore, the distance between an embedding and its target directly controls obfuscation success. In the untargeted case, obfuscation success rate measures the proportion of perturbed faces that evade their source identity. We expect the untargeted obfuscation rate to increase as a function of the threshold τ.

2.2.2 An analytical model for face obfuscation

Consistent with usage of toy models in previous machine learning literature [56–58], we devise a hierarchical Gaussian distribution with which we explore fair face obfuscation. While our model does entirely capture neural network behavior, the model conveys intuition on how how discrepancies in demographic sampling lead to disparities we observe in face obfuscation

utility. To this end, the analytical model consists of a k-component PCA and hierarchical Gaussian distribution. The k-component PCA, which utilizes projections onto the leading k principal components to perform its dimensionality reduction, is the theoretically tractable proxy for a nonlinear metric embedding network. Use of the hierarchical Gaussian is inspired by both the hierarchical nature of popular face recognition datasets and the embedding space geometry. Though all samples drawn from our simplified model are vectors, we use terms *identity* and *image* to draw parallels between the hierarchical nature of our probabalistic model and the hierarchical structure of existing face recognition datasets.

Within the hierarchical Gaussian are two mutually exclusive groups: group a and group b. Sometimes a placeholder g is used to represent a group $\mathsf{g} \in \{\mathsf{a}, \mathsf{b}\}$. $\boldsymbol{\mu}_{\mathsf{a}} \in \mathbb{R}^d$ and $\boldsymbol{\mu}_{\mathsf{b}} \in \mathbb{R}^d$ denote the mean vectors for population groups a and b, respectively. Moreover, $\boldsymbol{\mu}_{\mathsf{a}} = -\boldsymbol{\mu}_{\mathsf{b}}$, $\|\boldsymbol{\mu}_{\mathsf{a}}\|_2 = 1$, and $\|\boldsymbol{\mu}_{\mathsf{b}}\|_2 = 1$. Figure 2.2 is a visual depiction of our analytical model.

The i^{th} identity in group g is denoted by $\boldsymbol{\nu}_{\mathsf{g},i} \in \mathbb{R}^d$. The j^{th} image representing identity $\boldsymbol{\nu}_{\mathsf{g},i}$ is denoted by $\boldsymbol{x}_{\mathsf{g},i,j} \in \mathbb{R}^d$. $\boldsymbol{\Sigma}_{\mathsf{a}} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{\Sigma}_{\mathsf{b}} \in \mathbb{R}^{d\times d}$ are diagonal covariance matrices. Furthermore, $\boldsymbol{\Sigma}_{\mathsf{a}} = \gamma\boldsymbol{\Sigma}_{\mathsf{b}}$ where $\gamma \in \mathbb{R}^+$.

By construction, the γ parameter captures demographic imbalance in sampling. Let us assume, that if both groups a and b are sampled equally in a manner matching the natural distribution, then $\boldsymbol{\Sigma}_{\mathsf{a}} = \boldsymbol{\Sigma}_{\mathsf{b}}$. Without loss of generality, let us assume $\gamma \in (0, 1]$. As γ decreases, the sampling from the natural distribution deviates further from the underlying distribution. Disparities in embedding density become apparent: When group a has few samples, embeddings of identities within group a are close together, similar to how minority groups are depicted in fig. 2.3.

An identity $\boldsymbol{\nu}_{\mathsf{g},i}$ is drawn from the identity distribution $\mathcal{N}(\boldsymbol{\mu}_{\mathsf{g}}, \boldsymbol{\Sigma}_{\mathsf{g}})$. The identity distribution may be thought of as a hyperprior on images. An image $\boldsymbol{x}_{\mathsf{g},i,j}$ is drawn from $\mathcal{N}(\boldsymbol{\nu}_{\mathsf{g},i}, \beta\boldsymbol{I})$ where β is a positive, real-valued number. For each identity $\boldsymbol{\nu}_{\mathsf{g},i}$, exactly m images are drawn from $\mathcal{N}(\boldsymbol{\nu}_{\mathsf{g},i}, \beta\boldsymbol{I})$. Lastly, we denote by $\mathcal{D}_{\mathsf{g}}$ the distribution of images in group g. Thus, given

$\alpha \in (0,1)$, we can represent, the hierarchical Gaussian distribution as $\mathcal{D} = \alpha\mathcal{D}_{\mathsf{a}} + (1-\alpha)\mathcal{D}_{\mathsf{b}}$.

In the context of fair face obfuscation, we concern ourselves with how well the embedding represents a particular group from the lens of that group. This is captured by relative projection distance. The relative projection distance measures how well the k-component PCA embedding function represents a particular group, from the lens of that group:

Definition 2.1. *Let S be a sample of images drawn from the overall synthetic distribution $\mathcal{D}$. The relative projection distance of a point $\boldsymbol{x}$, a member of group g, with respect to the leading k principal components of a dataset S is denoted by $\rho_{\mathrm{rp},S,\mathsf{g},k} : \mathcal{X} \to \mathbb{R}^+$. More precisely:*

$$\rho_{\mathrm{rp},S,\mathsf{g},k}(\boldsymbol{x}) \triangleq \frac{\left\| \left(\boldsymbol{x} - \sum_{i=1}^{k} \left\{ \frac{\boldsymbol{q}_i^{\mathsf{T}} \boldsymbol{x}}{\|\boldsymbol{q}_i\|_2 \|\boldsymbol{x}\|_2} \boldsymbol{q}_i \right\} \right) \right\|_2}{\sum_{j=1}^{d} \left(\boldsymbol{\Sigma}_{\mathsf{g}}\right)_{jj}}, \tag{2.7}$$

where the eigen-decomposition of the covariance of overall synthetic data distribution $\mathcal{D}$ is $\boldsymbol{\Sigma} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathsf{T}}$. Furthermore, $\boldsymbol{Q}$ may be decomposed as $\boldsymbol{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_d]^{\mathsf{T}}$.

The numerator of relative projection distance is the norm of the portion of sample $\boldsymbol{x}$ which is not captured by f_k. The denominator represents a group specific normalization factor representing the overall variance within all identity vectors $\boldsymbol{\nu}_{\mathsf{g},i}$. It is this normalization factor which allows us to show how error can be measured in the context of group g.

2.3 Fairness on Face Obfuscation Systems

Utilizing the experimental procedure from section 2.2.1, we answer the three questions appearing at the beginning of section 2.2. Our findings highlight demographic disparities in face obfuscation systems. In particular, we show that it is easier to impersonate identities in the same demographic than it is to impersonate identities in a different demographic.

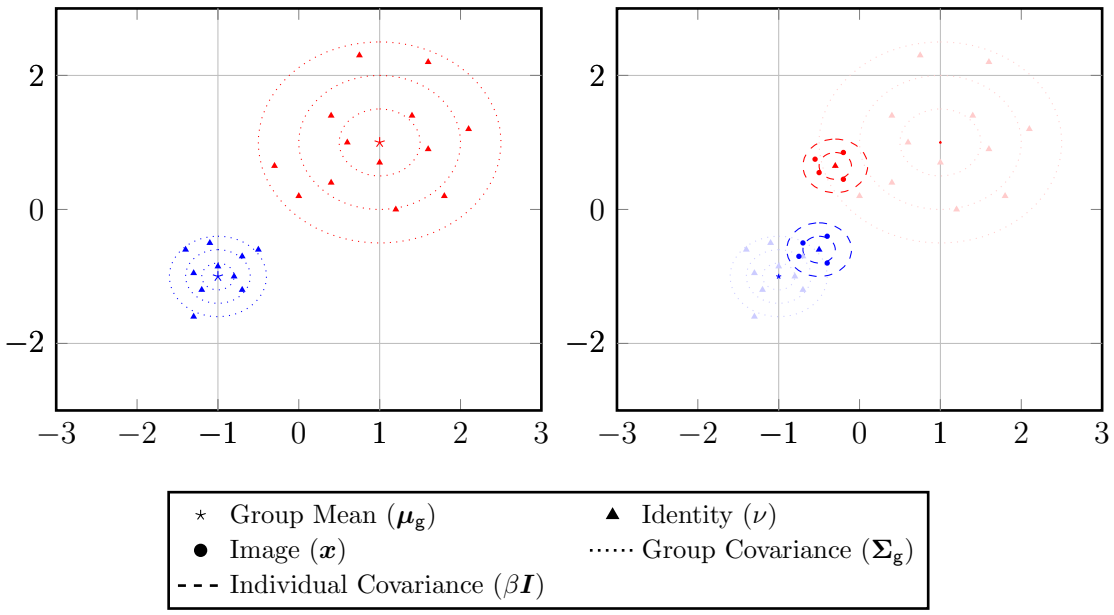


Figure 2.2: The hierarchical Gaussian distribution models patterns we observe with respect to faces in the embedding space. In the left plot, identities are drawn. Identities, $\boldsymbol{\mu}_{\mathsf{a}}$ and $\boldsymbol{\mu}_{\mathsf{b}}$ are depicted with a red star and a blue star, respectively. Identities $\boldsymbol{\nu}_{\mathsf{g},i}$ in group g are sampled from $\mathcal{N}(\boldsymbol{\mu}_{\mathsf{g}}, \boldsymbol{\sigma}_{\mathsf{g}})$. Identities are depicted as triangles. In the right plot, images are drawn. These images $\boldsymbol{x}_{\mathsf{g},i,j}$, depicted as circles, are drawn from $\mathcal{N}(\boldsymbol{\nu}_{\mathsf{g},i}, \beta \boldsymbol{I})$.

2.3.1 Error Disparity Among Groups

Face recognition systems are known to be biased with respect to population demographics. We show such demographic disparities exist in the embedding space, and can be traced through the learning process of the embedding network.

Face Recognition Empirical Performance: Table 2.1 demonstrates the demographic disparities in face recognition. When evaluating matching performance $\text{TPR}_{0.001}$ on the pretrained FaceNet model, pairs of identities selected within the same demographic group perform worse when compared to pairs selected without any such demographic restriction (table 2.1). Because identities in the same demographic group are closer together, as seen in fig. 2.3, false accepts are more likely thereby lowering $\text{TPR}_{0.001}$. This idea also explains demographic disparities in matching performance. With the exception of the Indian demographic group, which is too small to source any significant conclusions, minority groups

Pre-trained Facenet						
$TPR_{0.001}$	.9618	.8516	.9594	.8536	.9242	1.000
AUC	.9994	.9977	.9996	.9981	.9995	1.000
$TPR_{0.001}$	.9678	.9436	.9732	.9448	.9742	1.000
AUC	.9995	.9988	.9993	.9998	.9999	1.000
N	10000	5000	10000	2500	1240	20
	Male	**Female**	**White**	**Asian**	**Black**	**Indian**

Same Demographic ☐ Any Demographic

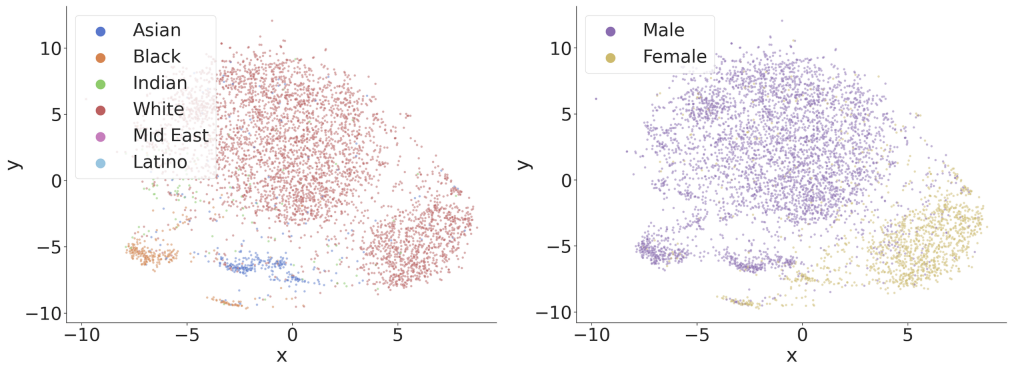
Table 2.1: Matching accuracy on LFW embeddings generated by FaceNet. Pairs within only the same demographic have lower accuracy compared to pairs matching any demographic. N: number of pairs.

have lower matching performance. This is attributed to the tightly clustered embeddings we observe for minority demographic groups.

Intuition from Analytical Model: Understanding why embedding networks have disparate demographic-wise behavior is intractable given the state of current literature in neural network analysis, so we turn to our analytical model and PCA for intuition. Given our interest in studying the impact of the frequency of each group in a training set on the efficacy of the learned embedding network, we formulate a proposition which relates a relative projection distance, our experiments, and our analytical model:

Proposition 2.2. *For fixed* $\boldsymbol{\mu}_{\mathsf{a}}$ *and fixed* $\boldsymbol{\Sigma}_{\mathsf{a}}$*, as* γ *approaches* 0*, the relative projection distance (defined in eq.* (2.7)*) of examples in group* a *increases.*

This proposition suggests minority demographic groups are poorly represented by metric embedding networks. The magnitude of performance disparities can be explained by γ, the parameter capturing demographic imbalance in sampling. For smaller values of γ, the sampled distribution does not represent the natural distribution of the minority group. For the metric embedding network, such poor representation leads the embedding network to



(a) LFW - Race **(b)** LFW - Sex

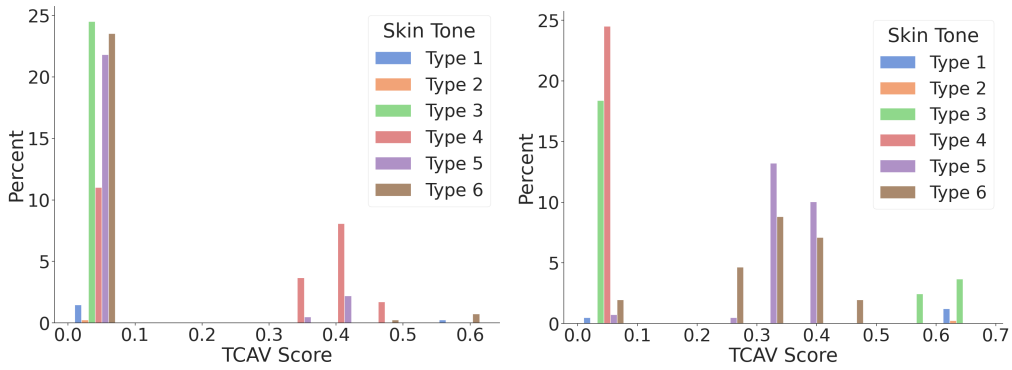
Figure 2.3: t-SNE [42] of LFW embeddings generated by FaceNet.

have trouble discerning identities in the minority demographic group. Hence, we see more errors and ease of impersonation for members of the minority group. Proposition 2.2 is discussed further in our long-form report.[4]

What the Metric Embedding Network Learns: To understand *what* the metric embedding network architecture learns, we plot the resulting embedding structure of FaceNet using t-SNE [42] in fig. 2.3. Embedding networks trained without explicit demographic information can discern demographic groups in the embedding space. For each demographic group, separate clusters appear within the embedding space. The male and female clusters appear almost linearly separable.

Where the Metric Embedding Network Learns: Given the demographic-wise clustering behavior we see in fig. 2.3, we determine the degree to which the network itself contributes to observed clustering behavior. We utilize TCAV [44] to investigate if intermediate layers of the network learn to distinguish demographic attributes. As it was originally defined for classification networks, we retrofit the TCAV framework to metric embedding networks; we associate each identity with an anchor embedding, corresponding to its embedding centroid. We then use the ℓ_2 distance between the embedding of an input face and its anchor embedding to estimate the gradient from the output layer to the relevant activation layer.

[4]Long-form report: arxiv.org/abs/2108.02707



(a) Block 35 Activation 3 **(b)** Conv2d 2a 3x3 Activation

Figure 2.4: TCAV score distribution of 408 identities in VGGFace2. Figures 2.4a and 2.4b learn concepts for lighter and darker skin tones, respectively.

We use skin tones as the demographic concepts. We annotate the skin tones with the Fitzpatrick scale [1], depicted in table A.1. At each activation layer, we train linear classifiers that distinguish the layer activation according each skin tone. Each classifier is trained on 6 skin tones, each containing 75 images. The TCAV score associated with each identity is the proportion of images for which the dot product of this linear classifier and gradient is positive. Nonzero TCAV scores indicate a concept heavily utilized in embedding construction.

Our evaluation involves annotated subsets of the VGGFace2 dataset, one containing 408 identities and another with 4102 identities, with each identity containing 100 images of a person's face. The subset with 408 identities contains a balanced subset of skin tones, whereas the subset with 4102 mainly consists of faces with type 4 and type 5 skin tones. Only the TCAV scores with exactly matching skin-tone annotations are reported in fig. 2.4. We see high utilization of the skin tone concepts by two early layers in the network. We see that darker skin tones are learned separately (Conv2d 2a) than lighter skin tones (Block 35). This suggests metric embedding networks differentiate between skin tones in early layers.

2.3.2 Effectiveness of Obfuscation

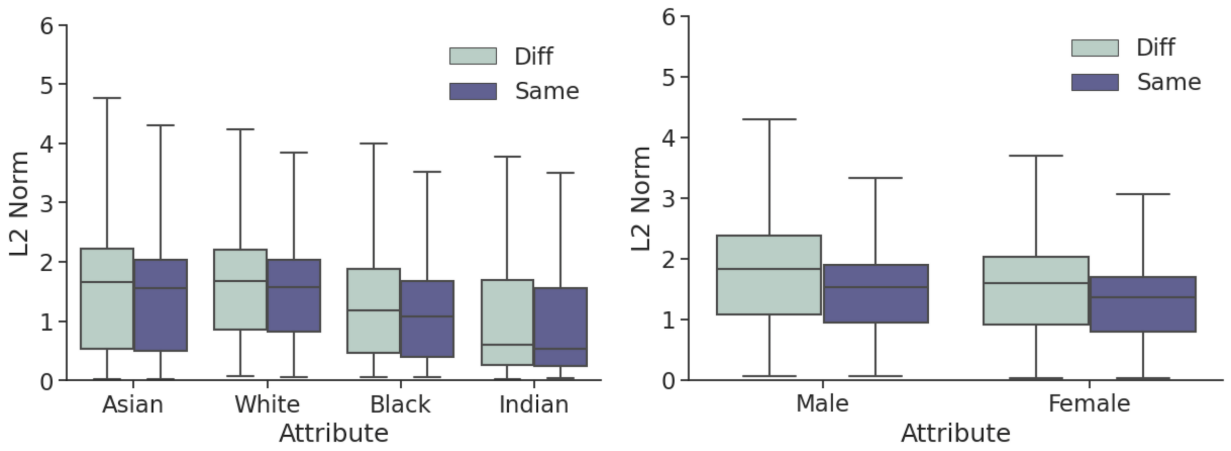
After characterizing the baseline disparities in face recognition systems, we study the subsequent disparities in face obfuscation. We assess, for each demographic group, the difficulty with which identities may be impersonated. We show embedding space geometry induces demographic disparities in the perturbation norm, $\|\boldsymbol{\delta}\|_2$, necessary to successfully impersonate an identity. Further, we study the demographic disparities in the obfuscation success rates. These disparities are present in both the white-box and black-box settings.

Perturbation Norms: To detect disparities in the difficulty of face obfuscation, we examine the norms of perturbations $\|\boldsymbol{\delta}\|_2$, needed to successfully impersonate an identity. Using the adversarial examples generated by targeted attacks, we depict in fig. 2.5 the distribution of perturbation norms conditioned by demographic. We observe that the perturbation strength necessary to successfully impersonate an identity depends on demographic. Further, we observe that for each demographic group, the strength of the perturbation necessary to impersonate an identity is larger if the target identity is in a demographic different from the source identity. We put forth null hypothesis 2.3 which formally addresses demographic disparities in perturbation strength.

Null Hypothesis 2.3.

For each demographic group, the mean perturbation ℓ_2 norm $\|\boldsymbol{\delta}\|_2$ necessary for targeted obfuscation is identical.

A formal analysis rejects null hypothesis 2.3: Differences in perturbation strength for impersonation of identities in the same and different sex demographic group, and the same and different race demographic group are statistically significant: p-values do not exceed 10^{-6}. Furthermore, there is a statistically significant difference between perturbations targeted within the same demographic and perturbations targeted outside the demographic for the Male, Female, Asian and White population groups with p-values of 6.61×10^{-24}, 2.00×10^{-26},



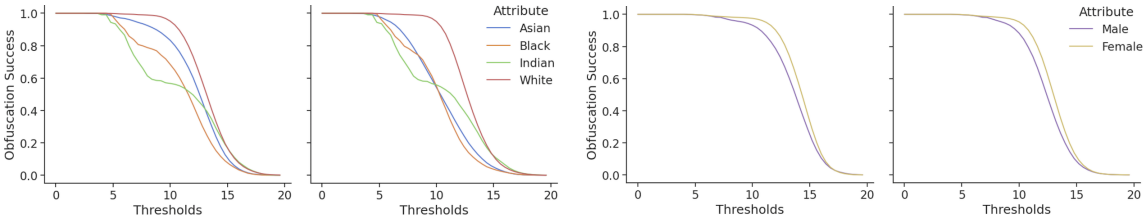
(a) ℓ_2 Norm: Race **(b)** ℓ_2 Norm: Sex

Figure 2.5: Distribution of perturbation norms generated by CW [59] on LFW.

0.00110, and 0.00180, respectively. These are the largest four demographic groups in our dataset. Compared to the Black and Indian minority groups, we expect a more significant difference in perturbation strength necessary to impersonate identities in the same demographic and that which is needed to impersonate identities in different demographics.

Obfuscation Success: Figure 2.6 shows that targeted perturbations' success rates are dependent on the demographic attribute. Faces with the White race attribute exhibit a higher obfuscation success rate. These results also hold in the black box settings: similar trends exist for embeddings generated on OpenFace, Deepface, ArcFace, and VGGFace models. Success rates for the OpenFace model are shown in fig. 2.8, and success rates for the remaining black box models may be found in our long-form report. We observe examples generated on faces from the majority group transfer better than those of minority groups. We conjecture that the larger perturbation norms of such faces contribute to improved transferability rates. This observation is consistent with an observation from Face-off [8]; increasing the norm of perturbations improved transferability to black-box models.

In addition to offline models, this trend also holds for commercially available online face recognition APIs. We test the success of the perturbed faces against the Face++, Amazon



(a) Targeted: Different Race (Left) vs. Same Race (Right)

(b) Targeted: Different Sex (Left) vs. Same Sex (Right)

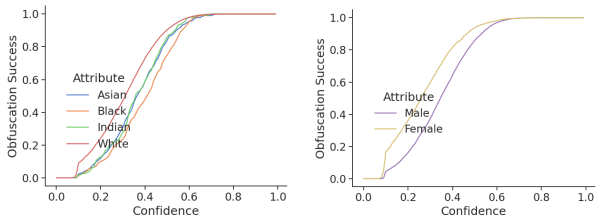
Figure 2.6: Targeted obfuscation success on FaceNet in a white-box setting.

AWS Rekognition, and Microsoft Azure Face APIs. In fig. 2.7, we present the CCDF of confidence scores for untargeted examples generated tested by Azure. For race, on a fixed obfuscation success rate we observe a generally higher confidence in impersonation for White faces. In contrast, Black faces see a generally lower confidence in impersonation for that same obfuscation success rate. Interestingly, for the sex attribute, we generally observe a slightly higher confidence for females. By examining the distance between an embedding and its target, we gain insight into the factor which directly controls obfuscation success rate. We put forth a null hypothesis to formally test our observations.

Null Hypothesis 2.4. *Across all source demographic groups, the distribution of distances between obfuscated embeddings and their targets is identical.*

This null hypothesis is easily rejected: Differences in distance between face embeddings and targets are statistically significant regardless of target demographic: The p-values do not exceed 0.0395. These results confirm the bias in obfuscation performance previously identified. From a practical perspective, our results may suggest that users should select an identity to impersonate of the same race or sex in order to optimize face obfuscation system utility. This is counterproductive to the user's privacy for two reasons: 1) adversarial perturbations with smaller ℓ_p norms will struggle in transferring to other models and 2) a user may leak demographic information to an adversary.

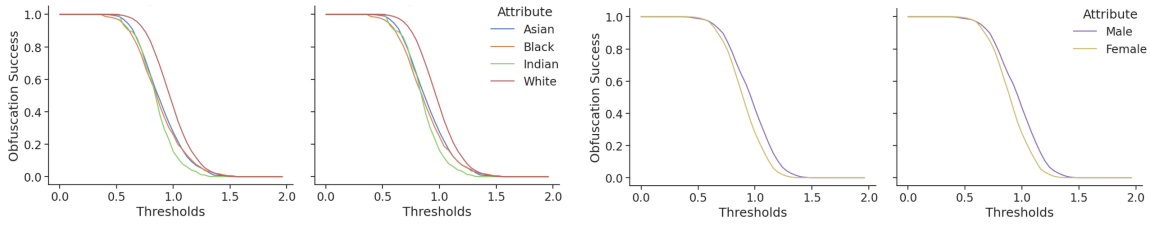
Intuition from Analytical Model: From fig. 2.5, we observe demographic disparities in



(a) Untargeted Azure: Race

(b) Untargeted Azure: Sex

Figure 2.7: Untargeted obfuscation success on Microsoft Azure Face API.



(a) OpenFace: Different Race (Left) vs. Same Race (Right)

(b) OpenFace: Different Sex (Left) vs. Same Sex (Right)

Figure 2.8: Targeted obfuscation success on OpenFace in a black-box setting.

the strength of perturbation necessary to successfully obfuscate an image. We also showed it is generally easier to impersonate identities in the same demographic group than it is to impersonate identities in different demographic groups. This translated into the disparate success rates observed in figs. 2.6 to 2.8. We use our analytical model to provide some mathematical intuition on this phenomenon.

Given an example image $\boldsymbol{x}$ in group g, we study how difficult it is to construct a perturbation $\boldsymbol{\delta}$ such that $\boldsymbol{x} + \boldsymbol{\delta}$ successfully impersonates an identity outside of group g. The analytical model is a natural medium in which quantifying the necessary strength of such a perturbation $\boldsymbol{\delta}$ may occur.

For the sake of simplicity, consider the 1-PCA embedding function $\mathrm{f}_1 : \mathbb{R}^d \to \mathbb{R}$. Let sample $\boldsymbol{x}$ be drawn from the overall synthetic data distribution $\mathcal{D}$ and without loss of generality assume this $\boldsymbol{x}$ is a member of group a. Denote by $p_{\mathcal{Q}}$ the probability density function for a probability distribution $\mathcal{Q}$. We will assume group a is the minority group and so we

assume that $\gamma \leq 1$, $p_{\mathcal{D}_\mathsf{a}}[\boldsymbol{\mu}_\mathsf{a}] > p_{\mathcal{D}_\mathsf{b}}[\boldsymbol{\mu}_\mathsf{a}]$, and $p_{\mathcal{D}_\mathsf{b}}[\boldsymbol{\mu}_\mathsf{b}] > p_{\mathcal{D}_\mathsf{a}}[\boldsymbol{\mu}_\mathsf{b}]$. Given a perturbation $\boldsymbol{\delta}$, we assume that perturbation $\boldsymbol{\delta}$ is norm-bounded and in the direction of group b, i.e. $\frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}$. That is, we assume $\|\boldsymbol{\delta}\|_2 \leq \epsilon$ where ϵ is a non-negative real number. We quantify the values of ϵ for which the following optimization problem is infeasible, thereby guaranteeing $\boldsymbol{x}+\boldsymbol{\delta}$ impersonates[5] an identity in group a:

$$\min_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_2\leq\epsilon} \quad \|\boldsymbol{\delta}\|_2 \qquad \text{s.t.} \quad \mathop{\mathbb{P}}_{\boldsymbol{x}\sim\mathcal{D}_\mathsf{b}}[\mathrm{f}(\boldsymbol{x}+\boldsymbol{\delta})] > \mathop{\mathbb{P}}_{\boldsymbol{x}\sim\mathcal{D}_\mathsf{a}}[\mathrm{f}(\boldsymbol{x}+\boldsymbol{\delta})] \tag{2.8}$$

$$\boldsymbol{\delta} = \eta \cdot \frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2} \text{ where } \eta \in \mathbb{R} \tag{2.9}$$

For notational compactness, we denote:

$$a = \mathrm{f}(\boldsymbol{x}) \text{ and } b = \mathrm{f}\left(\frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}\right).$$

This optimization objective in eq. (2.9) guaranteed to be infeasible when:

$$\epsilon < \max\left\{0, \frac{2b(\gamma-1)\sqrt{\frac{a^2\gamma}{b^2(\gamma-1)^2}} + a\gamma + a + \mathrm{f}(\boldsymbol{\mu}_\mathsf{b})(1-\gamma)}{b(\gamma-1)}\right\} \tag{2.10}$$

This result is further detailed in appendix A.1.

Therefore we conclude that $\boldsymbol{x}+\boldsymbol{\delta}$ impersonates an identity in a when inequality (2.10) holds. Furthermore, note that the bound in inequality (2.10) is not tight. The bound loosens as γ approaches 0. Within inequality (2.10), we notice that for a fixed $\boldsymbol{\Sigma}_\mathsf{a}$, $\boldsymbol{\mu}_\mathsf{a}$, and $\boldsymbol{x}$ which impersonates an identity in group a, as γ approaches 0, the set of perturbations $\boldsymbol{\delta}$ for which $\boldsymbol{x}+\boldsymbol{\delta}$ still impersonates an identity in group a, decreases in size.

Relating to experiments in this section on existing face recognition datasets, it is the disparities in sampling which affect the strength of the perturbation necessary to impersonate an identity in a different demographic group. These disparities are captured by γ. This

[5]Here, we assume it is possible for an example to impersonate itself.

analysis agrees with perturbation norms in fig. 2.5: impersonating an identity in a different demographic is more difficult than impersonating an identity in the same demographic group.

Stability Properties of Face Obfuscation: We further study the impact of γ on the stability of networks. By examining estimates of the local Lipschitz constants, we investigate the stability of metric embedding networks in relation to the demographic distribution of their training sets. A classifier's margin scales inversely with the Lipschitz constant, making classifiers with high local Lipschitz constants less stable and easier to attack [60–64]. We use the RecurJac [65] and Fast-Lin [66] bound algorithms to upper-bound local Lipschitz constants within small neural networks trained on datasets with uniform and non-uniform distributions of demographic groups. The level of uniformity in the demographic distribution is captured by the parameter γ. When the demographics are sampled uniformly, then $\gamma = 1$. If we assume the minority demographic group is a, then greater disparities in sampling mean γ tends to 0. Figure A.2 in appendix A.2.2 shows the distributions of the upper bounds on the estimated local Lipschitz constants for the non-uniform and uniform classifiers trained on 600 identities; this distribution is plotted for the training set. We observe larger upper bounds in non-uniform sampling of each demographic group. Further, identities corresponding to minority demographic groups have larger upper bounds on local Lipschitz constants than do majority identities. These results suggest networks generalize worse for demographic groups which are a minority in the training set, thus networks are less robust to perturbation for certain demographics.

Nearest Neighbors: In addition to performance of source-target matching, we consider the generalized setting in which a source image must be matched to the best candidate among multiple targets. This generalized scenario is studied by measuring accuracy in a nearest neighbors model. In the experiment, accuracy refers to the rate at which embeddings designed to impersonate are still classified as the original source identity. The training points

for the nearest neighbor model are embedding centroids for each identity.

For each demographic group, when impersonating identities in the same demographic group, generated embeddings are accurate on at least 66.7% of examples. For each demographic group, when impersonating identities in different demographic groups, the generated embeddings are accurate on at least 62.9% of examples.

As expected, the accuracy decreases or remains approximately the same for all demographic groups when impersonating identities from different demographics. The general trend is a decrease in accuracy.

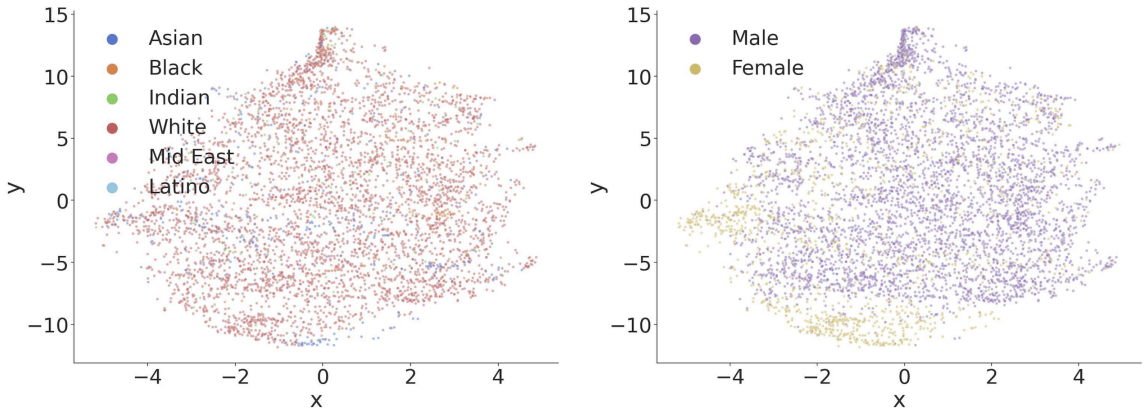
2.3.3 Bias Mitigation

Finally, we examine two bias mitigation techniques and study their impact on the demographic disparities of face obfuscation. The first technique is a training procedure designed by Xu et al. [32] to promote fairness, and the second technique is dataset balancing by demographic. Models of the FaceNet architecture, outputting embeddings in $\mathbb{R}^{512}$ are trained. Because we are interested in characterizing the best-case obfuscation performance in the presence of bias mitigation strategies, we focus exclusively on the white-box setting.

Training Procedure of Xu et al.

Xu et al. propose a technique, that within the context of face obfuscation, is designed to mitigate disparate susceptibility of individuals to impersonation. The training technique decomposes overall error into the natural error and boundary error. *Natural error* refers to the error on examples prior to the addition of a perturbation, and *boundary error* refers to the net increase in error induced when perturbing the natural example. The Xu et al. training procedure incentivizes training of an accurate model such that natural error and boundary error are each roughly equivalent across all identities.

Visualizing Embeddings: Upon examining the t-SNE plots in fig. 2.9, we observe the



(a) LFW - Race **(b)** LFW - Sex

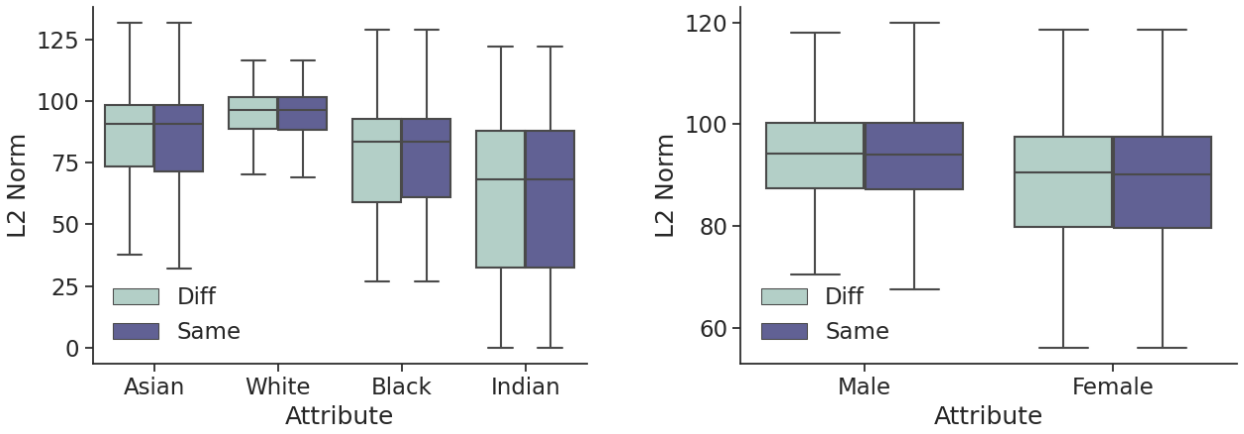
Figure 2.9: t-SNE of the embeddings for the LFW dataset. Embeddings are generated on a FaceNet model trained by Xu et al. procedure.

training procedure of Xu et al. does mitigate sources of bias in the embedding space geometry: Demographic clusters are less distinct and embeddings for each demographic group are somewhat interspersed amongst each other. However, clustering is not eliminated entirely; there is a cluster of females in fig. 2.9b, though many female embeddings are interspersed among embeddings corresponding to males.

Analyzing Obfuscation: Embedding space geometry impacts the perturbation norms needed for impersonation: Compared to the pre-trained model, we see less disparity between the perturbation norms needed to impersonate identities in the same and different demographic groups in fig. 2.10. We put forth a formal null hypothesis to assess our observations.

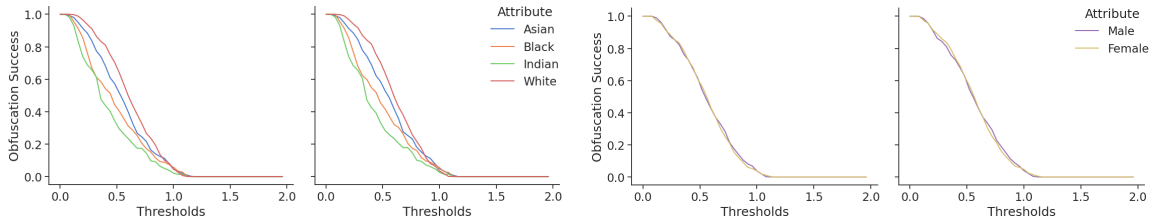
Null Hypothesis 2.5. *For the model trained by the Xu et al. procedure, the mean perturbation ℓ_2 norm $\|\boldsymbol{\delta}\|_2$ necessary to impersonate an identity in the same demographic is identical to the mean perturbation ℓ_2 norm $\|\boldsymbol{\delta}\|_2$ necessary to impersonate an identity in a different demographic group.*

Though we perceive a significant reduction in bias due to the Xu et al. training procedure, we can partially reject null hypothesis 2.5. There is a statistically significant difference



(a) ℓ_2 Norm: Race **(b)** ℓ_2 Norm: Sex

Figure 2.10: Distribution of perturbation norms generated by the CW attack on FaceNet model trained with Xu et al. procedure.



(a) Different (Left) Race vs. Same (Right) Race **(b)** Different (Left) Sex vs. Same (Right) Sex

Figure 2.11: Targeted white-box obfuscation success on a FaceNet model trained with Xu et al. procedure.

between perturbations targeted within the same demographic and perturbations targeted outside the demographic for only the White demographic, with a p-value less than 10^{-6}.

We also observe that the perturbation strength necessary to impersonate an identity has a dependence on demographic. We put forth a formal null hypothesis to test this observation:

Null Hypothesis 2.6. *For the model trained by the Xu et al. procedure, the mean perturbation ℓ_2 norm $\|\boldsymbol{\delta}\|_2$ necessary to impersonate an identity is identical for each source demographic group.*

As anticipated, we can reject this null hypothesis. Differences in perturbation strength between demographic groups are statistically significant for impersonation targeting iden-

tities in the same and different sex demographic groups, and the same and different race demographic groups. The p-values do not exceed 10^{-6}.

The significance of such disparities manifests itself in obfuscation obfuscation success rates as depicted in fig. 2.11, especially for success rates conditioned by race. Much like the pre-trained FaceNet, the White race group tends to have the highest obfuscation success rate. Looking at the sex demographic groups, obfuscation success rates appear to be similar. It is unclear what structurally about Xu et al procedure yields a model with obfuscation success rates disparate with respect to race, yet obfuscation success rates conditioned by sex are similar. For completeness, we put forth a null hypothesis to test the significance of this observation:

Null Hypothesis 2.7. *For the model trained by the Xu et al. procedure: Across all source demographic groups, the distribution of distance between each obfuscated face embedding and its target is identical.*

We partially reject this null hypothesis: Only when targeting identities outside the identity of source identity demographic group, differences in distance between face embeddings and targets are statistically significant. The p-values are each less than 10^{-6} when targeting identities in a different race and sex. Given these results, it is apparent that the Xu et al. training procedure does mitigate bias to some extent, but is far from a perfect solution.

Accuracy Trade-off: The observed bias mitigation also comes at the cost of model accuracy. We measure these costs by examining the matching performance and through accuracy of a nearest neighbors model. The matching performance, $\text{TPR}_{0.05}$, of a model trained with the Xu et al. procedure is depicted in the middle subtable of table 2.2. Matching performance of the reference FaceNet is depicted in the upper subtable of table 2.2. The matching performance across most demographics is reduced by 1-4% compared to the reference model. When measuring training accuracy of a nearest neighbors classifier, conditioned by demographic, the lowest accuracy in a model trained with the Xu et al. procedure is 55.2%. This

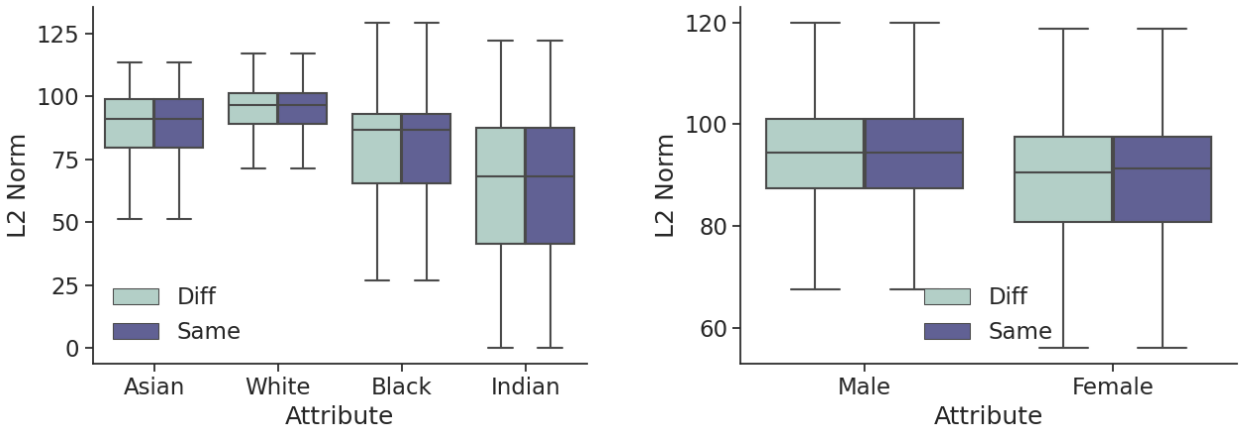
compares poorly to the reference FaceNet model, for which the worst natural accuracy, when conditioning by demographic, is 95.4%. Full results for nearest neighbors accuracy can be found in our long-form report.

Dataset Balancing

The second bias mitigation strategy we study is that of training face recognition models on demographically balanced datasets. To do so, we train three metric embedding networks of the FaceNet architecture. Each training procedure is identical, with the exception of the training dataset. The network trained on the first dataset acts as the reference: It is trained on the entirety of the VGGFace2 [24] training split, a total of 8631 identities. Experiments for this model appear in the appendix.

The remaining two datasets are used to train comparison models. One such dataset is Sex-Balanced VGGFace2, which contains a total of 4866 identities. The Sex-Balanced VGGFace2 dataset was created by removing original identites to create a training set which contains an equal number identities for the demographic of interest. Instead of performing data augmentation, we remove examples from VGGFace2 so as to not introduce any artifacts. Similarly, we construct a third dataset, which consists of 307 identities for each of the race subgroups. We call this dataset, containing 1842 total identities, Race-Balanced VGGFace2. Race-Balanced VGGFace2 and Sex-Balanced VGGFace2 are collectively referred to as "(demographically) balanced datasets". Race-Balanced VGGFace2 and Sex-Balanced VGGFace2 are the training sets used to obtain Race-Balanced FaceNet and Sex-Balanced FaceNet, respectively. We observe that the Balanced FaceNets have less demographic-wise disparity in both face recognition and face obfuscation performance.

Visualizing Embeddings from Models: Upon examining the t-SNE plots in fig. 2.13, we notice that minority demographic groups have larger clusters than do the clusters for the same demographic groups generated on the reference model. Compared to embeddings generated



(a) Race-Balanced FaceNet on race data **(b)** Sex-Balanced FaceNet on sex data

Figure 2.12: The distribution of adversarial perturbation sizes generated using the CW attack on FaceNet trained on demographically balanced datasets.

by the model trained with the Xu et al. procedure, demographic clusters generated by Sex-Balanced and Race-Balanced are more separate and distinct. Further, each demographic group generally appears more distinct.

Analyzing Obfuscation: Given the distinct demographic clusters we see in balanced embedding space geometry, we do not expect to see much bias mitigation. Before drawing such a conclusion, we formally test a null hypothesis:

Null Hypothesis 2.8. *For models trained on balanced datasets, the mean perturbation ℓ_2 norm $\|\boldsymbol{\delta}\|_2$ necessary to impersonate an identity in the same demographic is identical to the mean perturbation ℓ_2 norm $\|\boldsymbol{\delta}\|_2$ necessary to impersonate an identity in a different demographic group.*

We are able to partially reject null hypothesis 2.8. There is a statistically significant difference between perturbations targeted within the same demographic and perturbations targeted outside the demographic for the only male, female and white demographic groups with p-values of 3.20×10^{-6}, 5.28×10^{-5}, and 1.65×10^{-30}, respectively.

We observe disparities in the perturbation norms required for faces in each demographic group to be successfully obfuscated. We propose a null hypothesis to test this observation:

Null Hypothesis 2.9. *For models trained on balanced datasets, the mean perturbation ℓ_2 norm $\|\boldsymbol{\delta}\|_2$ necessary to impersonate an identity is identical for each source demographic group.*

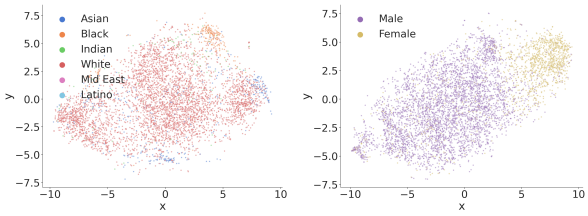
We are also able to reject this null hypothesis: For models trained on both sex-balanced and race-balanced data, differences in perturbation strength between demographic groups are statistically significant for the same sex demographic group, different sex demographic group, same race demographic group, and different race demographic group. The p-values do not exceed 10^{-6}.

The disparities in perturbation norm also manifest themselves in obfuscation success rates as depicted in fig. 2.14. Such success rates are disparate when conditioned by both sex and race. Much like the pre-trained FaceNet model, when conditioned by race, the White demographic group has the highest success rate. Unlike the pre-trained FaceNet, the Female demographic group has higher obfuscation success rate than the male demographic group. It is unclear why the Female demographic group now has higher obfuscation success rates. We put forth a null hypothesis to formally test our observations:

Null Hypothesis 2.10. *For models trained on balanced data: Across all source demographic groups, the distribution of distance between each obfuscated face embedding and its target is identical.*

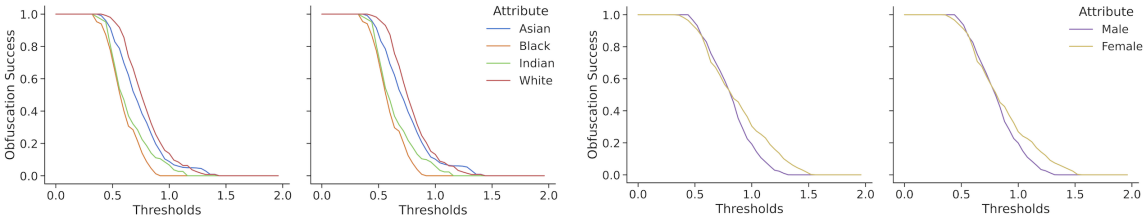
The null hypothesis is easily rejected: Differences in distance between face embeddings and targets are statistically significant in all settings. The p-values do not exceed 10^{-6}. Given the significant disparities in face obfuscation performance, it is apparent dataset balancing is ineffective at mitigating bias in face obfuscation.

Natural Accuracy Trade-off: The matching performance, $\text{TPR}_{0.05}$, of a models trained on balanced datasets is the lowest sub-table in table 2.2. The matching performance across most demographics is reduced by 0.5-3% compared to the reference model. When measuring



(a) LFW - Race **(b)** LFW - Sex

Figure 2.13: t-SNE of embeddings for the LFW dataset. Figure 2.13a is generated on Race-Balanced FaceNet. fig. 2.13b is generated on Sex-Balanced FaceNet.



(a) Different (Left) Race vs. Same (Right) Race. Race-Balanced FaceNet

(b) Different (Left) Sex vs. Same (Right) Sex. Sex-Balanced FaceNet

Figure 2.14: Targeted obfuscation success evaluated on Demographically Balanced FaceNets in a white-box setting.

training accuracy of a nearest neighbors classifier, conditioned by demographic, the lowest accuracy in a model trained with demographically balanced training data 94.5%. This is only a very slight performance reduction relative to the reference FaceNet model.

2.4 Discussion

Potential Remedies: In section 2.3.3, we discussed the impact of two bias mitigation strategies on face obfuscation. First, we examined a model training procedure by Xu et al. This procedure does not explicitly optimize for fairness, but instead implicitly relies on a documented connection between adversarial learning and fairness [57]. The second bias mitigation procedure involved training FaceNet models on balanced datasets. The Xu et al. procedure mitigated more bias than the balanced FaceNet models, but neither procedure completely mitigated disparities in face obfuscation performance. A training procedure which

Reference FaceNet Model						
$TPR_{0.05}$	.9046	.8620	.9164	.9152	.8565	.9000
AUC	.9808	.9864	.9826	.9820	.9711	1.000
$TPR_{0.05}$	.9164	.9328	.9294	.9720	.9919	.8000
AUC	.9828	.9864	.9847	.9938	.9967	1.000
Training Procedure of Xu et al.						
$TPR_{0.05}$	.8984	.8180	.8952	.8776	.8258	.6000
AUC	.9774	.9631	.9782	.9796	.9644	1.000
$TPR_{0.05}$	.9036	.9408	.9104	.9768	.9806	.9000
AUC	.9806	.9866	.9816	.9945	.9955	1.000
Balanced Training						
	Sex-Balanced		**Race-Balanced**			
$TPR_{0.05}$	.9184	.8804	.8716	.9160	.8726	.8000
AUC	.9837	.9733	.9719	.9825	.9738	.9800
$TPR_{0.05}$	.9330	.9244	.8978	.9672	.9823	.7000
AUC	.9867	.9842	.9772	.9938	.9950	1.000
N	10000	5000	10000	2500	1240	20
	Male	**Female**	**White**	**Asian**	**Black**	**Indian**

Same Demographic ☐ Any Demographic

Table 2.2: The matching performance on LFW on a reference FaceNet model (top), FaceNet trained with Xu et al. procedure (model), and FaceNet models on demographically balanced data.

better mitigates performance disparities in face obfuscation efficiently, is a topic for future work.

Beyond bias mitigation in the manner we studied in this chapter, balancing performance across combinations of multiple demographic attributes might be of interest. There exists work addressing this problem: Serna et al. propose Sensitive Loss, a "discrimination-aware" Triplet Loss tuning procedure for pre-trained models [67]. This tuning procedure involves adding a layer to the network, then tweaking this layer using only identities drawn from the same demographic group.

Threats to Validity: The intent of section 2.2.2 is the creation of a tractable analytical

model for embedding spaces and dimensionality reduction in general. Perhaps the most significant limitation of this analytical model is that PCA is a linear embedding function and is incapable of capturing non-linear effects present in metric embedding networks. Another threat to validity is the scarcity of publicly available face recognition datasets; datasets with labeled demographic attributes are even more rare. Such datasets are rather small, as such, our results may not generalize to datasets orders of magnitude larger than the available face recognition datasets.

2.5 Related Work

Buolamwini and Gebru study commercially available face recognition datasets and classifiers [30]. Their findings indicate that prominent commercial face classifiers exhibit disparate performance across demographic groups. Follow-up research has attempted to address these demographic disparities through balanced datasets [68, 69]. Such datasets improve model generalization but do not resolve disparities completely. These findings are consistent with results in appendix A.2.

Demographic disparities seem to contribute to the phenomenon of overlearning, a term coined by Song and Shmatikov [70]. Overlearning refers to the phenomenon where models implicitly learn to recognize sensitive patterns not part of the original learning objective. Metric embedding networks trained on faces do overlearn; they learn a demographically-aware dimensionality reduction on faces.

Cherepanova et al. [71] study the relationship between training data, test data, population demographics and face recognition performance. The authors, independent of our results, observe that models trained on demographically balanced datasets are not bias-free. They discuss neither face obfuscation, nor impacts on privacy and security.

Nanda et al. discuss the relationship between fairness and robustness of face recognition [31]. Our work differs in several respects: First, there is a distinction between the study

of robust metric learning. This chapter explores bias present in the embedding space and how it affects recently proposed face obfuscation systems. Through an analytical model in section 2.2.2, we show the amount of tolerable perturbation in a face recognition system depends on class imbalances. Second, we examine face obfuscation in the black-box setting; these experiments are performed on both open-source and commercial models as depicted in fig. 2.7. The results demonstrate transferability of biases in adversarial perturbations; we note the negatively-affected demographics can differ between the white-box and black-box settings depending on the perturbation strength. Third, we investigate root causes for demographic-wise performance disparities via TCAV. We also explore dataset balancing and the Xu et al. training procedure as bias mitigation techniques.

Cilloni et al. study a targeted face obfuscation which attempts to find the minimum strength successful perturbation via optimization procedure regularized by the a well-established image similarity scoring function known as DSSIM. Modulo the regularizer, our Carlini-Wagner based experiments achieve the same purpose.

Rajabi et al [72]. consider two methods of face obfuscation: the first is a universal ensemble perturbation, an untargeted obfuscatory method designed to be transferable; we explore transferability of untargeted perturbations in section 2.3.2 and fig. 2.8. The second technique is cryptographic in nature. Finally, a Master's thesis by Qin [73] evaluates discrepancies in a face obfuscation system called FAWKES [9], conditioned on skin tones. Qin finds differences in perturbation visibility for certain demographics. In comparison, our work characterizes these discrepancies and their impact on face obfuscation systems.

2.6 Conclusion

Face recognition systems have seen increased usage in online settings at the cost of heightened privacy concerns. Researchers have proposed face obfuscation systems that leverage evasion attacks against metric embedding networks. Our results show that, in an effort to

mitigate such privacy concerns, face obfuscation systems have performance characteristics that depend on demographic information, thereby creating a new privacy incursion. Such performance characteristics can not only leak demographic membership information and decrease the performance of face obfuscation among underrepresented demographic groups. Imbalances in training set demographics are only partially to blame for this privacy leak: remaining causes are yet to be discovered. To mitigate the effects of this incidental privacy leak, we must develop both loss functions for training fair metric embedding networks and techniques to characterize if such privacy leaks will occur.

Chapter 3

Synthetic Counterfactual Examples for Face Recognition Systems

Automated face recognition technology has rapidly expanded across various industries, including commercial and governmental domains. These systems enable many applications, such as identifying individuals on social media, locating missing persons, assisting in law enforcement and surveillance activities, and authenticating personal identities [20, 21]. This rapid adoption benefited from significant advances in face recognition systems, such as Amazon Rekognition, and the wide availability of labeled facial datasets [23, 74].

While these systems are often evaluated for the average accuracy on available datasets, more is needed to learn about their robustness and fairness against distributional shifts in input data. As such systems are deployed in applications with significant societal impact, ensuring their ethical and reliable use is essential. We posit that characterizing and improving the performance of the deployed systems requires understanding how they make decisions. Towards that end, we explore counterfactual explanation techniques, which aim to generate human-understandable input modifications that would have resulted in a different output decision by the face recognition system. By auditing which attributes of the input data were most influential in the system's decision-making process, we can identify unintended biases

and failure modes in face recognition systems. These explanations would also provide insights into improving face recognition performance, which enables a more transparent deployment.

We propose a new method to generate counterfactual examples for face recognition systems. The largest hurdle towards counterfactual analysis of existing face recognition systems is generating realistically modified faces that cover a range of demographic and semantic attributes. Sampling natural inputs that satisfy this condition, such as faces with different skin tones, lighting conditions, hairstyles, or accessories, is nearly impossible. We address this limitation by utilizing recent innovations in Text-to-Image generative models to semantically change faces until the face recognition system reaches a different decision. This method involves generating a large set of identities with diverse demographic attributes, generating multiple face images for each identity, and applying different semantic attributes to the generated faces.

To the best of our knowledge, this chapter represents work amongst the first to: (1) *provide an end-to-end pipeline*, from a text-to-image diffusion model, to generate synthetic faces annotated with fine-grained attributes at a large scale; (2) *evaluate the quality* of the generated faces using vision question answering (VQA) models and a user study; and (3) *perform counterfactual evaluation* of two commercial face recognition systems: AWS Rekognition and Face++. We also provide a theoretical narrative which explains the observed model behavior.

Our evaluation shows that both face recognition systems exhibit performance disparities under semantic changes, with significant differences across gender and ethnic groups. Unusual features like facial hair on women or long hair on men notably impacted accuracy. Amazon Rekognition consistently performed better, though both systems struggled with specific demographics, especially Face++ with Asian faces. These insights would not have been possible without counterfactual examples.

3.1 Background and Related Work

In this section, we discuss prior research in characterizing face recognition, generating counterfactual examples, and synthesizing face datasets.

3.1.1 Failure in Face Recognition

Race and gender bias in facial recognition has been studied extensively over the years [14, 75–78]. It is widely reported that face recognition systems can be biased towards male and light-skinned faces while demonstrating the lowest accuracy on dark-skinned female faces [14, 15, 78]. Other facial semantics like age [78–80], pose [81] and hair [80, 82] have also been shown to contribute to facial recognition performance. Terhörst et al. [83] does a comprehensive analysis of 47 different attributes using the manually annotated MAAD-Face dataset [84].

The prominent natural datasets used to train face recognition models include LFW [33], CASIA WebFace [85], VGGFace [74, 86], and Flickr-Faces-HQ [87]. These datasets are sourced from the web, mostly contain celebrity faces, and can be biased towards certain demographics [88]. Many balanced datasets have been proposed [89–91] to overcome problems with the above datasets. However, even a demographically balanced dataset can show disparate facial demographic performance [92] due to limiting factors like lightning, pose, and image quality.

Our contributions: Our work differs from Terhörst et al. in two aspects: (i) we use a synthetic dataset ensuring a balanced coverage of different attributes, and (ii) we study the performance of online API systems, instead of open-source models trained on limited datasets.

3.1.2 Counterfactual Explanation

Auditing is Challenging. The inner workings of neural networks are neither human-understandable nor theoretically tractable. Consequently, the best way to understand model performance is empirical validation. In traditional machine learning settings, practitioners understand the model performance evaluating it on a validation set. This validation set follows the same distribution as the training data. This approach provides no insight into performance on out-of-distribution data, yet machine learning models often operate on out-of-distribution data.

Counterfactual Explanation. One way to achieve this goal is through explainability methods, which explore the failure modes and spurious correlations inherent in the model. Counterfactual explanation (CE) is a post-hoc technique that aims to perform hypothetical input modifications that would have resulted in a different decision by the model. By revealing which features of the input data were most influential in the model's outcome, CE identifies unintended biases and spurious correlations and provides insights into how to improve the model's performance.

Recent works [93, 94] have explored diffusion models for generating counterfactual examples, aiming to identify spurious correlation and failure modes in vision classifiers trained on ImageNet. Vendrow et al. [94] defined 23 shifts such as 'at night,' 'blue,' 'in the beach,' 'in the snow,' and 'sketch.' They performed textual inversion, a few-shot fine-tuning step, to encode these shifts in the diffusion model, limiting the method's scalability. They then explored the failure modes of the model under these 23 shifts. On the other hand, Wiles et al.[93] explored the failure modes using an automated approach. They utilized a diffusion model to generate test images, clustered semantically close errors, and employed a captioning model to label these failure clusters. However, it's worth noting that neither study explored the application of face recognition models, where demographic disparity and the intricacies

of facial features pose unique challenges for CE edits.

CE for Face Applications. Conventional studies [95–97] that perform CE on faces focus on feature classifiers such as *Young* or *Smile* classifier. In the context of face recognition, CE often involves limited features, such as the mouth, eyes, or eyebrows, inpainting using a mix of images. Other works focus on the creation of feature saliency maps [98]. These maps highlight facial regions crucial to the recognition decision but may fall short of revealing the specific changes to regions that would alter the decision.

Our contributions: In this work, we edit the semantic attributes of faces, for example, by growing facial hair to a clean-shaven face. We pass the transformed image to the face recognition model and analyze the change in model outputs as a function of the change in semantic attributes. By design, we associate each face with fine-grained semantic and demographic labels.

3.1.3 Synthetic Face Datasets

Existing natural face datasets suffer from many limitations with respect to demographic diversity, semantic attributes annotation, resolution, dataset size, and data privacy [99,100]. These limitations have motivated our choice for composing synthetic datasets to generate counterfactual examples. Two architectures dominate the face generation landscape: GANs (Generative Adversarial Networks) and Diffusion Models. StyleGAN [87, 101] is among the best-known face generation architectures. While StyleGAN is a capable image generator, it has several drawbacks: editing existing images is tremendously difficult, GAN inversion is non-trivial, and original images may not be recoverable. Furthermore, there is no user-friendly method to induce specific semantics on a face. Importantly, GAN images exhibit low variability [102], limiting their utility.

Diffusion models [103,104] are another architecture used in face generation. They rely on

internal Gaussian randomness and the construction of Markov Chains to construct images. Many diffusion models utilize text guidance. Such prompting provides a highly functional user interface for generating images with fine-grained semantics. Further, diffusion models are more easily invertible than GANs, enabling more tractable image editing. Recent works utilized diffusion models to generate faces with different styles (DCFace [105]), poses, accessories, and expressions (GANDiffFace [106]), and age [107]. However, these models must be fine-tuned on, usually synthetic, face datasets representing the required style and semantics.
Our contributions: We perform large-scale one-shot generation of face images with various semantic changes and diverse demographic representation. Fine-tuning or inverting the diffusion model for each identity or transformation does not scale. As such, we devise a novel pipeline that employs semantically guided attention [108] and name prompts, as will be described in the Framework section (section 3.2).

3.2 Framework

We have developed a framework to produce targeted counterfactual examples, which reflect various semantic transformations at a large scale.[1] This framework allows for the automated evaluation of face recognition model performance and the identification of biases and failure cases. The framework consists of three main components: (1) facial semantic attributes definition, (2) counterfactual images generation pipeline (depicted in fig. 3.1), and (3) face recognition models assessment.

3.2.1 Semantic Features Definition

Existing research [14, 78] has examined sex and ethnicity as the main attributes affecting face recognition. The problem with such an approach is that demographic attributes can be ambiguous and are not necessarily the root causes affecting face recognition. Terhörst et

[1] We will make the source code of the framework and generated dataset publicly available.

al. were the only ones to analyze a wide range of attributes [83], such as face accessories and age. However, their analysis considered annotations from a dataset [84] that can be subjective, such as "attractive face." When analyzed attributes are not measurable, it is hard to determine how to correct the failure modes in face recognition. Prior research sheds light on this phenomenon [109]: faces labeled as "Indian" exhibit high variance in face recognition performance, which suggests ethnic labels do not fully represent facial diversity of the group.

Therefore, we treat demographic attributes as a "high-level" factor contributing to face recognition performance. Additionally, we study "low-level" semantic attributes that are quantifiable, objective, independent of demographics, and pertinent to any human face. Through this approach, we can investigate how combinations of these semantic attributes and demographic factors affect the performance of face recognition systems.

Demographic Attributes. Similar to leading studies on face recognition fairness [14,78], we consider ethnicity and sex as the defining factors for the demographic group. We consider two labels for the sex attribute: 'Male' and 'Female.' As for ethnicity, we consider four labels: 'Black,' 'East Asian,' 'Indian,' and 'White.' Therefore, we study eight distinct ethnicity-sex combinations.

Semantic Attributes. We create a thorough set of attributes by incorporating insights from prior research [83] and drawing from research in face morphology [110]. This set includes the following groups of attributes: accessories, age, skin complexion, nose shape, eye shape, face proportions, facial expression, facial hair, hair color, and hairstyle. The full list of attributes is in the second column of table 3.2.

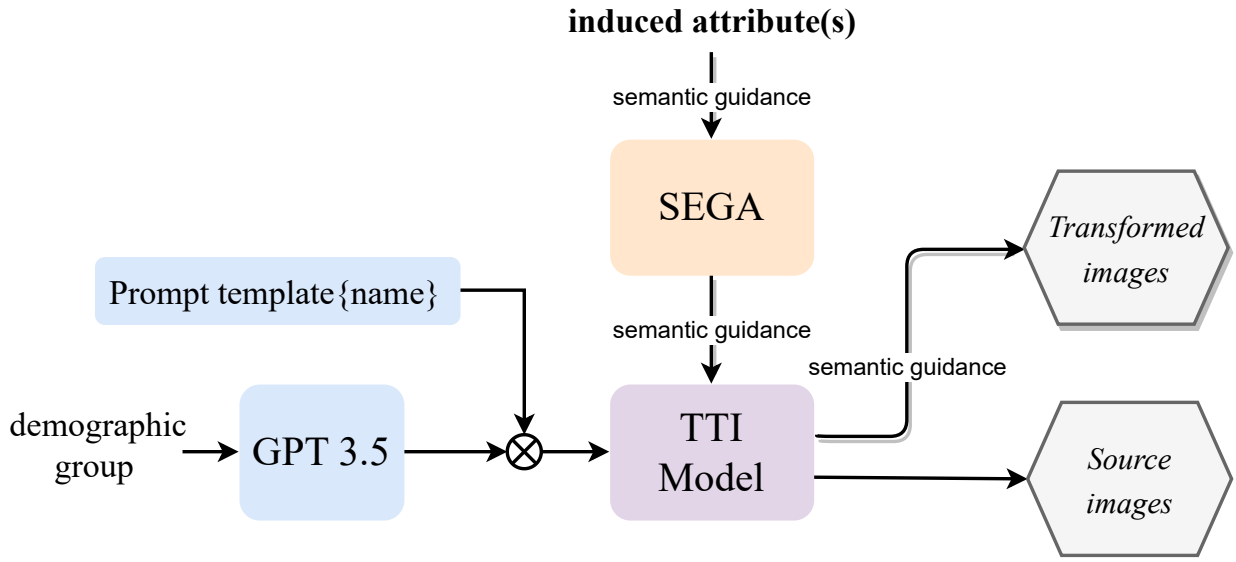


Figure 3.1: Our counterfactual image generation pipeline.

3.2.2 Counterfactual Examples Generation

The second stage of our framework is a pipeline to generate *Source Faces* and *Transformed Faces*. The Source Faces correspond to demographically-aware synthetic faces. Their counterfactual examples, given a set of attributes, are the Transformed Faces. Realizing this pipeline requires (1) generating identities with diverse demographic attributes, (2) generating multiple face image variations for each identity, and (3) applying semantic attributes to the generated faces while preserving their identity.

We generate synthetic faces using a Text-to-Image (TTI) diffusion model: a finetuned Realistic Vision Model,[2] hereafter referred to as *Realism*. Realism is among the many openly available finetuned models from the checkpoints of the Stable Diffusion (SD) model for face image generation. While SD is trained on the open-source LAION-5B dataset, Realism's exact fine-tuning details are not known. Our framework is agnostic to the generative model, and we treat it as a grey-box component.

Our counterfactual face-generation pipeline takes two inputs: the demographic group and

[2] `https://huggingface.co/SG161222/Realistic_Vision_V4.0_noVAE`

the counterfactual semantic attributes. It employs customized prompts to generate identities belonging to the demographic group, referred to as '*Source Faces*.' For each identity, the pipeline generates multiple images to diversify the conditions of Source Faces, including different poses, lighting, backgrounds, expressions, etc. Finally, it utilizes image editing tools [108] to apply the specified semantic attributes to the images of each identity. The resulting faces, referred to as '*Transformed Faces*,' resemble a natural face image dataset that is both demographically balanced and semantically rich. fig. 3.2 provides an example of a synthetic dataset with eight identities corresponding to the eight ethnicity-sex combinations and multiple images per identity, each representing a different semantic attribute. We explain in section 3.2.2 why we synthesize source face images instead of using natural ones.

Below, we break down the details of the counterfactual pipeline, which includes prompt design, identity selection, source and transformed face generation, and manual validation of generated images.

Prompt Design

TTI (text-to-image) generators, especially diffusion models, are highly sensitive to prompts [111], posing a challenge for aligning images to text (intent). In the context of face generation, an optimal prompt should satisfy three criteria: (i) encode the concept of identity: for a fixed prompt, the model consistently generates images depicting the same *facial* identity despite randomness and stochasticity of the generation process, (ii) encode the high-level concepts of a demographic group, including ethnicity and sex, and (iii) maintain good perceptual image quality, avoiding cartoonish, anime-style, fantasy, or cropped representations. However, currently, there are no computationally efficient approaches to align the prompt with these intents. Known prompt design strategies, such as prompt space search [112] and text-inversion [113], are too slow or highly constrained, making them inappropriate for the task of large-scale generation of diverse face images.

Prior works employing TTI diffusion models for large-scale face image generation [114–117] utilize prompts specifying demographic information, such as '*A photo of the face of an Asian man*,' or '*Photo portrait of an Asian man*.' While these prompts encode identity and high-level concepts, the generated images lack diversity, the faces look quite similar, undermining the model's potential to generate distinct identities [17]. Most importantly, they fail to recall and generate images of a specific identity. This criterion is crucial for generating the Transformed Faces; the prompt has to encode an underlying identity.

To address this issue, we follow the work of Rosenberg et al. [17], where they use names in the prompt to encode an identity associated with a certain demographic group. Specifically, we use this prompt template: '*A photo of the face of* ⟨*name*⟩,' where ⟨name⟩ can be *Emily Thompson* for a white woman identity, or *Raj Patel* for an Indian man identity, for example. For each demographic group, we compile a list of celebrity names associated with the group to replace the ⟨name⟩ placeholder. Our empirical evaluation confirms that including a name within the prompt effectively encodes both identity and demographic group.

Identity Selection

Diffusion models are typically trained on web-scraped datasets, where a significant proportion of face images come from celebrities. They demonstrate better composition of, and can even memorize, their training data distribution [118]. Hence, we employ celebrity names that show frequently in Stable Diffusion's training data to guide identity generation.

Stable Diffusion [104] (the base of the Realism model) is trained on the LAION-5B [119] dataset, which has 5 billion images, including human faces, among many other images and scenes. While LAION-5B is an open-source dataset, directly extracting face images or querying the dataset captions to identify frequently occurring celebrity names is a non-trivial task. Consequently, we adopt a multi-step approach to perform this task. First, we compile an extensive list of celebrity names representing the eight demographic groups using a combi-

nation of IMDB celebrity lists and LFW [33], a standard faces dataset. Second, we integrate these names into the previously described prompt template. Third, we use a CLIP-based retrieval tool[3] to search LAION-5B using the prompt template. This tool measures the CLIP embedding similarity between the prompt and LAION-5B images and their captions, returning images from LAION-5B that are semantically similar to the provided prompt. To further refine the training data, we filter out images with aesthetic scores lower than five. Finally, for each demographic group, we select 50 identities (celebrity names)[4] with the highest number of retrieved instances where the celebrity name appears in the caption.

Source Faces Generation

We instruct the Realism model using the prompt template to generate six images for each of the 50 identities per demographic group. Due to the inherent randomness of the generation process, the generated images would vary across multiple conditions. However, we observed that some variations might drift slightly from the underlying identity. Additionally, prior work [120] reports on the significance of the generator's random seed choice in ensuring high image quality and high prompt fidelity.

For better control over the Source Faces fidelity, we generate 20 images per identity using 20 random generator *seed numbers*. We filter out the outlier images for each identity and retain the identities whose images exhibit high semantic consistency. To identify outliers, we calculate the similarity scores of the 20 images using a face embedding network, FaceNet [7]. We then keep the six images (along with their corresponding seeds) that have the highest average similarity scores. Subsequently, we select ten identities out of the initial 50 based on the highest average similarity across their six images. As a result, the Source Faces dataset comprises eight demographic groups, each with ten identities and six images per identity, resulting in a total of 480 images. The prompt template, identity name, and seed

[3] `https://rom1504.github.io/clip-retrieval`,

[4] The lists of names will be provided in the supplementary material.

selection are essential in reproducing the semantics of the source images while applying the transformations on the Transformed Faces.

Transformed Faces Generation

We employ a latent manipulator to introduce the semantic attributes (section 3.2.1, table 3.2) into the generated Source Faces, such as incorporating sunglasses onto the face. In particular, we use the semantic-guidance image generation technique SEGA [108]. SEGA steers the TTI model towards generating images that incorporate specific semantic concepts based on user-provided textual edits while keeping the rest of the image semantics intact, all without the need to finetune the TTI model. SEGA integrates with any generative model and allows for subtle and extensive edits, making it a scalable and user-friendly technique.

We integrate SEGA with Realism to apply a range of semantic attributes to the Source Faces. We prompt Realism with the exact template, identity names, and seeds used to generate the source images and instruct SEGA with the required attribute as an edit concept. As SEGA does not support image inversion (mapping an image to its noise latents), we could not have used natural source images. Instead, we use the same prompts for the source image generation and attribute transformation. Another advantage of this approach is that all source images are consistent in terms of resolution, size, and generation.

The attributes we consider (section 3.2.1) fall into ten main groups: accessories, age, skin complexion, nose shape, eye shape, sex, facial expression, facial hair, hair color, and hairstyle. For each attribute, we manually tune SEGA's hyperparameters, such as image guidance, text guidance, edit threshold, and momentum, to achieve the best transformation quality on average.[5] In total, we applied 28 attributes (listed in table 3.2) on all 480 Source Faces, resulting in a total of 13,440 Transformed Faces.

[5]Ideally, hyperparameters should be tuned for each source image and attribute pair. However, for scalability, we select the best hyperparameters on average.

Figure 3.2: A subset of the counterfactual examples obtained through the generation pipeline. Source Faces in the first column. Subsequent columns display transformed faces corresponding to the column headers representing various attributes.

Dataset Validation

Next, we validate that the edits performed by SEGA have resulted in an acceptable image quality with no distortions. This step is important to decouple potential failures in the generative pipeline from the assessment of the face recognition model. The supplementary material shows an example of faces distorted due to transformations. To perform scalable validation, we leverage a vision-language model for the detection of corrupted or distorted images. We manually verify the correctness of this vision-language model through a small-scale study.

Vision-Language Model. We use LLaVA-1.5[6] (Large Language-and-Vision Assistant), which is a multimodal model that combines a vision encoder and a text decoder [121]. LLaVA is capable of understanding both language and vision, making it a powerful visual question-answering (VQA) tool. We ask LLaVA to detect whether the Transformed Faces are corrupted using this question: "*Is this a distorted face? Give a Yes or No answer.*"

Manual Validation. Next, we validate LLaVA's performance via a user survey on a random subset of the Transformed Faces. We selected 10% of the images, resulting in 1195 images. Four annotators (who are not the authors) annotated each transformed image as distorted vs. not distorted by visual inspection. We compared LLaVA's answers to the manual annotation to validate its performance. The results, depicted in the supplementary material, indicate that LLaVA detects all the distorted images with a false negative rate of 1.4%. LLaVA has a slightly higher false positive rate of 22.8%, where it labels some undistorted images as distorted. This results in removing some valid images from our analysis. However, due to its low false negative rate, LLAVA successfully filters out the distorted images. Moreover, the survey confirms the majority of images (94.4%) are undistorted, providing evidence of SEGA's capability to execute high-quality semantic edits.

[6] `https://huggingface.co/liuhaotian/llava-v1.5-7b`

		LLaVA Response	
		Class 0	**Class 1**
Survey Response	Class 0	856 (71.63%)	272 (22.76%)
	Class 1	17 (1.42%)	50 (4.18%)

Table 3.1: Confusion Matrix of LLaVA's image distortion prediction against survey results. Class 0: not distorted, Class 1: distorted.

3.2.3 Assessment of Face Recognition Systems

Regardless of the choice of face recognition system, our framework can perform a targeted assessment of the system's performance. We examine two commercially available and widely used systems for face recognition: Face++[7] and Amazon Rekognition.[8] Both online APIs provide face similarity scores.

On each face recognition system, we measure face similarity associated with pairs of faces. We measure the similarity between each source face and each of the transformed images derived from that source image's identity. We refrain from setting a threshold to classify identities and opt to use similarity scores, which better represent the recognition model's understanding of the image identity.

3.3 Results

Our objective is to evaluate the efficacy of semantic transformations in generating counterfactual examples for face recognition systems. In particular, we use the images generated from our framework to characterize the performance of two face recognition systems thoroughly: Face++ and Amazon Rekognition. In particular, we analyze the performance of these systems through the lens of the eight sex-ethnicity combinations discussed earlier. In-

[7] `https://www.faceplusplus.com/`

[8] `https://aws.amazon.com/getting-started/hands-on/detect-analyze-compare-faces-rekognition/`

formally, we want to investigate how robust these systems are to semantic shifts for each group. We formalize our evaluation through the following six null hypotheses:

Null Hypothesis 3.1. *The per-sex distributions of image similarity on ⟨System⟩ are identical.*

Null Hypothesis 3.2. *The per-ethnicity distributions of image similarity on ⟨System⟩ are identical.*

Null Hypothesis 3.3. *The per-demographic distributions of image similarity on ⟨System⟩ are identical.*

Null Hypothesis 3.4. *The per-sex distributions of image similarity degradation on ⟨System⟩ are identical.*

Null Hypothesis 3.5. *The per-ethnicity distributions of image similarity degradation on ⟨System⟩ are identical.*

Null Hypothesis 3.6. *The per-demographic distributions of image similarity degradation on ⟨System⟩ are identical.*

Null hypotheses 3.1 to 3.3 capture whether or not performance disparities associated with face recognition appear among semantically manipulated faces. Null hypotheses 3.4 to 3.6 capture whether or not performance disparities can be associated with the semantic manipulation process itself. Each null hypothesis is evaluated with a one-way ANOVA. The statistical significance of null hypotheses 3.1 to 3.6, using a significance threshold of 0.01 (after correcting for multiple hypothesis testing), for AWS are indicated in table 3.2. An analogous table for Face++ appears in table 3.3. The rejection of null hypotheses 3.1 to 3.3 each requires a statistical test on raw face similarity scores, and the rejection of null hypotheses 3.4 to 3.6 each requires a statistical test on the difference between two sets of similarity scores: 1) similarity scores over pairs of source faces and 2) similarity scores over pairs consisting of one source and one transformed face.

Attribute Group	Attribute	AM	AF	BM	BF	IM	IF	WM	WF	NH 3.1	NH 3.2	NH 3.3	NH 3.4	NH 3.5	NH 3.6
SOURCE	N/A	99.9299	99.9286	99.9951	99.9829	99.9707	99.9902	99.9943	99.9722	✗	✓	✓	N/A	N/A	N/A
no operation	no operation	99.95	99.91	100.0	99.98	99.97	99.99	99.99	99.97	✗	✓	✓	✗	✓	✓
accessories	facemask	89.29	85.84	99.98	93.07	99.95	95.34	99.97	89.85	✗	✓	✓	✓	✓	✓
	glasses	99.26	97.82	99.95	99.53	99.3	99.68	99.85	99.61	✗	✓	✓	✓	✗	✓
	head_band	99.48	98.5	99.94	99.53	99.37	98.29	99.81	99.5	✓	✓	✓	✓	✓	✓
	scarf	99.9	99.7	100.0	99.97	99.94	99.94	99.99	99.94	✓	✓	✓	✓	✓	✓
	sunglasses	91.41	70.45	99.64	99.21	92.92	96.39	99.11	94.69	✓	✓	✓	✓	✓	✓
age	old	98.28	92.34	99.99	99.93	99.25	99.82	99.89	99.79	✓	✓	✓	✓	✓	✓
	young	98.1	97.99	99.12	99.59	94.19	98.9	94.68	98.64	✓	✓	✓	✓	✓	✓
complexion	dark_colored_skin_tone	99.05	96.92	99.94	99.85	99.67	99.77	99.87	99.04	✓	✓	✓	✓	✓	✓
	heavy_makeup	99.77	99.86	99.95	99.97	97.26	99.98	99.83	99.95	✗	✓	✓	✗	✓	✓
	light_colored_skin_tone	98.96	99.17	99.98	99.89	98.25	99.73	99.35	99.65	✓	✓	✓	✓	✓	✓
	natural_makeup	98.77	99.36	99.5	99.96	95.93	99.96	99.46	99.89	✓	✓	✓	✓	✓	✓
	red_lipstick	99.9	99.78	99.99	99.98	99.95	99.97	99.99	99.94	✓	✓	✓	✓	✓	✓
expression	smile	99.14	98.79	99.92	99.82	98.94	99.46	99.87	99.5	✓	✓	✓	✓	✓	✓
facial hair	goatee	97.24	75.5	99.96	89.33	99.18	76.08	99.76	85.07	✓	✓	✓	✓	✓	✓
	mustache	97.81	93.8	99.91	99.83	99.69	99.26	99.22	99.68	✓	✓	✓	✓	✓	✓
	patchy_beard	99.94	98.94	100.0	98.43	99.96	99.47	99.99	99.76	✓	✓	✓	✓	✓	✓
	thick_beard	64.67	30.71	99.29	30.82	93.69	43.23	96.53	31.29	✓	✓	✓	✓	✓	✓
	thick_eyebrow	98.98	98.32	99.99	99.79	99.64	99.84	99.94	99.78	✗	✓	✓	✗	✓	✓
hair color	blue_hair	99.88	97.52	99.99	99.76	99.92	99.6	99.98	99.47	✓	✓	✓	✓	✓	✓
	red_hair	98.63	96.18	99.79	99.4	98.27	98.72	98.05	99.21	✗	✓	✓	✗	✓	✓
hairstyle	afro	99.6	94.44	99.97	99.49	99.71	96.24	99.97	99.04	✓	✓	✓	✓	✓	✓
	buzz_cut	96.21	78.84	99.97	99.19	99.58	99.59	99.79	99.33	✗	✓	✓	✗	✓	✓
	curly_hair	99.12	99.29	99.54	99.89	99.55	99.88	99.91	99.83	✓	✗	✓	✓	✗	✓
	pigtails	98.4	99.03	99.94	99.84	96.03	99.31	99.72	99.79	✗	✓	✓	✗	✓	✓
	shoulder_length_hair	79.34	94.57	64.92	99.64	79.25	99.06	86.41	98.82	✓	✗	✓	✓	✗	✓
sex	female	85.41	99.72	95.19	99.94	54.52	99.93	83.39	99.89	✗	✓	✓	✗	✓	✓
	male	98.42	82.33	99.97	45.43	99.58	73.86	99.73	79.18	✓	✗	✓	✓	✗	✓

Table 3.2: AWS Rekognition Similarity Scores. Range is from 0 to 100. The first letter of each demographic abbreviation stands for one of (East) Asian, Black, Indian or White. The second letter stands for sex: one of Male or Female. A ✓ indicates statistical significance, and an ✗ indicates no statistical significance.

We refer to the eight demographic combinations as *AM*–(East) Asian Male, *AF*–(East) Asian Female, *BM*–Black Male, *BF*–Black Female, *IM*–Indian Male, *IF*–Indian Female, *WM*–While Male, and *WF*–White Female.

3.3.1 Notable Performance Changes

Because all semantic changes alter images, similarity scores are affected regardless of semantic shift, yet some demographics incur greater significant performance degradation:

For example, the addition of a goatee or a thick beard significantly affects the similarity scores of female faces more than males. On both AWS Rekognition and Face++, similarity scores for AM, BM, IM, and WM exceed the similarity scores for AF, BF, IF, and WF, respectively. Likewise, the addition of shoulder-length hair more drastically affects similarity scores for males than females. On both AWS Rekognition and Face++, similarity scores for AF, BF, IF, and WF exceed the similarity scores for AM, BM, IM, and WM, respectively. In these cases, the induced semantic shift produces out-of-distribution examples: females typically do not grow facial hair, and males typically have shorter hair.

Attribute Group	Attribute	AM	AF	BM	BF	IM	IF	WM	WF	NH 3.1	NH 3.2	NH 3.3	NH 3.4	NH 3.5	NH 3.6
SOURCE	N/A	91.31	91.16	93.9	94.88	93.19	94.39	94.44	94.62	✓	✓	✓	N/A	N/A	N/A
no operation	no operation	92.36	91.87	94.69	95.17	93.89	94.74	94.82	94.95	✓	✓	✓	✓	✓	✓
accessories	facemask	64.32	64.73	93.1	72.57	92.34	79.78	93.73	78.71	✓	✓	✓	✓	✓	✓
	glasses	86.4	84.54	92.25	92.16	89.5	90.46	90.86	91.15	✗	✓	✓	✓	✓	✓
	head_band	88.23	85.81	93.16	93.17	91.89	91.9	92.85	92.22	✗	✓	✓	✗	✓	✓
	scarf	90.99	89.97	94.27	94.62	92.94	93.64	94.08	94.13	✗	✓	✓	✗	✓	✓
	sunglasses	78.83	74.92	91.38	89.72	82.83	83.44	87.42	85.75	✗	✓	✓	✗	✓	✓
age	old	83.74	81.29	93.65	93.72	89.43	92.49	92.82	92.85	✓	✓	✓	✓	✓	✓
	young	83.98	83.25	89.02	92.79	86.39	91.77	87.92	91.42	✗	✓	✓	✗	✓	✓
complexion	dark_colored_skin_tone	88.9	85.94	93.32	93.96	91.19	92.8	92.29	90.65	✓	✓	✓	✓	✓	✓
	heavy_makeup	88.56	91.04	92.85	94.96	90.81	94.38	93.04	94.62	✗	✓	✓	✓	✓	✓
	light_colored_skin_tone	88.57	88.07	93.39	93.71	89.45	92.26	90.59	93.03	✗	✓	✓	✓	✓	✓
	natural_makeup	88.32	87.78	93.49	94.59	90.84	93.95	93.8	94.21	✓	✓	✓	✓	✓	✓
	red_lipstick	91.25	90.14	94.42	94.56	93.2	93.91	93.91	94.09	✓	✓	✓	✓	✓	✓
expression	smile	85.24	85.02	91.84	93.23	89.63	90.76	92.01	91.83	✓	✓	✓	✓	✓	✓
facial hair	goatee	83.57	73.66	92.86	88.65	90.41	85.83	91.27	88.74	✗	✓	✓	✗	✓	✓
	mustache	84.01	74.6	92.54	91.12	89.56	86.64	89.44	90.62	✓	✓	✓	✓	✓	✓
	patchy_beard	92.0	89.05	94.36	93.78	93.53	93.09	94.6	94.0	✓	✓	✓	✓	✓	✓
	thick_beard	68.63	44.36	90.81	64.93	87.51	76.51	88.12	65.0	✗	✓	✓	✓	✓	✓
	thick_eyebrow	87.47	86.74	93.83	93.44	92.18	92.76	93.08	93.05	✗	✓	✓	✗	✓	✓
hair color	blue_hair	90.78	84.96	94.1	93.13	93.18	91.79	94.04	92.39	✗	✓	✓	✗	✓	✓
	red_hair	86.93	85.29	92.04	92.66	89.96	90.81	91.68	91.78	✗	✓	✓	✓	✓	✓
hairstyle	afro	89.21	82.4	93.29	92.93	91.99	89.8	93.7	90.81	✓	✓	✓	✓	✓	✓
	buzz_cut	84.92	77.53	93.76	92.37	91.46	92.86	92.21	91.61	✓	✓	✓	✓	✓	✓
	curly_hair	87.48	87.0	92.86	93.67	91.63	92.92	93.13	93.02	✓	✓	✓	✓	✓	✓
	pigtails	87.61	87.15	93.1	93.58	89.79	91.94	93.03	92.54	✓	✓	✓	✓	✓	✓
	shoulder_length_hair	75.01	83.55	74.37	92.82	83.09	91.41	84.62	91.3	✓	✓	✓	✓	✓	✓
sex	female	81.5	89.55	89.8	94.49	78.49	93.87	84.26	94.07	✓	✓	✓	✓	✓	✓
	male	87.56	79.03	93.23	71.72	92.02	86.2	92.97	85.71	✓	✓	✓	✓	✓	✓

Table 3.3: Face++ Similarity Scores. Range is from 0 to 100.

However, counterfactual examples can generate in-distribution examples that expose the failure modes of face recognition systems. For both AWS Rekognition and Face++, AM similarity scores drop significantly for the thick beard attribute, and it is conceivable for an Asian Male to grow a thick beard. However, both systems fail to recognize these cases correctly, probably due to insufficient representation in their training data.

3.3.2 Face Recognition Performance Disparity

Our experiments show that face recognition systems incur statistically significant performance disparities on generated faces. In tables 3.2 and 3.3, null hypotheses 3.1 to 3.3 are consistently rejected. We observe this to be the case more for Face++ than AWS Rekognition. In Face++, the Asian demographic consistently suffers from a worse performance. Interestingly, unlike the performance disparities previously observed by Buolamwini and Gebru [14], the Black ethnic group appears to have strong performance across different semantic attributes. This observation suggests developers of face recognition systems have addressed fairness concerns since the publication of Buolamwini and Gebru's work.

3.3.3 Utility of Counterfactual Examples

Counterfactual examples allow us to compare the performance of AWS Rekognition and Face++ across multiple targeted semantic changes. Because each element of the face pairs used to evaluate models is derived from the same identity, we generally expect a strong face recognition System to yield high similarity scores across all semantic changes. In comparing table 3.2 with table 3.3, we see that AWS is far superior to Face++. Despite this, high-level trends that appear in Face++ performance still manifest in AWS Recognition, albeit to a lesser degree.

Counterfactual examples also allow us to delve deeper into performance differences between face recognition systems. At first glance, Face++ appears to have acceptable (albeit disparate) performance for all demographic groups, as evident from the SOURCE row in table 3.3. However, when applying the semantic attributes to the images, Face++ exhibits more distinguishable performance disparities. This observation suggests a difference in the training of both systems. While both systems might have corrected for bias for standard face images, Amazon has put more effort into making their system robust in real-world deployment. Such insight is only possible with counterfactual examples.

3.4 Analytical Model

We observed that verification accuracy degrades upon manipulation of face images, particularly when a generated image is anomalous on a population level. We attribute this to machine learning models being trained on finite data. Typically these datasets are drawn from the internet. Generative models, such as diffusion models, thusly learn to generate images patterned on their internet-based dataset. Because the internet is well-understood to be a biased sample of the universe, a diffusion network trained on an internet-sourced dataset is itself a biased sample generator. To understand how a biased, finite training set

can yield biased sample generation, we utilize a Gaussian Mixture Model (GMM). A GMM is theoretically tractable proxy through which we gain intuition about generative models.

In that model, an image in demographic group $\mathtt{g}$ is drawn from $\mathcal{N}(\boldsymbol{\mu}_{\mathtt{g}}, \boldsymbol{\Sigma}_{\mathtt{g}})$. For brevity, we denote the distribution of examples in group $\mathtt{g}$ by $\mathcal{D}_{\mathtt{g}}$. Without loss of generality, our analysis considers two groups: a and b. Group a occurs with probability α where $\alpha \in (0, 1)$. Thus, group b occurs with probability $1 - \alpha$. The universe's distribution can be written as $\mathcal{D} = \alpha\mathcal{D}_a + (1 - \alpha)\mathcal{D}_b$. As previously identified, generative models are typically trained on a biased dataset. To model this bias, we assume training dataset S is drawn i.i.d. from biased data distribution $\mathcal{D}_S$. Samples in S are assumed to be d-dimensional. The biased data distribution $\mathcal{D}_S$ is a possibly re-weighted mixture of Gaussians $\mathcal{D}_a$ and $\mathcal{D}_b$. That is to say, $\mathcal{D}_S = \beta\mathcal{D}_a + (1 - \beta)\mathcal{D}_b$ where $\beta \in (0, 1)$. Distributions $\mathcal{D}_S$ and $\mathcal{D}$ are only equivalent if $\alpha = \beta$. For notational brevity, $S_{\mathtt{g}}$ denotes examples in S drawn from $\mathcal{D}_{\mathtt{g}}$.

The estimator of $\mathcal{D}$ learned from S is denoted $\hat{\mathcal{D}}_S$. The quality of estimator $\hat{\mathcal{D}}_S$ is measured with total variation distance, a notion of distributional discrepancy. Our use of total variation distance as a notion of distribution estimator quality is motivated by its use in generative model literature [122, 123].

Definition 3.7 (Discrepancy)**.** *Consider a measure space $(\Omega, \mathcal{F})$. If $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are continuous probability distributions, then the discrepancy is computed as*

$$\rho_{\mathrm{TV}}(\mathcal{Q}_1, \mathcal{Q}_2) = \sup_{z \in \mathcal{F}} |\mathcal{Q}_1(z) - \mathcal{Q}_2(z)| \tag{3.1}$$

Utilizing definition 3.7, we can show that two distinct factors contribute to discrepancy. The first factor is the finite size of S. The second factor is S being a non-representative sample of $\mathcal{D}$. Both factors are formalized in proposition 3.8, its proof is in appendix B.1. We assume the process yielding $\hat{\mathcal{D}}_S$ is an empirical Bayes estimator, such as the Expectation-Maximization algorithm. This process learns five parameters about the distribution: the group proportion β, the means and covariances of groups a and b: $\boldsymbol{\mu}_{\mathtt{a}}$, $\boldsymbol{\Sigma}_{\mathtt{a}}$, $\boldsymbol{\mu}_{\mathtt{b}}$, and $\boldsymbol{\Sigma}_{\mathtt{b}}$.

Proposition 3.8 (Proposition)**.** *Let $\delta \in (0, 1)$. If*

$$|S| = O\left(d^2 \log(2) \log(2/\delta) \left[H^2(\mathcal{D}_a, \mathcal{D}_b)\right]^{-4}\right) \tag{3.2}$$

then we have

$$\mathbb{P}\left[\rho_{\mathrm{TV}}(\hat{\mathcal{D}}_S, \mathcal{D}) > \frac{|\alpha - \beta|}{2} H^2(\mathcal{D}_a, \mathcal{D}_b)\right] \geq 1 - \delta \tag{3.3}$$

where H^2 is the squared Hellinger distance, and

$$\begin{aligned} H^2(\mathcal{D}_a, \mathcal{D}_b) &= \left(1 - \frac{|\boldsymbol{\Sigma}_a|^{1/4} |\boldsymbol{\Sigma}_b|^{1/4}}{\left|\frac{\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b}{2}\right|^{1/2}}\right) \\ &\times \exp\left\{ -\frac{1}{8} (\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^{\mathsf{T}} \left(\frac{\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b}{2}\right)^{-1} (\boldsymbol{\mu}_a - \boldsymbol{\mu}_b) \right\} \end{aligned} \tag{3.4}$$

if $\mathcal{D}_a, \mathcal{D}_b$ are each multivariate Gaussians.

This proposition suggests that even with a large number of samples in S, it can be impossible to learn $\mathcal{D}$ exactly. This is due to non-representative proportions being drawn from each demographic group, i.e. when $|\alpha - \beta| > 0$. On the other hand, when $\alpha = \beta$, and the number of samples in S is infinite, $\hat{\mathcal{D}}_S$ and $\mathcal{D}$ are equal, so $\rho_{\mathrm{TV}}(\hat{\mathcal{D}}_S, \mathcal{D})$ tends to 0. The bound also has a dependence on dimension squared: large dimension inputs require many more samples to train high-fidelity generative models. Thus, we conjecture faces generated by diffusion models trained upon larger datasets should improve upon the most degraded AWS Rekognition similarity scores in table 3.2.

3.5 Discussion

Our work puts forth a framework to audit the performance of face recognition systems via counterfactual examples. Several limitations within current diffusion models limit the effectiveness of our technique. The most consequential limitation is that prompt and hyper-parameter selection is non-trivial. We found their selection to be an incredibly time-intensive

manual process. This is further complicated by the need for our parameter choices to yield sufficient perceptual quality across an entire batch of images.

Despite the identified limitations, we foresee many directions of future work. First, a feature allowing for fine-grained control of semantic shift intensity would be of significant interest to the community, allowing an almost continuous exploration of the decision space. Second, our work would greatly benefit from computational techniques to verify that resulting images match the semantic transformation intent and the underlying source images, and do not produce other confounding factors. Finally, further generalizing our technique to other vision tasks, and perhaps even other modalities, would enable the counterfactual auditing of many models and multimodal models.

3.6 Conclusion

In this work, we put forth a technique to perform a counterfactual audit of face recognition systems. The technique depends on text-to-image models, so any counterfactual that is text-describable may be evaluated. Our technique allows us to demonstrate that face recognition systems perform poorly on out-of-distribution examples. Despite advancements in machine learning fairness, we show that face recognition systems still have statistically significant performance disparity across demographic groups.

Chapter 4

An Exploration of Sample Complexities for Fair Machine Learning

Recent advances in Generative Artificial Intelligence (AI) have turbocharged the performance of Machine Learning (ML) systems, leading to broad-scope adoption including healthcare, education, financial, and legal settings. This intertwining of society and machine learning has surfaced the issue of fairness of such ML systems. ML fairness studies whether a given predictor achieves performance parity with respect to a sensitive attribute at the individual or group level [124]. In this chapter, we characterize the role of synthetic data on fairness metric generalization behavior.

Synthetic data has tremendous utility: it is often cheaper to acquire synthetic data than natural data. Furthermore, synthetic data can realistically emulate examples that would be difficult or otherwise impossible to collect naturally. This is especially true for image data depicting organic subjects. Consider the following pathological example task: Training a gesture recognition system that works well on both amputees and non-amputees. To best

guarantee strong performance on both groups, a model should be trained on large samples from each. Fortunately for society – and unfortunately for this training task – non-amputees are far more common than amputees. Hence, a practitioner may need to curate samples for the amputee group. Setting aside potential Institutional Review Board (IRB) concerns, collecting natural data necessary to satisfactorily train such a gesture recognition system could be an expensive, ethically-fraught task[1]. For this task, and many others which are far less pathological, synthetic data generation is an effective, humane replacement for natural data curation.

Even though synthetic data is often cheaper than natural data, it is not free. Synthetic data generation pipelines have computational cost. Hence for any machine learning task, understanding how much data is needed to complete that task is still tremendously valuable. We turn our attention to sample complexity bounds.

Among the most common characterizations of machine learning efficacy are sample complexities for learning. Informally speaking, sample complexities are functions that return the number of examples necessary to achieve a specified performance guarantee associated with Empirical Risk Minimization (ERM). In the typical machine learning setting, sample complexity offers a probabilistic certificate on the difference between a predictor's true and empirical risk. The fundamental assumption underlying most sample complexity bounds is both train and test data are drawn from the same distribution. When synthetic data is used to train models, the assumption is violated, so new bounds must be derived. Conveniently, *domain adaptation* literature characterizes the behavior of a classifiers learned from one distribution, and tested on another distribution. In this chapter, we apply insights from domain adaptation to study the role of synthetic data in fairness metric generalization sample complexity bounds. We refer to such bounds as *categorywise risk sample complexity bounds*. Categorywise risk sample complexity bounds may be used to quantify how many samples

[1]We do not endorse this natural data collection strategy. It is a pathological example for demonstrative purposes.

are needed to achieve an equitable performance of ML predictors over specified population categories.

Previous works have sought to determine sample complexities for fairness metrics [125–128]. Such an approach is problematic for two reasons. Firstly, existing bounds do not account for synthetic data in model training: training and test data are assumed to be drawn from the same distribution. Hence, existing bounds do not appropriately capture generalization behavior associated with leading ML systems. Secondly, the derivation of such sample complexity bounds for a choice of fairness metric often requires intensive mathematical analysis. Such analysis must take place for each notion of predictor class complexity, such as VC-dimension [129] or Rademacher complexity [130]. In some cases, this same analysis must be repeated for each predictor class [127]. To understand the generalization behavior of a single fairness metric, sample complexity bounds for a linear Support Vector Machine (SVM) may require a separate derivation from that of a two-layer Rectified Linear Unit (ReLU) network.

In this chapter, we present a solution to both problems: by reparametrizing sample complexities for ERM Learning, we show how to achieve generalization sample complexity bounds for fairness metrics on ML systems under a variety of training regimes. Because ERM learning sample complexity bounds appear frequently in literature, our main theorem (theorem 4.8) grants ML practitioners a plug-and-play technique to broadly capture generalization behavior of fairness metrics of interest. Its corollary (corollary 4.9) characterizes fairness metric behavior in the presence of synthetic data.

Our results also suggest dataset composition influences fairness metric generalization. In particular, the least frequent category controls the sample complexity necessary to see generalization; balanced datasets lead to faster generalization. In particular, when synthetic data is drawn from balanced version of the natural distribution, we show faster model generalization. We also show that a model achieving faster generalization does not mean the model is

more performant on all categories. We experimentally validate our findings with leading data synthesis techniques. Our results in section 4.3 highlight the utility of near-infinite synthetic data stream: for each category, as the number of samples increases, categorywise measurements of risk converge to their expectation on the synthetic data distribution $\mathcal{D}'$. When the synthetic data distribution closely imitates the natural data distribution, categorywise measurements of risk also converge to their expectation on the natural data distribution $\mathcal{D}$. What does it mean to "closely imitate" a natural distribution? This is explored in our theory and evaluation.

Our experiments focus on tabular data. In particular, we focus on UCI Adult [131]. Our evaluation on UCI Adult serves as proxy for results on more computationally expensive data generation tasks.

To summarize, we put forth a framework for ML practitioners to construct generalization bounds on fairness metrics from sample complexities for ERM learning. As we will show, synthetic data can hasten convergence to generalization performance. This chapter makes the following contributions:

1. We capture the role of synthetic data on fairness metric generalization behavior. The effectiveness of the synthetic data is closely tied to how well it represents natural data.

2. We show that sample complexities for state-of-the-art generalization bounds on fairness metrics are essentially reparametrizations of sample complexities for ERM learning of hypothesis classes.

3. We draw insights into dependencies of uniform generalization bounds on categorywise risk. We find that structural issues in data synthesis, data itself, and the choice of classifier architecture induce a baseline of unfairness. Some unfairness can be overcome by training a model on well-crafted synthetic data.

4.1 Preliminaries

We define the notation necessary to analyze categorywise risk generalization: $\mathcal{X} \subseteq \mathbb{R}^d$ denotes the space of examples and $\mathcal{Y}$ denotes the set of possible labels. A set of examples is denoted as $S \subseteq \mathcal{X} \times \mathcal{Y}$ and contains N total items. S is $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$. Sometimes we abuse notation and use S to represent an unlabeled set of training examples, i.e. $S \triangleq \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. We may also refer to S as a dataset. S is drawn i.i.d. from a natural distribution $\mathcal{D}$, to which practitioners are blind. The classifier architecture defines a predictor class $\mathcal{H}$, from which a predictor $\mathrm{h} : \mathcal{X} \to \mathcal{Y}$ is selected. The terms hypothesis, classifier, and predictor are used interchangably.

Recent advances in machine learning have demonstrated the utility of training models on synthetic data, yet previous fairness bounds literature only considers natural data. Hence new bounds are required which account for synthetic training data. Care must be taken in derivation: Generated data, while realistic, is not drawn from the natural distribution. In this chapter, we assume synthetic data is drawn from from a distribution $\mathcal{D}'$ which imitates natural distribution $\mathcal{D}$.

Domain adaptation literature characterizes the behavior of a classifiers, learned from a source domain, then applied to a target domain. Error bounds for domain adaptation are well studied in literature [132, 133]. Typically, domain adaptation is associated with transfer learning: For example it may be appropriate to utilize a sentiment classification model trained on X platform posts to analyze Amazon customer feedback on computer parts. This may be because source data is be more readily available than target data, or it may be more efficient to utilize an already trained model. Despite the association with transfer learning, domain adaption bounds still apply when a model is trained on synthetic data $\mathcal{D}'$, and applied to natural data $\mathcal{D}$: The synthetic data distribution $\mathcal{D}'$ is the source distribution, and the natural data distribution $\mathcal{D}$ represents the target data distribution.

We denote a loss function by $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Common examples of loss functions

include hinge loss and cross-entropy loss. Generally speaking, loss functions are used as an optimizable proxies for raw error rate, sometimes referred to as the 0-1 loss:

$$\ell_{0\text{-}1}(\mathrm{h}(\boldsymbol{x}), y) \triangleq \mathbb{1}\left[\mathrm{h}(\boldsymbol{x}) \neq y\right] \tag{4.1}$$

where $\mathbb{1}$ is the indicator function. Paired with the 0-1 loss is accuracy: accuracy is $1 - \ell_{0\text{-}1}$. While we focus on categorywise generalization properties associated with the 0-1 loss, existing domain adaptation literature shows our insights extend to additional loss functions [134].

The *risk* refers to the expected loss of a predictor h. More specifically, there two related varieties of risk to which we refer: the *empirical risk* and the *true risk*. Let h be a predictor. The empirical risk refers to the expected loss of h on a finite sample S drawn. In the typical machine learning setting, S is assumed to be drawn from the natural distribution $\mathcal{D}$; however, in this work, a dataset S may be drawn from the synthetic distribution $\mathcal{D}'$. The true risk refers to the expected loss of h on distribution of interest $\mathcal{D}$ itself:

$$\mathcal{L}_{S,\ell}(\mathrm{h}) \triangleq \frac{1}{|S|} \sum_{(\boldsymbol{x},y)\in S} \ell\left(\mathrm{h}(\boldsymbol{x}), y\right) \tag{4.2}$$

$$\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h}) \triangleq \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{D}} \ell\left(\mathrm{h}(\boldsymbol{x}), y\right) \tag{4.3}$$

When referring to the risk associated with a general loss function, we sometimes omit the subscript ℓ.

If S is drawn i.i.d. from $\mathcal{D}$, $\mathcal{L}_{S,\ell}(\mathrm{h})$ is a consistent estimator of $\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h})$. If S is instead drawn from $\mathcal{D}'$, we must leverage insights from domain adaptation [132, 133] to understand inconsistencies in estimating $\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h})$.

The overall population S breaks into potentially overlapping population groups denoted as g_i. The set of population groups is denoted G and can be decomposed as $\{\mathsf{g}_1, \ldots, \mathsf{g}_{|\mathsf{G}|}\}$.

When referring to a general population group, we may use g without subscript.

A *sensitive* feature captures an individual's membership in one or more protected population groups g [45, 135]. Membership in a protected group often of interest to society or an institution. Examples of protected groups include those delineated by sex, income, and religion. We assume our learner is blind to protected attributes. That is to say, given an example $\boldsymbol{x}$, the learner (i.e. trained classifier) is *blind* to which population group(s) g the example $\boldsymbol{x}$ belongs, but the practitioner is *aware* of the population group to which the example belongs.

With the population group g defined, we now formally define the category:

Definition 4.1 (Category)**.** *A category is a pair* $(\mathsf{g}, \hat{y})$ *of population group* $\mathsf{g} \in \mathsf{G}$ *and a predicted label* $\hat{y} \in \mathcal{Y}$*.*

Remark 4.2. *Note that one may wish to define a category as the pair,* $(\mathsf{g}, \hat{Y})$*, consisting of population group* $\mathsf{g} \in \mathsf{G}$ *and set of predicted labels* $\hat{Y} \subseteq \mathcal{Y}$*. Such a category definition is useful when analyzing ML models for regression problems.*

The frequency with which a predictor h renders prediction $\hat{y}$ on an example in group g_i is denoted by $\gamma_{\mathrm{h},\mathsf{g},\hat{y}}$. More precisely, $\gamma_{\mathrm{h},\mathsf{g},\hat{y}} = \mathbb{P}_{\boldsymbol{x}\sim\mathcal{D}}\left[\mathrm{h}(\boldsymbol{x}) = \hat{y}, \boldsymbol{x} \in \mathsf{g}\right]$. For notational convenience, we denote $\gamma = \min \gamma_{\mathrm{h},\mathsf{g},\hat{y}}$. That is, the lowest frequency amongst categories $(\mathsf{g}, \hat{y}) \subseteq \mathsf{G} \times \mathcal{Y}$ is denoted γ. Sometimes we utilize $\gamma_{(\mathsf{g},\hat{y})}$ which refers to the frequency of category $(\mathsf{g}, \hat{y})$.

Though ERM optimizes the training loss, a practitioner may be interested quantifying or otherwise having guarantees on alternative notions of classifier performance. We define the categorywise risk, which captures many well known fairness metrics.

Definition 4.3 (Categorywise Risk)**.** *Let* $(\mathsf{g}, \hat{y})$ *be a category, let* $\mathrm{h} \in \mathcal{H}$ *be a predictor, and let* ℓ *be a loss function. Further, let* $\mathcal{D}$ *be an unknown data distribution and let* S *be a sample from that distribution. The* Categorywise Empirical Risk $\mathcal{L}_{S,\ell}(\mathrm{h}, \mathsf{g}, \hat{y})$ *refers to the*

expected loss of h *on examples in* S *belonging to category* $(\mathsf{g}, \hat{y})$. *The* Categorywise True Risk $\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h}, \mathsf{g}, \hat{y})$ *refers to the expected loss of* h *over examples drawn from* $\mathcal{D}$ *conditioned on their membership in category* $(\mathsf{g}, \hat{y})$*:*

$$\mathcal{L}_{S,\ell}(\mathrm{h}, \mathsf{g}, \hat{y}) \triangleq \sum_{(\boldsymbol{x},y)\in S} \frac{\ell\left(\mathrm{h}(\boldsymbol{x}), y\right) \cdot \mathbb{1}\left[\boldsymbol{x} \in \mathsf{g}, \mathrm{h}(\boldsymbol{x}) = \hat{y}\right]}{\sum_{(\boldsymbol{x}',y')\in S} \mathbb{1}\left[\boldsymbol{x}' \in \mathsf{g}, \mathrm{h}(\boldsymbol{x}') = \hat{y}\right]} \tag{4.4}$$

$$\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h}, \mathsf{g}, \hat{y}) \triangleq \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{D}} \Big[\ell\left(\mathrm{h}(\boldsymbol{x}), y\right) \;\Big|\; \boldsymbol{x} \in \mathsf{g}, \mathrm{h}(\boldsymbol{x}) = \hat{y}\Big] \tag{4.5}$$

The categorywise risk encompasses many fairness metrics and constraints, including demographic parity [136], Equal Opportunity [47], and Multicalibration [126]. For instance, a predictor h achieves true demographic parity if for all $\mathsf{g}, \mathsf{g}' \in \mathsf{G}$ and all $\hat{y}, \hat{y}' \in \mathcal{Y}$ we have $\mathcal{L}_{\mathcal{D},\ell_{0\text{-}1}}(\mathrm{h}, \mathsf{g}, \hat{y})$ is equivalent to $\mathcal{L}_{\mathcal{D},\ell_{0\text{-}1}}(\mathrm{h}, \mathsf{g}', \hat{y}')$. Assuming $\mathsf{G} = \{\mathsf{g}, \mathsf{g}'\}, \mathcal{Y} = \{y, y'\}$ and y is the "advantaged" outcome, a binary predictor h would satisfy True Equal Opportunity if $\mathcal{L}_{\mathcal{D}_y,\ell_{0\text{-}1}}(\mathrm{h}, \mathsf{g}, y)$ is equivalent to $\mathcal{L}_{\mathcal{D}_y,\ell_{0\text{-}1}}(\mathrm{h}, \mathsf{g}', y)$ where $\mathcal{D}_y$ denotes the distribution of examples in $\mathcal{D}$ conditioned on the examples holding a ground-truth "advantaged" outcome or label.

The focus of this chapter is characterizing how empirical categorywise risk $\mathcal{L}_{S,\ell}$ relates to categorywise risk $\mathcal{L}_{\mathcal{D},\ell}$. In particular, we focus on sample-complexity bounds. When training data S is drawn from the natural distribution, sample-complexity bounds on a hypothesis $\mathrm{h} \in \mathcal{H}$ provide the number of i.i.d. samples drawn from a distribution $\mathcal{D}$ necessary to estimate, with a sufficiently small error ϵ and suitably high probability $1 - \delta$, the true risk $\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h})$ of predictor h over that distribution $\mathcal{D}$. It is defined as follows.

Definition 4.4 (Sample Complexity Bound)**.** *Given a loss function, error tolerance* $\epsilon \in [0, \infty)$ *and confidence parameter* $\delta \in (0, 1]$*, a sample complexity* $m_{\mathcal{H},\ell}(\epsilon, \delta)$ *is the minimum number of samples such that the following holds:*

$$\mathbb{P}\left\{|\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h}) - \mathcal{L}_{S,\ell}(\mathrm{h})| \leq \epsilon\right\} \geq 1 - \delta \tag{4.6}$$

It is important to note that in some contexts, $m_{\mathcal{H},\ell}(\epsilon, \delta)$ is a *distribution specific* bound.

This means that it has a dependence on the data distribution, $\mathcal{D}$. When referring to the sample complexity associated with a general loss function, or if the loss function is otherwise clear from context, we may omit the subscript ℓ.

If S is drawn i.i.d. from $\mathcal{D}$, $\mathcal{L}_{S,\ell}(\mathrm{h})$ is a consistent estimator of $\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h})$. If S is instead drawn from $\mathcal{D}'$, we must leverage insights and definitions from domain adaptation [132, 133] to understand inconsistencies in estimating $\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h})$. We need some additional notation before describing domain adaptation bounds. $\mathcal{H}$ is minimum sum of realizable error among predictors both $\mathcal{D}$ and $\mathcal{D}'$:

$$\lambda \geq \inf_{\mathrm{h}\in\mathcal{H}} \left[\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h}) + \mathcal{L}_{\mathcal{D}',\ell}(\mathrm{h})\right] \tag{4.7}$$

Given a collection of subsets $\mathcal{A}$ of $\mathcal{X}$ such that $\mathcal{A}$ is measurable with respect to both $\mathcal{D}, \mathcal{D}'$, the $\mathcal{A}$-distance between $\mathcal{D}, \mathcal{D}'$ is defined as

$$\rho_{\mathcal{A}}(\mathcal{D}, \mathcal{D}') = 2 \sup_{A\in\mathcal{A}} |\mathop{\mathbb{P}}_{\mathcal{D}}[A] - \mathop{\mathbb{P}}_{\mathcal{D}'}[A]| \tag{4.8}$$

Lastly, we need to adapt the $\mathcal{A}$-distance to hypotheses. For a binary-valued $f(\boldsymbol{x})$, let $\mathcal{X}_f \subseteq \mathcal{X}$ be the subset with characteristic function f

$$\mathcal{X}_f = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) = 1\} \tag{4.9}$$

Then $\rho_{\mathcal{H}}(\cdot, \cdot)$ denotes the $\mathcal{A}$-distance on the class of subsets whose characteristic functions are functions in $\mathcal{H}$. In slight abuse of notation, we may use $\rho_{\mathcal{H}}$, without arguments, to denote $\rho_{\mathcal{H}}(\mathcal{D}, \mathcal{D}')$.

Definition 4.5 (Domain Adaptation Sample Complexity)**.** *Let* h *be a hypothesis in hypothesis class $\mathcal{H}$. Let training data S be drawn from synthetic distribution $\mathcal{D}'$. Further, let error tolerance $\epsilon \in [0, \infty)$, confidence parameter $\delta \in (0, 1]$, and λ, $\rho_{\mathcal{H}}$ be as defined above. Then the domain adaptation sample complexity $m_{\mathcal{H},\ell,\lambda,\rho_{\mathcal{H}}}(\epsilon, \delta)$ is the minimum number of samples*

such that the following holds for every $\mathrm{h} \in \mathcal{H}$*:*

$$\mathbb{P}\left[|\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h}) - \mathcal{L}_{S,\ell}(\mathrm{h})| \leq \epsilon + \rho_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') + \lambda\right] \geq 1 - \delta \tag{4.10}$$

Note ϵ *has several dependencies, including the number of samples in* S*, and may be distribution specific.*

The machinery of definition 4.5 is nearly identical to the machinery underlying definition 4.4. Because S is drawn from $\mathcal{D}'$ in definition 4.5, we may decompose inequality 4.10 as follows: $|\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h}) - \mathcal{L}_{S,\ell}(\mathrm{h})| \leq |\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h}) - \mathcal{L}_{\mathcal{D}',\ell}(\mathrm{h})| + |\mathcal{L}_{\mathcal{D}',\ell}(\mathrm{h}) - \mathcal{L}_{S,\ell}(\mathrm{h})|$. Ben-David et al. show $|\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h}) - \mathcal{L}_{\mathcal{D}',\ell}(\mathrm{h})| \leq \rho_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') + \lambda$ [132], and $|\mathcal{L}_{\mathcal{D}',\ell}(\mathrm{h}) - \mathcal{L}_{S,\ell}(\mathrm{h})|$ is governed by ERM bounds which appear in literature.

For our purposes, it will also be useful to study the sample complexity for risk generalization on a population category $(\mathsf{g}, \hat{y})$.

Definition 4.6 (Category Sample Complexity Bound)**.** *Let* $(\mathsf{g}, \hat{y})$ *be a category. Then given a loss function* ℓ*, error tolerance* $\epsilon \in [0, \infty)$ *and confidence parameter* $\delta \in (0, 1]$*, a category sample complexity* $m_{\mathcal{H},\mathsf{g},\hat{y}}(\epsilon, \delta)$ *is the number of samples in category* $(\mathsf{g}, \hat{y})$*, such that the following holds:*

$$\mathbb{P}\left\{|\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h}, \mathsf{g}, \hat{y}) - \mathcal{L}_{S,\ell}(\mathrm{h}, \mathsf{g}, \hat{y})| \leq \epsilon\right\} \geq 1 - \delta \tag{4.11}$$

If training data is instead drawn from a synthetic distribution, we may also use domain adaptation techniques to derive category sample complexity bounds $m_{\mathcal{H},\mathsf{g},\hat{y},\ell,\lambda,\rho_{\mathcal{H}}}(\epsilon, \delta)$.

Definition 4.7 (Category Domain Adaptation Sample Complexity Bound)**.** *Let* $(\mathsf{g}, \hat{y})$ *be a category. Then given a loss function* ℓ*, error tolerance* $\epsilon \in [0, \infty)$ *and confidence parameter* $\delta \in (0, 1]$*, a category domain adaptation sample complexity* $m_{\mathcal{H},\mathsf{g},\hat{y},\ell,\lambda,\rho_{\mathcal{H}}}(\epsilon, \delta)$ *is the number of samples in category* $(\mathsf{g}, \hat{y})$*, such that the following holds:*

$$\mathbb{P}\left\{|\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h},\mathsf{g},\hat{y}) - \mathcal{L}_{S,\ell}(\mathrm{h},\mathsf{g},\hat{y})| \leq \epsilon + \rho_{\mathcal{H},\mathsf{g},v}(\mathcal{D},\mathcal{D}') + \lambda_{\mathsf{g},v}\right\} \geq 1-\delta \tag{4.12}$$

where $\rho_{\mathcal{H},\mathsf{g},v}(\mathcal{D},\mathcal{D}')$ *denotes the* $\mathcal{A}$*-distance on the class of subsets of category* (g,v) *whose characteristic functions are functions in* $\mathcal{H}$ *and:*

$$\lambda_{\mathsf{g},v} \geq \inf_{\mathrm{h}\in\mathcal{H}} \left[\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h},\mathsf{g},v) + \mathcal{L}_{\mathcal{D}',\ell}(\mathrm{h},\mathsf{g},v)\right] \tag{4.13}$$

Synthetic data can hasten convergence to generalization for infrequent categories: a synthetic data distribution $\mathcal{D}'$ can be a re-balanced version of $\mathcal{D}$. If infrequent categories in $\mathcal{D}$ are made to be more frequent in $\mathcal{D}'$, the upper-bound on overall categorywise sample complexity bounds will become smaller.

Our goal is to determine a bound on categorywise risk sample complexity. That is, we are interested in bounds for $m_{\mathcal{H},\ell}(\epsilon,\delta,\gamma)$ and $m_{\mathcal{H},\ell,\lambda,\rho_{\mathcal{H}}}(\epsilon,\delta,\gamma)$.

4.2 Relating Categorywise Risk and ERM

With notation defined, we now focus on our contribution regarding categorywise risk generalization sample complexity bounds. The main contribution of this section is the ease with which such sample complexity bounds may be derived: *Categorywise risk generalization sample complexity bounds are a reparametrization of ERM learning sample complexity bounds.* We begin with a theorem which formalizes the relationship between ERM and categorywise risk. The theorem has implications on the relationship between risk, hypothesis class complexity, and dataset balancing.

Theorem 4.8. *Let* S *be an i.i.d. sample drawn from distribution* $\mathcal{D}$*, let* G *be the set of groups, and let* γ *be the category frequency parameter described in section 4.1. Given a loss function* ℓ *bounded between* 0 *and* 1*, the categorywise risk complexity can be bounded with the maximum category sample complexity. That is,*

$$m_{\mathcal{H}}(\epsilon, \delta, \gamma) \leq \max_{\mathsf{g} \in \mathsf{G}, \hat{y} \in \mathcal{Y}} \frac{2}{\gamma} m_{\mathcal{H}, \mathsf{g}, \hat{y}, \ell} \left(\epsilon, \frac{\delta}{2|\mathsf{G}||\mathcal{Y}|} \right). \tag{4.14}$$

Proof Ideas: The main idea of proving theorem 4.8 is to decompose the problem into $|\mathsf{G}||\mathcal{Y}|$ separate instances. Each instance corresponds to a category $(\mathsf{g}, \hat{y})$. Then, for each category, we show that risk for the category can be achieved by using standard ERM bounds for that category. In particular, the category $(\mathsf{g}, \hat{y})$ can be bounded in terms of $m_{\mathcal{H}, \mathsf{g}, \hat{y}}$. We then determine the number of samples needed so that each category will have a sufficient number of samples to draw from (providing the $\frac{2}{\gamma}$ term). A full proof of this theorem appears in appendix C.1.

This theorem implies that category sample complexity is *sufficient* for categorywise risk generalization. Further, this theorem implies that any bounds that can be instantiated for empirical risk minimization will imply bounds on categorywise risk. Theorem 4.8 also gives rise to a corollary apt to learning on synthetic data:

Corollary 4.9. *Let S be an i.i.d. sample drawn from distribution $\mathcal{D}'$, let λ and $\rho_{\mathcal{H}}$ be as defined in section 4.1, let G be the set of groups, and let γ be the category frequency parameter described in section 4.1. Given a loss function ℓ bounded between 0 and 1, then for each category, synthetic-data categorywise complexity can be bounded with the maximum synthetic data category sample complexity. That is,*

$$m_{\mathcal{H}, \ell, \lambda, \rho_{\mathcal{H}}}(\epsilon, \delta, \gamma) \leq \max_{\mathsf{g} \in \mathsf{G}, \hat{y} \in \mathcal{Y}} \frac{2}{\gamma} m_{\mathcal{H}, \mathsf{g}, \hat{y}, \ell, \lambda, \rho_{\mathcal{H}}} \left(\epsilon, \frac{\delta}{2|\mathsf{G}||\mathcal{Y}|} \right). \tag{4.15}$$

We will explore the utility of theorem 4.8 and corollary 4.9 in sections 4.2.1, 4.2.2 and 4.3.

4.2.1 Bound Implications

Demographic (im)balance within the dataset has been well discussed in fairness literature [45, 136, 137]. Interestingly, theorem 4.8 bounds have implications in dataset balancing as

well. Demographic balance manifests itself as γ. For fixed $\mathcal{H}, \epsilon, \delta, \gamma$ and G, sample complexity for categorywise risk generalization $m_{\mathcal{H}}(\epsilon, \delta, \gamma)$ is minimized when γ is maximized. This occurs when $\gamma = \frac{1}{|\mathsf{G}||\mathcal{Y}|}$, which is when the number of examples in each category is the same. A similar concept is true for domain adaptation sample complexities: $m_{\mathcal{H},\ell,\lambda,\rho_{\mathcal{H}}}(\epsilon, \delta, \gamma)$ is minimized when γ is maximized.

The predictor class complexity factors into the categorywise risk generalization sample complexity bounds via the ERM category sample complexity bound $m_{\mathcal{H},\mathsf{g},\hat{y}}$. The category sample complexity bound itself is a standard ERM sample complexity bound $m_{\mathcal{H}}(\epsilon, \delta)$ conditioned on all samples being members of the specified category $(\mathsf{g}, \hat{y})$. In general it takes fewer samples to learn a less complicated predictor class than it does to learn a more complicated predictor class. On the other hand, even if a more expressive predictor class is not fully learned, the resulting classifier can be more performant than a classifier learned on a less complicated class. This somewhat counterintuitive property will be shown in our evaluation.

4.2.2 Illustrative Bounds

The choice of ERM sample complexity term also allows machine learning practitioners to loosen or tighten categorywise risk generalization bounds depending on how much information about the underlying classification scenario. The more information a practitioner has, the sharper their bounds can be. For illustrative purposes, let us consider the following three scenarios.

Scenario 1 The practitioner is aware of the predictor class $\mathcal{H}$, but the practitioner has yet to receive any data.

Scenario 2 The practitioner is aware of the predictor class, and training data is known to be drawn from a balanced synthetic distribution, but the practitioner has yet to receive any data.

Scenario 3 The practitioner is aware of the predictor class $\mathcal{H}$ and has received the labeled natural dataset S.

Scenario 1: Vapnik-Chervonenkis (VC) Dimension. In this first setting, the practitioner has enough information to invoke the VC-dimension based ERM learning sample-complexity bounds. VC-dimension [129] is the canonical distribution agnostic tool used to bound the sample complexity for learning a hypothesis class $\mathcal{H}$. In particular, if $\mathcal{H}$ has VC-dimension p, then $m_{\mathcal{H}}(\epsilon, \delta) = O\left(\frac{p+\log\frac{1}{\delta}}{\epsilon^2}\right)$ [138]. As such, the practitioner can use theorem 4.8 to achieve a sample-complexity bound on loss bounded between 0 and 1. When we have knowledge of the VC-dimension, the sample complexity bound is:

$$m_{\mathcal{H}}(\epsilon, \delta, \gamma) = O\left(\frac{1}{\gamma\epsilon^2}\left(\mathrm{d} + \log\left(\frac{2|\mathsf{G}||\mathcal{Y}|}{\delta}\right)\right)\right) \tag{4.16}$$

Scenario 2: Balanced Synthetic Data In this case, the practitioner is able to utilize their knowledge of conditions upon the synthetic data distribution to yield faster categorywise risk generalization. While it is possible to balance natural data across protected attributes, many existing techniques rely on data subtraction to achieve balancing. Instead, we show how sharper-than-natural sample complexity bounds can be obtained by implementing data augmentation with well-crafted data. In particular, to achieve the sharpest generalization bound, categories should be rendered uniformly likely in training data, which forces $\gamma = \frac{1}{|\mathsf{G}||\mathcal{Y}|}$. This value of γ was discussed in section 4.2.1

We showcase a similar scenario in our evaluation (section 4.3).

Scenario 3: Rademacher Complexity. Because the practitioner now has access to both dataset S and the hypothesis class itself, a Rademacher-complexity based ERM learning sample-complexity bound may be invoked. Some predictor classes have finite Rademacher complexity, but infinite or otherwise undefined VC-dimension. The most notable such hy-

pothesis class is the Radial Basis Function (RBF) kernel SVM. Hence, the ability to invoke a Rademacher complexity ERM learning bound may grant the ability to construct a category-wise risk generalization bound in a scenario for which a VC-dimension is uninformative. In appendix C.3, we illustrate this more precisely with two examples: Kernel SVM predictors and ReLU activated neural networks. Consider the Kernel SVM predictor class $\mathcal{H}_K$:

$$\mathcal{H}_K \triangleq \left\{ \boldsymbol{x} \mapsto \operatorname{sgn}\left(\boldsymbol{w}^{\mathsf{T}} \Phi_K(\boldsymbol{x})\right) : |\boldsymbol{w}^{\mathsf{T}} \Phi_K(\boldsymbol{x})| \geq 1 \right\} \tag{4.17}$$

where $\Phi : \mathbb{R}^d \to \mathbb{R}^{d'}$ is the kernel mapping function and $d' \in \mathbb{N} \cup \infty$. Often, the dimension of the kernel d' is exceedingly large, so the *kernel trick* is utilized to reduce the computational complexity of both the training of and prediction with a kernel predictor class. The kernel trick takes predictor classes of the form depicted in eq. (4.17) and represents them as $\mathrm{h}(\boldsymbol{x}) = \sum_{i=1}^{N} w_i y_i K\left(\boldsymbol{x}_i, \boldsymbol{x}\right)$.

Notably, for the RBF Kernel, d' is ∞. Hence, the kernel trick is the only way to pragmatically deploy an RBF kernel. By computing the Rademacher complexity (details are in section C.3 for this hypothesis class), we have that $m_{\mathcal{H}_K}(\epsilon, \delta) = O\left(\frac{B^2 + \log \frac{1}{\delta}}{\epsilon^2}\right)$, where B denotes the maximum value of $K(\boldsymbol{x}, \boldsymbol{x})$ over all $\boldsymbol{x}$. This bound holds over all categories as the value of B is unchanged between categories. Substituting this into theorem 4.8, we find a categorywise risk complexity of:

$$m_{\mathcal{H}_K}(\epsilon, \delta, \gamma) = O\left(\frac{B^2 + \log\left(\frac{2n|\mathsf{G}||\mathcal{Y}|}{\delta}\right)}{\epsilon^2 \gamma}\right) \tag{4.18}$$

4.3 Evaluation

Our evaluation aims to show the generalization behavior of the categorywise risk on the 0-1 loss. The evaluation utilizes the UCI Adult dataset. The Adult dataset, comprises 14 attributes and 48842 records, with split between 32561 train records and 16281 test records.

Entries in the dataset are extracted from the 1994 Census database. A classification task is associated with the UCI Adult dataset: A classifier trained to predict whether or not a person's income exceeds \$50K. Table 4.1 details the sex and label distribution of UCI Adult.

The UCI Adult dataset has been the subject of numerous ML fairness publications [45, 125, 139]. Examining table 4.1, we see how dependencies on sex exist: the dataset is about 33% female, yet of the people earning over $> \$50k$, just over 15% are female. Though our experiments focus primarily on sex disparities, there are similar concerns about UCI adult with respect to race. The social concerns surrounding the UCI Adult dataset have brought forth its use in ML fairness literature.

Because the original UCI Adult dataset is finite, we capture its underlying distribution by training the Synthetic Data Vault (SDV) implementation of conditional tabular generative adversarial network (CTGAN) on the UCI Adult Training Data [140, 141]. Infinite data allows us to capture generalization properties indicated by our framework: CTGAN grants us access to an infinite data stream that is intended to imitate UCI Adult. This allows us to examine bounds presented in theorem 4.8 and corollary 4.9 currently. Our bounds indicate that more synthetic data indicates better generalization on the source distribution. Combined with theoretical techniques from domain adaptation, we can formally link accuracy on synthetic data to accuracy on natural data.

To study category-wise generalization properties, we train classifiers on increasingly training sets. All training sets are drawn from the synthetic distribution. We consider three types of training sets: data drawn unconditionally from the CTGAN, data conditioned to have equiprobable labels, data conditioned to have equiprobable sex-label pairs.

We consider 40 cardinalities of unconditional training data. For each value of z, we draw 100 unique training sets.:

$$|S| \in \left\{ \left\lfloor 10^{z/10} \right\rfloor | z \in \{10, 11, \ldots, 48, 49\} \right\} \tag{4.19}$$

For conditionally generated training data with equiprobable labels, we consider approximately 40 cardinalities. For each value of z, we again draw 100 unique training sets:

$$|S| \in \left\{ |\mathcal{Y}| \left\lfloor \frac{10^{z/10}}{|\mathcal{Y}|} \right\rfloor | z \in \{10, 11, \ldots, 48, 49\} \right\} \tag{4.20}$$

For conditionally generated training data with equiprobable sex-label pairs, we consider approximately 40 cardinalities. For each value of z, yet again we draw 100 unique training sets:

$$|S| \in \left\{ |\mathsf{G}||\mathcal{Y}| \left\lfloor \frac{10^{z/10}}{|\mathsf{G}||\mathcal{Y}|} \right\rfloor | z \in \{10, 11, \ldots, 48, 49\} \right\} \tag{4.21}$$

We assign an individual earning over \$50K the label $y = 1$, and if they do not, we assign them a label $y = 0$. Classifiers for the Adult dataset are trained on the following features: `Age`, `Workclass`, `fnlwgt`, `Education-Num`, `Martial Status`, `Occupation`, `Relationship`, `Capital Gain`, `Capital Loss`, `Hours per week`, and `Country`. The sex and race features are withheld from training and are reserved as protected attributes.

For each unique training set, we learn 3 classifier architectures: A linear SVM, a sigmoid kernel SVM, and PyTorch [142] implementation of Tabnet [143]. Each classifier is evaluated on its training data, a large sample from its source synthetic data distribution (100k examples), and the original UCI Adult test set.

Q1. Do models learned on synthetic data perform effectively on natural data?

Q2. Does category balancing within the training set translate to faster categorywise generalization?

Q3. How well does the observed generalization behavior resemble our bounds?

Figure 4.1 demonstrates our results on the UCI Adult dataset. Each column represents a model architecture, and each row represents a sampling strategy. Each subfigure contains

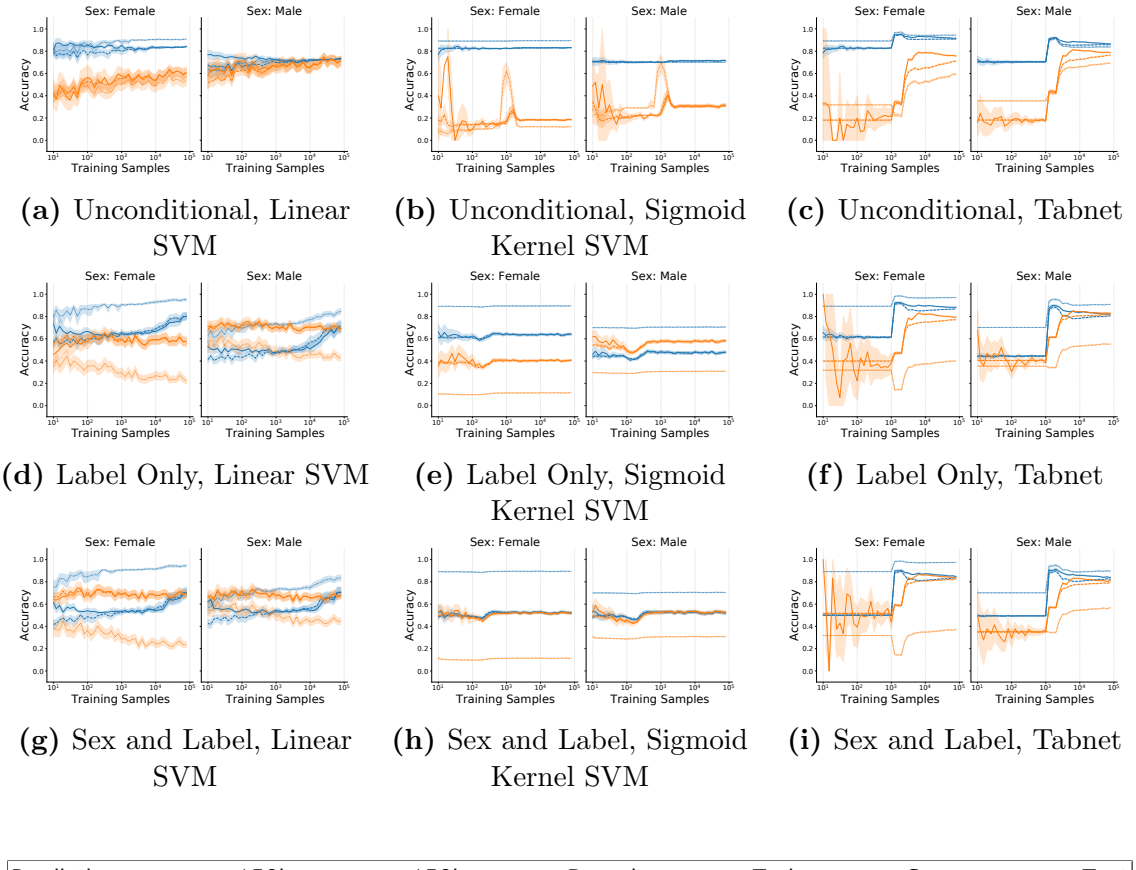
	$>$ \$50k	$\leq$ \$50k	TOTAL
Female (Train)	1179	9592	10771
Male (Train)	6662	15128	21790
TOTAL (Train)	7841	24720	32561
Female (Test)	590	4831	5421
Male (Test)	3256	7604	10860
TOTAL (Test)	3846	12435	16281

Table 4.1: UCI Adult Dataset Demographic Distribution as conditioned by label and sex. The top half of the table depicts the demographics of training data, and the lower half displays the demographics of the test data.

two plots. Classification on female examples is depicted on the left of each subplot, and classification on male examples appears on the right. The blue color represents a category in which each example is predicted to earn $\leq$\$50k. Orange represents the category in which each example is predicted to earn $>$\$50k.

We answer our research questions below by examining the performance of models by examining their generalization and convergence behavior. Observed trends help us understand how faithful theory is to experiments.

Q1. Learning on synthetic data does appear to have some utility with respect to natural data performance. The best way to understand natural data performance is to examine a model's performance on the UCI Adult's test dataset. As more synthetic data is used to train a model, performance on the test dataset appears to increase on at least one category. Synthetic data use is not free: it is not entirely representative of the underlying dataset. Generally speaking, the less representative the synthetic data distribution is of natural data, the greater the divergence term $\rho_{\mathcal{H}}$. Hence we see insurmountable performance gaps between performance on synthetic data, and on natural (test) data itself. This performance gap is especially prominent on Tabnet.



(a) Unconditional, Linear SVM **(b)** Unconditional, Sigmoid Kernel SVM **(c)** Unconditional, Tabnet

(d) Label Only, Linear SVM **(e)** Label Only, Sigmoid Kernel SVM **(f)** Label Only, Tabnet

(g) Sex and Label, Linear SVM **(h)** Sex and Label, Sigmoid Kernel SVM **(i)** Sex and Label, Tabnet

Figure Legend

Figure 4.1: Generalization behavior is depicted on the UCI Adult dataset. Each column depicts a classifier architecture. Each row represents a data sampling strategy.

Q2. Categorywise dataset balancing leads enhance convergence rate towards the model's generalization behavior on the least frequent category, a female earning less than $50k. For each classifier architecture, we examine performance on that category. In particular, we compare the performance of models trained unconditionally drawn data with that of models trained on data balanced by both sex and label. The width of the shaded region is a proxy for convergence to generalization performance. On all three architectures, the female earning less than $50k, has fastest convergence when trained on balanced data. Note that fast convergence to generalization behavior does not mean the generalization behavior is performant. Models trained on unconditionally drawn data have the strongest performance

on UCI Adult test set. This is likely because synthetic data best represents natural data when synthetic data is drawn unconditionally. Conditioning data drawn from the synthetic distribution $\mathcal{D}'$ is necessary to balance to balance training data, yet the same conditioning process generally results in greater divergence from the natural distribution as captured by the $\rho_{\mathcal{H}}$ term.

Q3. To understand how well generalization behavior resembles our bounds, we examine the relationship between training data performance and source data performance in the context of hypothesis class. Our bounds suggest that a less expressive model architectures see faster convergence to generalization behavior. Indeed that is true in our experiments. Model performance on training data converges to performance on the source dataset. The gap between source data performance and training data performance is largest for Tabnet, the most expressive hypothesis class we consider. Even though the gap between source and training performance is largest on Tabnet, the overall performance is generally stronger than Linear SVM and Sigmoid Kernel SVM.

4.4 Related Work

The following discusses works related to this chapter, including synthetic data, ERM sample complexity, and other fairness notions.

Synthetic Data

Synthetic data is often used to evaluate and train machine learning models when natural data is difficult, expensive, or otherwise impossible to obtain. Two model architecture families are often associated with synthetic data: Generative Adversarial Networks (GANs) [144] and diffusion models [103]. In addition to ML models, 3D rendering techniques and classical statistical techniques are also utilized to curate synthetic data.

Synthetic data has gained significant traction in the healthcare and finance sectors, where privacy is paramount. Because synthetic data can have fewer legal and regulatory constraints, it is often used as a proxy for natural data [145,146]. Perhaps the best known use of synthetic data in a general machine learning setting occurs in data augmentation [147, 148]. Data augmentation typically increase the size of the training set by transforming, or otherwise manipulating natural examples. Well-designed data augmentation processes can improve overall model performance.

ERM Learning

Our results demonstrate how sample complexity bounds for ERM learning translate into categorywise risk generalization bounds. ERM bounds are common in literature, including bounds on decision trees, neural networks, and other families of classifiers. Given the prevalence of machine learning classifiers and associated bounds, theorem 4.8 makes the rendering of categorywise risk sample complexity bounds accessible for ML practitioners using ERM bounds.

Much of modern ML fairness literature is focused on an application of constrained ERM known as learning fair classifiers. This is accomplished by performing ERM, or some variant thereof, subject to what is known as a fairness constraint. A fairness constraint is often formulated as an optimization constraint and is dependent on population group $\mathtt{g}$, or another protected attribute. Examples of fairness constraints include *statistical parity* (also referred to as *demographic parity*) introduced by Dwork et al. [136] and *equality of opportunity* introduced by Hardt et al. [47]. These works differ from ours in that a learner is assumed to have access to protected attributes: they assume the fairness constraint(s) can be enforced during training.

Multicalibration.

Multicalibration error is a well-studied quantity measuring group-wise performance parity. This notion is useful when classifiers do not have access to group information; Hebert-Johnson et al. were the first to develop and study the notion of multicalibration [126] in the context of machine learning. Liu et al. consider a two-protected group binary decision problem[149]. The authors relate the multicalibration of a learned selection policy subject to a fairness constraint to the prediction utility. Liu et al. also study the relationship between multicalibration and unconstrained machine learning [125]. They show that for a large class of loss functions, if a learned classifier has a small excess risk over the Bayes risk when conditioning on the protected attribute(s), the learned classifier will also be well-calibrated. More recently, Shabat et al. utilize the graph dimension [150], a multi-class generalization of the VC dimension, to determine the sample-complexity necessary for uniform convergence guarantees on multicalibration error [127]. Their results have no dependence on the underlying learning algorithm. Jung et al. [151] design algorithms which yield predictors that have guarantees on higher central moments of multicalibration error. This provides information on the geometry of the multicalibration error beyond what is given by standard mean estimates.

Other Related Work

Wald et al. studied calibration model calibration on across multiple domains [152]. The authors study model performance on out-of-distribution data, but they do not study population categories or subsets.

Related to the multicalibration error is the notion of Outcome Indistinguishability. Outcome Indistinguishability, as introduced by Dwork et al. [128], measures the degree to which a set of examples as labeled by nature differs from the same set of examples labeled as a synthetic or trained classifier. Furthermore, Dwork et al. demonstrate that enforcing out-

come indistinguishability transitively enforces multicalibration and vice-versa. Kearns et al. utilized the VC dimension [129] in their study of auditing algorithms for the prevention of fairness gerrymandering [153]. A closely related paper is Rothblum and Yona's [154] study of Multi-group Probably Approximately Correct (PAC) Learnability. For a fixed loss function, the authors provide an algorithm that produces a classifier that guarantees that each population group has a similar average loss. The authors also provide sample complexity bounds for Multi-group PAC learning.

4.5 Conclusion

This chapter explored machine learning fairness from the perspective of sample complexities for categorywise risk convergence on both natural and synthetic data. We proved that categorywise risk convergence bounds are reparametrizations of ERM learning sample complexity bounds. This reparametrization subsumes existing state-of-the-art sample complexity bounds for loss notions. We demonstrated the utility of our framework in the era of highly-capable generative models which provide more-practical avenues for data augmentation. Furthermore, utilizing our main theorem, we put forth a novel Rademacher complexity style bound on categorywise risk convergence. Beyond these theoretical contributions, we experimentally explored our bounds on linear and sigmoid kernel SVM. Our results provide ML practitioners, equipped with ERM learning bounds, a plug-and-play technique yielding sample complexity bounds for categorywise risk generalization.

Chapter 5

Conclusion

In this dissertation we explore both the societal and privacy implications of modern AI systems. We used modern automated face recognition systems as the medium of exploration. We also presented theory which applies not only to face recognition systems, but to general machine learning models, including models trained on synthetic data. We re-summarize major research contributions and put forth research directions.

5.1 Contributions

Each of chapters 2 to 4 study how representation within training data affects AI system performance:

Chapter 2 shows how, even when steps are taken to preserve privacy in the presence of face recognition, the underlying data distribution still leads to privacy leaks. Because intra-demographic impersonation generally requires weaker perturbation than inter-demographic impersonation, demographic imbalance in enrolled individuals can reveal the composition of the dataset. Furthermore, without explicit access to demographics, face recognition systems are shown to cluster identities by demographic group. The most prominent clustering is by sex: t-SNE clusters for male and female are visually distinct.

Chapter 3 uses generative data to probe for limitations in the most performant commercially available face recognition systems. Discovered failures are associated with uncommon attribute combinations both in society and on the internet. Notably the addition of facial hair to a female, and the addition of longer hair to a male, pose difficulties for leading face recognition systems.

Chapter 4 discusses categorywise generalization bounds for AI systems, including AI systems trained on generated data. The bounds suggest the least frequent category has outsized influence on how much data is needed to generalize a classifier. Well-constructed synthetic data shows promise in rectifying the deficiency, but extreme care must be taken. Poor quality synthetic data can diminish model performance.

5.2 Future Directions

The combination of generative AI, fairness, and privacy avails many future research directions. At a high level, poor generative model implementations, and even those which are well-designed, can leak private information. This information leakage can be disparate along semantic or social lines. Further, information leakage can manifest as performance disparities. Potential mitigation strategies are numerous, each meriting study. Here are a few questions which come to mind: Can additional training on synthetic data mitigate performance disparities or information leakage? Alternatively, can unlearning mitigate this information leakage? Relatedly, what does unlearning mean in the context of generative models? Because humans and machines often perceive sensory inputs differently, we can also investigate discrepancies in measured privacy leakage with respect to generative models.

Privacy considerations do not negate the utility of synthetic data. In this dissertation, we showed how synthetic data can help identify model limitations. Thus far, model limitations were only found in a more human-supervised manner than is desirable: semantics of interest were entirely human specified. Automated techniques to identify any and all potential model

limitations with respect to input semantics would be of significant utility to the community, and to research community broadly.

Once performance disparity, or perhaps another form of privacy leakage, is discovered within a classifier, what is the best way to resolve it? Is finetuning the classifier with synthetic data a good strategy to improve model performance? How well does this strategy scale when addressing multiple conditional failures?

The vast majority of this dissertation focused on face recognition; however, it would be useful to repeat a similar study in other domains. Representation in training data is likely to be the cause of performance disparities in multiple domains, including audio and text. Better characterizations of AI system conditional failure modes are almost certainly to be of interest to machine learning practitioners.

Bibliography

[1] J. D'Orazio, S. Jarrett, A. Amaro-Ortiz, and T. Scott, "Uv radiation and the skin," *International journal of molecular sciences*, vol. 14, no. 6, pp. 12 222–12 248, Jun 2013. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23749111

[2] F. von Luschan, *Beiträge zur Völkerkunde der deutschen Schutzgebiete*. D. Reimer, 1897.

[3] A. Hadden, "The identification of criminals by the bertillon system note," *Western Reserve Law Journal*, vol. 3, no. 7, pp. 165–, 1897-1898. [Online]. Available: https://heinonline.org/HOL/P?h=hein.journals/wrlj3&i=181

[4] J. E. Hoover, "Criminal identification," *The ANNALS of the American Academy of Political and Social Science*, vol. 146, no. 1, pp. 205–213, 1929.

[5] L. Debter, "Retailers quietly deploying controversial technology to combat crime spree," Jan 2022. [Online]. Available: https://www.forbes.com/sites/laurendebter/2022/01/31/retailers-quietly-deploying-controversial-technology-to-combat-crime-spree/?sh=56265c787689

[6] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. Am. A*, vol. 4, no. 3, pp. 519–524, Mar 1987. [Online]. Available: https://opg.optica.org/josaa/abstract.cfm?URI=josaa-4-3-519

[7] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[8] V. Chandrasekaran, C. Gao, B. Tang, K. Fawaz, S. Jha, and S. Banerjee, "Face-off: Adversarial face obfuscation." in *2021 Proceedings on Privacy Enhancing Technologies*, vol. 2021, no. 2, 2021, pp. 369–390.

[9] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting privacy against unauthorized deep learning models," in *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, S. Capkun and F. Roesner, Eds. USENIX Association, 2020, pp. 1589–1604. [Online]. Available: https://www.usenix.org/conference/usenixsecurity20/presentation/shan

[10] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[11] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[12] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 1528–1540.

[13] M. Ngan, P. Grother, and K. Hanaoka, "Part6b: Face recognition accuracy with face masks using post-covid-19 algorithms," *Ongoing Face Recognition Vendor Test*, pp. 1–87, 2021.

[14] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.

[15] V. Albiero, K. Ks, K. Vangara, K. Zhang, M. C. King, and K. W. Bowyer, "Analysis of gender inequality in face recognition accuracy," in *Proceedings of the ieee/cvf winter conference on applications of computer vision workshops*, 2020, pp. 81–89.

[16] H. Rosenberg, B. Tang, K. Fawaz, and S. Jha, "Fairness properties of face recognition and obfuscation systems," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 7231–7248.

[17] H. Rosenberg, S. Ahmed, G. V. Ramesh, R. K. Vinayak, and K. Fawaz, "Unbiased face synthesis with diffusion models: Are we there yet?" *arXiv preprint arXiv:2309.07277*, 2023.

[18] H. Rosenberg, R. Bhattacharjee, K. Fawaz, and S. Jha, "An exploration of multicalibration uniform convergence bounds," *arXiv preprint arXiv:2202.04530*, 2022.

[19] R. B. Fosdick, "Passing of the bertillon system of identification," *J. Am. Inst. Crim. L. & Criminology*, vol. 6, p. 363, 1915.

[20] L. Feiner and A. Palmer, "Rules around facial recognition and policing remain blurry," *CNBC Tech*, 2021.

[21] A. Roussi, "Resisting the rise of facial recognition," *Nature*, 2020.

[22] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019.

[23] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[24] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[25] V. Cherepanova, M. Goldblum, H. Foley, S. Duan, J. P. Dickerson, G. Taylor, and T. Goldstein, "Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition," *CoRR*, vol. abs/2101.07922, 2021. [Online]. Available: https://arxiv.org/abs/2101.07922

[26] I. Evtimov, P. Sturmfels, and T. Kohno, "Foggysight: A scheme for facial lookup privacy," in *2021 Proceedings on Privacy Enhancing Technologies*. PoPETs, 2021.

[27] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=rJzIBfZAb

[28] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.

[29] E. Radiya-Dixit and F. Tramèr, "Data poisoning won't save you from facial recognition," in *ICML Workshop on Adversarial Machine Learning (AdvML)*, 2021. [Online]. Available: https://arxiv.org/abs/2106.14851

[30] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability*

and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. PMLR, 2018, pp. 77–91. [Online]. Available: http://proceedings.mlr.press/v81/buolamwini18a.html

[31] V. Nanda, S. Dooley, S. Singla, S. Feizi, and J. P. Dickerson, "Fairness through robustness: Investigating robustness disparity in deep learning," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 466–477. [Online]. Available: https://doi.org/10.1145/3442188.3445910

[32] H. Xu, X. Liu, Y. Li, A. Jain, and J. Tang, "To be robust or to be fair: Towards fairness in adversarial training," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 11 492–11 501. [Online]. Available: https://proceedings.mlr.press/v139/xu21b.html

[33] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and Simile Classifiers for Face Verification," in *IEEE International Conference on Computer Vision (ICCV)*, Oct 2009.

[34] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, 2006, pp. 1735–1742.

[35] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014. [Online]. Available: http://arxiv.org/abs/1312.6199

[36] A. Holmes, "533 million facebook users' phone numbers and personal data have been leaked online," *Business Insider Tech*, 2021.

[37] D. Harwell, "This facial recognition website can turn anyone into a cop — or a stalker," May 2021. [Online]. Available: https://www.washingtonpost.com/technology/2021/05/14/pimeyes-facial-recognition-search-secrecy/

[38] G. L. Goodwin, "Facial recognition technology: Federal law enforcement agencies should have better awareness of systems used by employees," GAO Report GAO-21-105309, Washington, DC: USA, 2021 [Online]. [Online]. Available: https://www.gao.gov/products/gao-21-105309

[39] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=Sys6GJqxl

[40] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

[41] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[42] L. van der Maaten and G. Hinton, "Viualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 11 2008.

[43] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems*, S. Becker, S. Thrun, and K. Obermayer, Eds.,

vol. 15. MIT Press, 2003. [Online]. Available: https://proceedings.neurips.cc/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf

[44] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 2668–2677. [Online]. Available: http://proceedings.mlr.press/v80/kim18d.html

[45] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019, http://www.fairmlbook.org.

[46] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ser. ITCS '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 214–226. [Online]. Available: https://doi.org/10.1145/2090236.2090255

[47] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, pp. 3315–3323, 2016.

[48] B. L. WELCH, "The Generalization of 'Student's' Problem When Several Different Population Variances Are Involved," *Biometrika*, vol. 34, no. 1-2, pp. 28–35, 01 1947. [Online]. Available: https://doi.org/10.1093/biomet/34.1-2.28

[49] R. A. Alexander and D. M. Govern, "A new and simpler approximation for anova under variance heterogeneity," *Journal of Educational Statistics*, vol. 19, no. 2, pp. 91–101, 1994. [Online]. Available: http://www.jstor.org/stable/1165140

[50] Face++, "Face++," https://www.faceplusplus.com.

[51] "Facial Recognition | Microsoft Azure," Jun. 2022, [Online; accessed 5. Jun. 2022]. [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/face

[52] "Detecting and analyzing faces - Amazon Rekognition," Jun. 2022, [Online; accessed 5. Jun. 2022]. [Online]. Available: https://docs.aws.amazon.com/rekognition/latest/dg/faces.html

[53] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[54] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[55] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[56] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry, "Adversarially robust generalization requires more data," *arXiv preprint arXiv:1804.11285*, 2018.

[57] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 1929–1938. [Online]. Available: https://proceedings.mlr.press/v80/hashimoto18a.html

[58] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proceedings of the 36th International Conference on*

Machine Learning, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 1310–1320. [Online]. Available: https://proceedings.mlr.press/v97/cohen19c.html

[59] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. IEEE, 2017, pp. 39–57.

[60] P. L. Bartlett, D. J. Foster, and M. Telgarsky, "Spectrally-normalized margin bounds for neural networks," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6241–6250.

[61] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, "Exploring generalization in deep learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 5949–5958.

[62] K. Scaman and A. Virmaux, "Lipschitz regularity of deep neural networks: Analysis and efficient estimation," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 3839–3848.

[63] T.-W. Weng, H. Zhang, P.-Y. Chen, J. Yi, D. Su, Y. Gao, C.-J. Hsieh, and L. Daniel, "Evaluating the robustness of neural networks: An extreme value theory approach," in *International Conference on Learning Representations (ICLR)*, may 2018.

[64] U. V. Luxburg and O. Bousquet, "Distance-based classification with lipschitz functions," in *J. Mach. Learn. Res.*, 2003.

[65] H. Zhang, P. Zhang, and C.-J. Hsieh, "Recurjac: An efficient recursive algorithm for bounding jacobian matrix of neural networks and its applications," in *AAAI Conference on Artificial Intelligence (AAAI), arXiv preprint arXiv:1810.11783*, dec 2019.

[66] T.-W. Weng, H. Zhang, H. Chen, Z. Song, C.-J. Hsieh, D. Boning, and I. S. D. A. Daniel, "Towards fast computation of certified robustness for relu networks," in *International Conference on Machine Learning (ICML)*, july 2018.

[67] I. Serna, A. Morales, J. Fierrez, M. Cebrian, N. Obradovich, and I. Rahwan, "Sensitiveloss: Improving accuracy and fairness of face representations with discrimination-aware deep learning," 2020.

[68] K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1548–1558.

[69] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, p. 64–73, Jan. 2016. [Online]. Available: https://doi.org/10.1145/2812802

[70] C. Song and V. Shmatikov, "Overlearning reveals sensitive attributes," *arXiv preprint arXiv:1905.11742*, 2019.

[71] V. Cherepanova, S. Reich, S. Dooley, H. Souri, M. Goldblum, and T. Goldstein, "A deep dive into dataset imbalance and bias in face identification," *arXiv preprint arXiv:2203.08235*, 2022.

[72] A. Rajabi, R. B. Bobba, M. Rosulek, C. V. Wright, and W. chi Feng, "On the (im)practicality of adversarial perturbation for image privacy," *Proceedings on Privacy Enhancing Technologies*, vol. 2021, no. 1, pp. 85–106, 2021. [Online]. Available: https://doi.org/10.2478/popets-2021-0006

[73] S. Qin, "Bias and fairness of evasion attacks in image perturbation," Master's thesis, Central Washington University, 2021.

[74] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[75] P. J. Phillips, P. Grother, R. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, "Face recognition vendor test 2002," in *2003 IEEE International SOI Conference. Proceedings (Cat. No. 03CH37443)*. IEEE, 2003, p. 44.

[76] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic bias in biometrics: A survey on an emerging challenge," *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89–103, 2020.

[77] S. H. Abdurrahim, S. A. Samad, and A. B. Huddin, "Review on the effects of age, gender, and race demographics on automatic face recognition," *The Visual Computer*, vol. 34, pp. 1617–1630, 2018.

[78] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Transactions on information forensics and security*, vol. 7, no. 6, pp. 1789–1801, 2012.

[79] V. Albiero, K. Bowyer, K. Vangara, and M. King, "Does face recognition accuracy get better with age? deep face matchers say no," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 261–269.

[80] B. Lu, J.-C. Chen, C. D. Castillo, and R. Chellappa, "An experimental evaluation of covariates effects on unconstrained face verification," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 42–55, 2019.

[81] P. G. Schyns and H. H. Bulthoff, "Viewpoint dependence and face recognition," in *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. Routledge, 2019, pp. 789–793.

[82] A. Bhatta, V. Albiero, K. W. Bowyer, and M. C. King, "The gender gap in face recognition accuracy is a hairy problem," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 303–312.

[83] P. Terhörst, J. N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. M. Moreno, J. Fierrez, and A. Kuijper, "A comprehensive study on face recognition biases beyond demographics," *IEEE Transactions on Technology and Society*, vol. 3, no. 1, pp. 16–30, 2021.

[84] P. Terhörst, D. Fährmann, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "Maad-face: A massively annotated attribute dataset for face images," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3942–3957, 2021.

[85] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[86] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015.

[87] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[88] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep face recognition: A survey," in *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*. IEEE, 2018, pp. 471–478.

[89] I. Hupont and C. Fernández, "Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019, pp. 1–7.

[90] K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 1548–1558.

[91] J. P. Robinson, C. Qin, Y. Henon, S. Timoner, and Y. Fu, "Balancing biases and preserving privacy on balanced faces in the wild," *IEEE Transactions on Image Processing*, 2023.

[92] H. Wu and K. W. Bowyer, "A real balanced dataset for understanding bias? factors that impact accuracy, not numbers of identities and images," *arXiv preprint arXiv:2304.09818*, 2023.

[93] O. Wiles, I. Albuquerque, and S. Gowal, "Discovering bugs in vision models using off-the-shelf image generation and captioning," *arXiv preprint arXiv:2208.08831*, 2022.

[94] J. Vendrow, S. Jain, L. Engstrom, and A. Madry, "Dataset interfaces: Diagnosing model failures using controllable counterfactual generation," *arXiv preprint arXiv:2302.07865*, 2023.

[95] P. Jacob, É. Zablocki, H. Ben-Younes, M. Chen, P. Pérez, and M. Cord, "Steex: steering counterfactual explanations with semantics," in *European Conference on Computer Vision*. Springer, 2022, pp. 387–403.

[96] E. Denton, B. Hutchinson, M. Mitchell, T. Gebru, and A. Zaldivar, "Image counterfactual sensitivity analysis for detecting unintended bias," *arXiv preprint arXiv:1906.06439*, 2019.

[97] G. Jeanneret, L. Simon, and F. Jurie, "Diffusion models for counterfactual explanations," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 858–876.

[98] J. R. Williford, B. B. May, and J. Byrne, "Explainable face recognition," in *European conference on computer vision*. Springer, 2020, pp. 248–263.

[99] T. Kay Lively, "Facial recognition in the us: Privacy concerns and legal developments," 2021. [Online]. Available: https://www.asisonline.org/security-management-magazine/monthly-issues/security-technology/archive/2021/december/facial-recognition-in-the-us-privacy-concerns-and-legal-developments/

[100] S. Zhang, Y. Feng, and N. Sadeh, "Facial recognition: Understanding privacy concerns and attitudes across increasingly diverse deployment scenarios," in *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, 2021, pp. 243–262.

[101] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. CVPR*, 2020.

[102] L. Colbois, T. de Freitas Pereira, and S. Marcel, "On the use of automatically generated synthetic image datasets for benchmarking face recognition," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2021, pp. 1–8.

[103] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[104] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[105] M. Kim, F. Liu, A. Jain, and X. Liu, "Dcface: Synthetic face generation with dual condition diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 715–12 725.

[106] P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, D. Lawatsch, F. Domin, and M. Schaubert, "Gandiffface: Controllable generation of synthetic datasets for face recognition with realistic variations," *arXiv preprint arXiv:2305.19962*, 2023.

[107] S. Banerjee, G. Mittal, A. Joshi, C. Hegde, and N. Memon, "Identity-preserving aging of face images via latent diffusion models," *arXiv preprint arXiv:2307.08585*, 2023.

[108] M. Brack, F. Friedrich, D. Hintersdorf, L. Struppek, P. Schramowski, and K. Kersting, "Sega: Instructing diffusion using semantic dimensions," *arXiv preprint arXiv:2301.12247*, 2023.

[109] H. Rosenberg, B. Tang, K. Fawaz, and S. Jha, "Fairness properties of face recognition and obfuscation systems," in *32nd USENIX Security Symposium (USENIX Security 23)*. Anaheim, CA: USENIX Association, Aug. 2023, pp. 7231–7248. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/rosenberg

[110] L. G. Farkas, M. J. Katic, and C. R. Forrest, "International anthropometric study of facial morphology in various ethnic groups/races," *Journal of Craniofacial Surgery*, vol. 16, no. 4, pp. 615–646, 2005.

[111] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, and D. H. Chau, "Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models," *arXiv preprint arXiv:2210.14896*, 2022.

[112] Y. Zhang, K. Zhou, and Z. Liu, "Neural prompt search," *ArXiv*, vol. abs/2206.04673, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249538657

[113] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022.

[114] 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 *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 1493–1504.

[115] A. S. Luccioni, C. Akiki, M. Mitchell, and Y. Jernite, "Stable bias: Analyzing societal representations in diffusion models," *arXiv preprint arXiv:2303.11408*, 2023.

[116] P. Seshadri, S. Singh, and Y. Elazar, "The bias amplification paradox in text-to-image generation," *arXiv preprint arXiv:2308.00755*, 2023.

[117] F. Friedrich, P. Schramowski, M. Brack, L. Struppek, D. Hintersdorf, S. Luccioni, and K. Kersting, "Fair diffusion: Instructing text-to-image generation models on fairness," *arXiv preprint arXiv:2302.10893*, 2023.

[118] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace, "Extracting training data from diffusion models," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 5253–5270.

[119] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278–25 294, 2022.

[120] S. Karthik, K. Roth, M. Mancini, and Z. Akata, "If at first you don't succeed, try, try again: Faithful diffusion-based text-to-image generation by selection," *arXiv preprint arXiv:2305.13308*, 2023.

[121] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," *arXiv preprint arXiv:2310.03744*, 2023.

[122] Z. Lin, A. Khetan, G. Fanti, and S. Oh, "Pacgan: The power of two samples in generative adversarial networks," *Advances in neural information processing systems*, vol. 31, 2018.

[123] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, "Assessing generative models via precision and recall," *Advances in neural information processing systems*, vol. 31, 2018.

[124] R. Binns, "On the apparent conflict between individual and group fairness," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 514–524. [Online]. Available: https://doi.org/10.1145/3351095.3372864

[125] L. T. Liu, M. Simchowitz, and M. Hardt, "The implicit fairness criterion of unconstrained learning," in *International Conference on Machine Learning*, 2019, pp. 4051–4060.

[126] U. Hebert-Johnson, M. Kim, O. Reingold, and G. Rothblum, "Multicalibration: Calibration for the (Computationally-identifiable) masses," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 1939–1948. [Online]. Available: https://proceedings.mlr.press/v80/hebert-johnson18a.html

[127] E. Shabat, L. Cohen, and Y. Mansour, "Sample complexity of uniform convergence for multicalibration," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 13 331–13 340. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/9a96876e2f8f3dc4f3cf45f02c61c0c1-Paper.pdf

[128] C. Dwork, M. P. Kim, O. Reingold, G. N. Rothblum, and G. Yona, "Outcome indistinguishability," in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021, pp. 1095–1108.

[129] V. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability & Its Applications*, vol. 16, no. 2, pp. 264–280, 1971.

[130] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 463–482, 2002.

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

[132] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., vol. 19. MIT Press, 2006. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2006/file/b1b0432ceafb0ce714426e9114852ac7-Paper.pdf

[133] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1, pp. 151–175, 2010. [Online]. Available: https://doi.org/10.1007/s10994-009-5152-4

[134] K. Crammer, M. Kearns, and J. Wortman, "Learning from multiple sources," *Journal of Machine Learning Research*, vol. 9, no. 57, pp. 1757–1774, 2008. [Online]. Available: http://jmlr.org/papers/v9/crammer08a.html

[135] C. Dwork and C. Ilvento, "Fairness under composition," in *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*, 2019.

[136] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[137] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. PMLR, 23–24 Feb 2018, pp. 77–91. [Online]. Available: https://proceedings.mlr.press/v81/buolamwini18a.html

[138] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[139] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, S. Dasgupta and D. McAllester, Eds., vol. 28, no. 3. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, pp. 325–333. [Online]. Available: https://proceedings.mlr.press/v28/zemel13.html

[140] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Oct 2016, pp. 399–410.

[141] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," *Advances in neural information processing systems*, vol. 32, 2019.

[142] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[143] S. Ö. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 8, 2021, pp. 6679–6687.

[144] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[145] J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S. N. Cohen, and A. Weller, "Synthetic data–what, why and how?" *arXiv preprint arXiv:2205.03257*, 2022.

[146] S. James, C. Harbron, J. Branson, and M. Sundler, "Synthetic data use: exploring use cases to optimise data utility," *Discover Artificial Intelligence*, vol. 1, no. 1, p. 15, 2021.

[147] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[148] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019. [Online]. Available: https://doi.org/10.1186/s40537-019-0197-0

[149] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt, "Delayed impact of fair machine learning," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 3150–3158. [Online]. Available: https://proceedings.mlr.press/v80/liu18c.html

[150] A. Daniely, S. Sabato, S. Ben-David, and S. Shalev-Shwartz, "Multiclass learnability and the erm principle," in *Proceedings of the 24th Annual Conference on Learning Theory*, 2011, pp. 207–232.

[151] C. Jung, C. Lee, M. Pai, A. Roth, and R. Vohra, "Moment multicalibration for uncertainty estimation," in *Proceedings of Thirty Fourth Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, M. Belkin and S. Kpotufe, Eds., vol. 134. PMLR, 15–19 Aug 2021, pp. 2634–2678. [Online]. Available: https://proceedings.mlr.press/v134/jung21a.html

[152] Y. Wald, A. Feder, D. Greenfeld, and U. Shalit, "On calibration and out-of-domain generalization," *Advances in neural information processing systems*, vol. 34, pp. 2215–2227, 2021.

[153] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *International Conference on Machine Learning*, 2018, pp. 2564–2572.

[154] G. N. Rothblum and G. Yona, "Multi-group agnostic pac learnability," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 9107–9115. [Online]. Available: https://proceedings.mlr.press/v139/rothblum21a.html

[155] S. I. Serengil and A. Ozpinar, "Lightface: A hybrid deep face recognition framework," in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2020, pp. 23–27.

[156] H. Ashtiani, S. Ben-David, and A. Mehrabian, "Sample-efficient learning of mixtures," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[157] C. Cortes, M. Mohri, and A. Rostamizadeh, "Generalization bounds for learning kernels," in *Proceedings of the 27th Annual International Conference on Machine Learning (ICML 2010)*, 2010. [Online]. Available: http://www.cs.nyu.edu/~mohri/pub/lk.pdf

Appendix A

Chapter 2 Appendix

A.1 Mathematics for the Analytical Model

To solve the optimization problem posed in eq. (2.9), we examine the likelihood function. Let $p_{\mathcal{D}_\mathsf{g}}$ now denote the PDF of the image of distribution $\mathcal{D}_\mathsf{g}$ as it appears in the 1-D PCA embedding space. We assume $\gamma \leq 1$. We aim to find the strength of the perturbation necessary to push an example $\boldsymbol{x}$ in a $\frac{p_{\mathcal{D}_\mathsf{a}}[\mathsf{f}(\boldsymbol{x}+\boldsymbol{\delta})]}{p_{\mathcal{D}_\mathsf{b}}[\mathsf{f}(\boldsymbol{x}+\boldsymbol{\delta})]}$ exceeds one, where:

$$1 \leq \left(\frac{p_{\mathcal{D}_\mathsf{a}}\left[\mathrm{f}(\boldsymbol{x}+\boldsymbol{\delta}) \mid \boldsymbol{\delta} \propto \frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}\right]}{p_{\mathcal{D}_\mathsf{b}}\left[\mathrm{f}(\boldsymbol{x}+\boldsymbol{\delta}) \mid \boldsymbol{\delta} \propto \frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}\right]} \right) \tag{A.1}$$

$$\begin{aligned} &= \left((2\pi)^{-1/2}(\mathrm{f}(\boldsymbol{q}_1))^{-1}\right) \\ &\quad \times \exp\left\{-\frac{1}{2}\left[\mathrm{f}\left(\boldsymbol{x}+\eta\frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}-\boldsymbol{\mu}_\mathsf{a}\right)\right]^2(\mathrm{f}(\boldsymbol{q}_1))^{-2}\right\} \\ &\quad \left[\left((2\gamma\pi)^{-1/2}(\mathrm{f}(\boldsymbol{q}_1))^{-1}\right)\right. \\ &\quad \left.\times \exp\left\{-\frac{\gamma}{2}\cdot\left[\mathrm{f}\left(\boldsymbol{x}+\eta\frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}-\boldsymbol{\mu}_\mathsf{b}\right)\right]^2(\mathrm{f}(\boldsymbol{q}_1))^{-2}\right\}\right]^{-1} \end{aligned} \tag{A.2}$$

$$\begin{aligned} &\leq \exp\left\{-\frac{1}{2}\left[\mathrm{f}\left(\boldsymbol{x}+\eta\frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}\right)+\mathrm{f}\left(\boldsymbol{\mu}_\mathsf{b}\right)\right]^2\right\} \\ &\quad \times \exp\left\{-\frac{\gamma}{2}\cdot\left[\mathrm{f}\left(\boldsymbol{x}+\eta\frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}\right)-\mathrm{f}\left(\boldsymbol{\mu}_\mathsf{b}\right)\right]^2\right\}\Bigg]^{-1} \end{aligned} \tag{A.3}$$

We now solve the following the following inequality for η

$$\begin{aligned} 1 \leq &\exp\left\{-\frac{1}{2}\left[\mathrm{f}\left(\boldsymbol{x}+\eta\frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}\right)+\mathrm{f}\left(\boldsymbol{\mu}_\mathsf{b}\right)\right]^2\right\} \\ &\times \exp\left\{-\frac{\gamma}{2}\cdot\left[\mathrm{f}\left(\boldsymbol{x}+\eta\frac{(\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b}-\boldsymbol{x}\|_2}\right)-\mathrm{f}\left(\boldsymbol{\mu}_\mathsf{b}\right)\right]^2\right\}\Bigg]^{-1} \end{aligned} \tag{A.4}$$

After some algebra, we arrive at the following interval solution for η.

$$\eta > \frac{2b(1-\gamma)\sqrt{\frac{a^2\gamma}{b^2(\gamma-1)^2}}+a\gamma+a+\mathrm{f}(\boldsymbol{\mu}_\mathsf{b})(1-\gamma)}{b(\gamma-1)} \tag{A.5}$$

AND

$$\eta < \frac{2b(\gamma-1)\sqrt{\frac{a^2\gamma}{b^2(\gamma-1)^2}}+a\gamma+a+\mathrm{f}(\boldsymbol{\mu}_\mathsf{b})(1-\gamma)}{b(\gamma-1)} \tag{A.6}$$

Where, for notational compactness, we have:

$$a = \mathrm{f}(\boldsymbol{x}) \tag{A.7}$$

$$b = \mathrm{f}\left(\frac{(\boldsymbol{\mu}_\mathsf{b} - \boldsymbol{x})}{\|\boldsymbol{\mu}_\mathsf{b} - \boldsymbol{x}\|_2}\right) \tag{A.8}$$

Since the right-side of inequality (A.4) upper bounds the right-side of inequality (A.1), we know that any solution for inequality (A.4) is also a solution for inequality (A.1).

Since inequality (A.6) is a bound on ϵ which provides a guarantee on when eq. (2.9) is infeasible. Hence we conclude that eq. (2.9) may be feasible only when

$$\epsilon \geq \max\left\{0, \frac{2b(\gamma-1)\sqrt{\frac{a^2\gamma}{b^2(\gamma-1)^2}} + a\gamma + a + \mathrm{f}(\boldsymbol{\mu}_\mathsf{b})(1-\gamma)}{b(\gamma-1)}\right\} \tag{A.9}$$

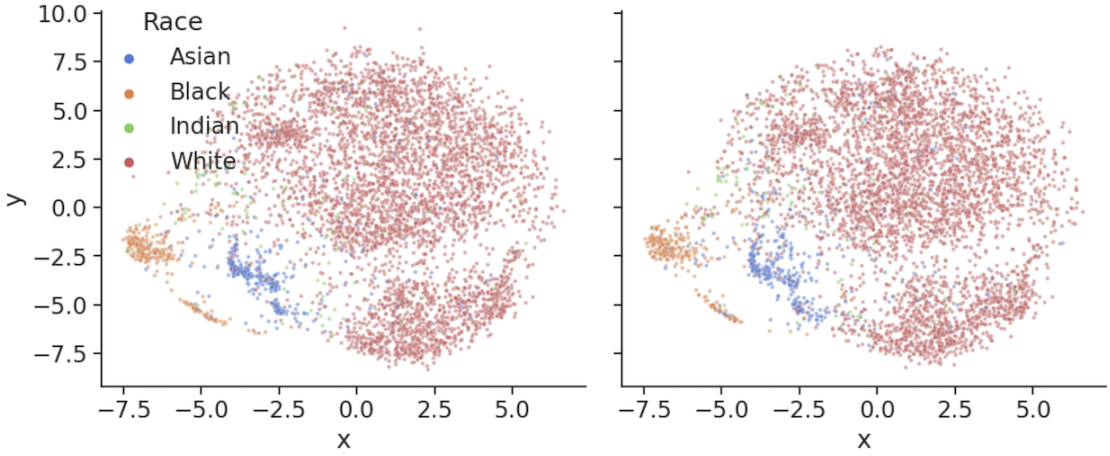
A.2 Additional Experiments

This section contains additional experiments furthering our understanding of face recognition/obfuscation fairness issues.

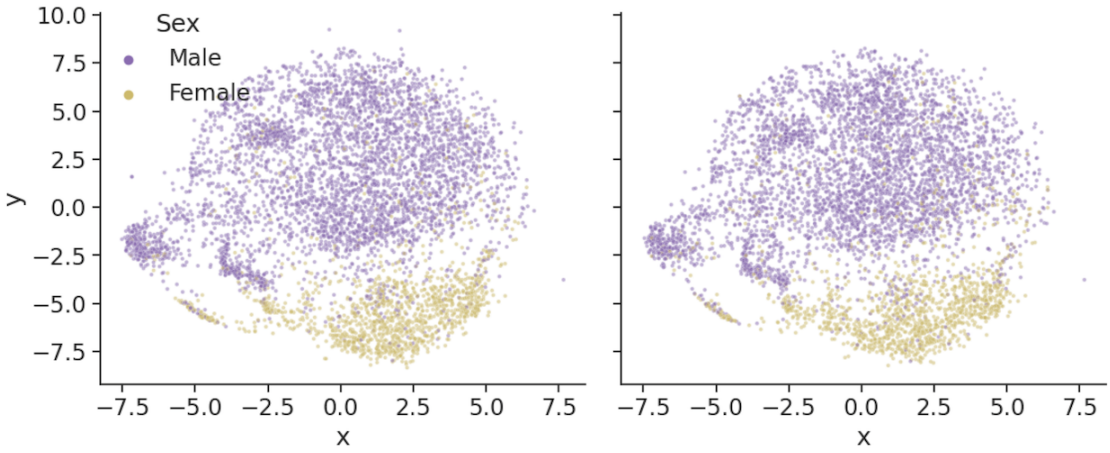
A.2.1 Face Obfuscation and Demographics

While adversarial examples cause a misclassification on a particular identity, these adversarial examples do not necessarily cause their demographic attributes to change. To validate this claim, we generate untargeted adversarial examples on the LFW dataset and classify these perturbed images using a classifier that predicts the race attribute from a face [155]. Of each identity in the dataset, only 8% of the identities' race attribute change after adversarial perturbations are added. We visualize this in fig. A.1, where the embeddings of the natural examples and adversarial examples are plotted according to their ground truth demographics using t-SNE. Comparing the plots within figs. A.1a and A.1b, a heavy overlap is

observed between the embeddings of both natural and adversarial examples of the same race and the same sex. These results further addressed confirm our intuition from section 2.3.2; an obfuscated face is unlikely to depart it's demographic group within the embedding space. This is especially true when the groups are clustered in the embedding space.



(a) Race: Natural (Left) vs. Adversarial (Right)



(b) Sex: Natural (Left) vs. Adversarial (Right)

Figure A.1: t-SNE of the natural and adversarial embeddings. Note that the clusters do not change, identities are still rooted within their own demographic.

I	II	III	IV	V	VI
0xF4F2F5	0xFAF0EF	0xFFF9E2	0xE4C567	0x8F573C	0x2D2024
0xEDECEA	0xF3EBE6	0xF1E8C4	0xE2C26A	0x7A4C2C	0x14152A
0xFAF9F7	0xF4F1EB	0xF0E3AE	0xE0C27C	0x642D0E	
0xFDFBE7	0xFBFCF4	0xE1D394	0xDFB978	0x652C1A	
0xFDF6E7	0xFCF8EE	0xF2E398	0xC8A664	0x602D1B	
0xFEF7E6	0xFEF6E2	0xECD7A0	0xBD9862	0x562E24	
	0xFFF9E2	0xECDA86	0x9D6B41	0x3E1A0D	

Table A.1: This chart depicts colors in the Fitzpatrick scale [1], which is derived from Von Luschan's chromatic scale [2]. Colors are in hexadecimal.

A.2.2 Black-Box Obfuscation Disparities

To understand whether the performance disparities between demographics (as discussed in section 2.3.2) manifest in commercial face recognition systems, we tested the success of the perturbed faces against three commercial face recognition systems. These Face Recognition APIs were Face++, Microsoft Azure Face API, and Amazon AWS Rekognition. In fig. 2.7, we observe large differences in obfuscation success for each race demographic. This performance disparity can be attributed to the larger perturbation sizes associated with Asian and White identities (fig. 2.5). Black identities were observed to have smaller perturbations, and higher local Lipschitz constants (fig. A.2), suggesting that decreased obfuscation success can stem from disparities in the robustness of each demographic.

A.2.3 Targeted vs. Untargeted Obfuscation

Figure A.3 portrays the untargeted obfuscation success rates for the pre-trained FaceNet model. Untargeted success is also provided for the black-box setting with the OpenFace model (fig. A.4). In fig. 2.14, the targeted success rates are provided for the Race-Balanced and Sex-Balanced FaceNet models. We observe similar trends as discussed in sections 2.3.2 and 2.3.3.

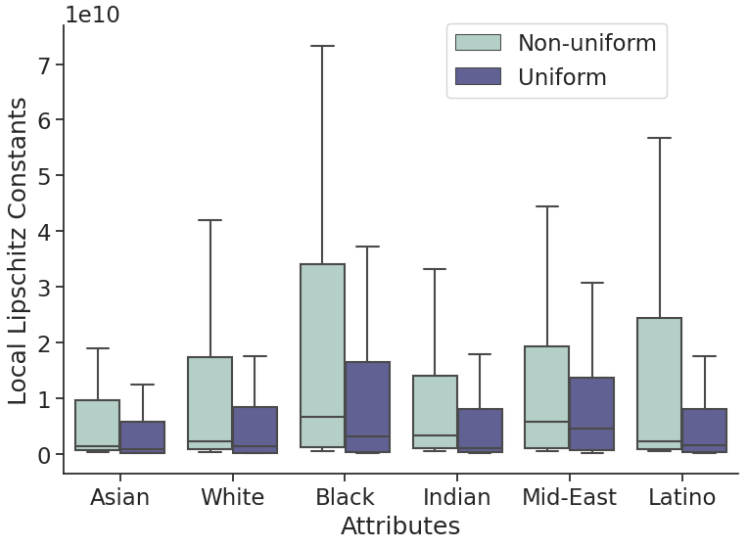
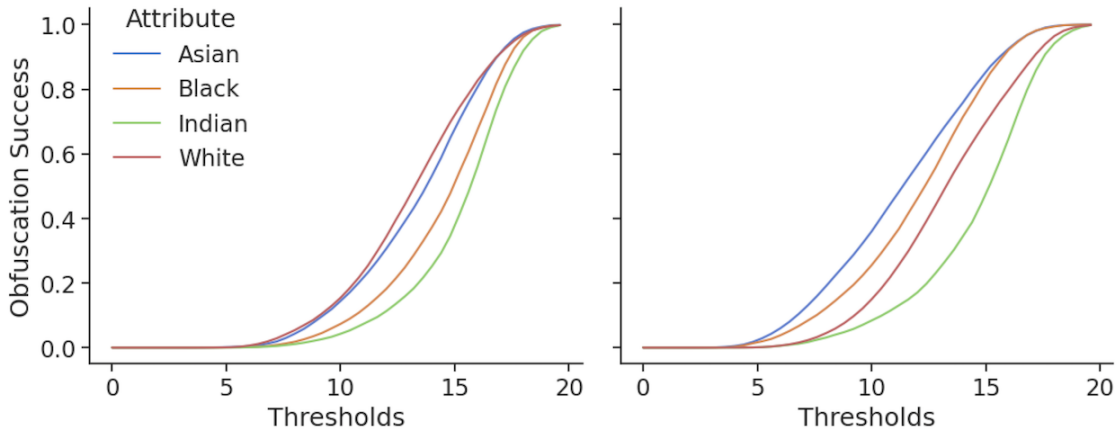
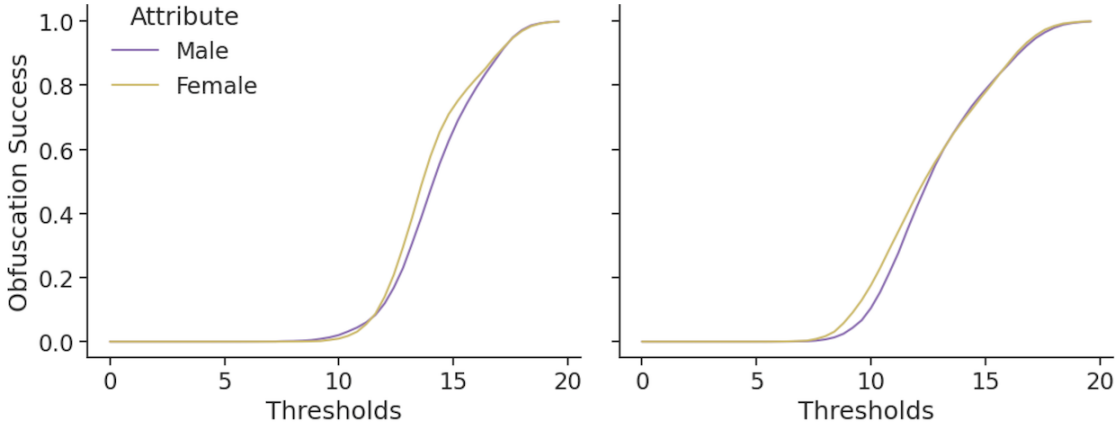


Figure A.2: Upper bounds on local Lipschitz constants estimated using Fast-Lin and RecurJac. Models trained on non-uniform demographic distributions suffer from higher instability. The maximum constants for each demographic is 2-3 times larger in the non-uniform model than in the uniform model.

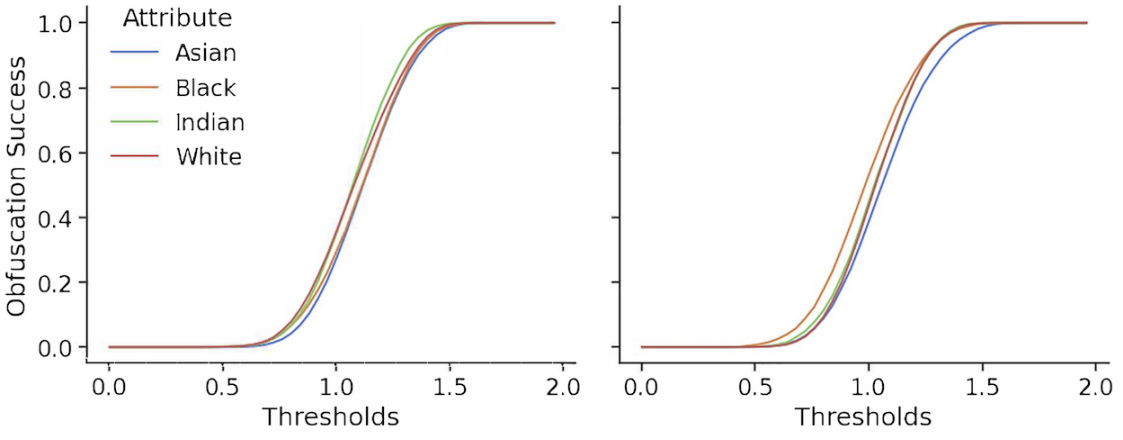


(a) Obfuscation success on FaceNet: Different (Left) vs. Same (Right) Race

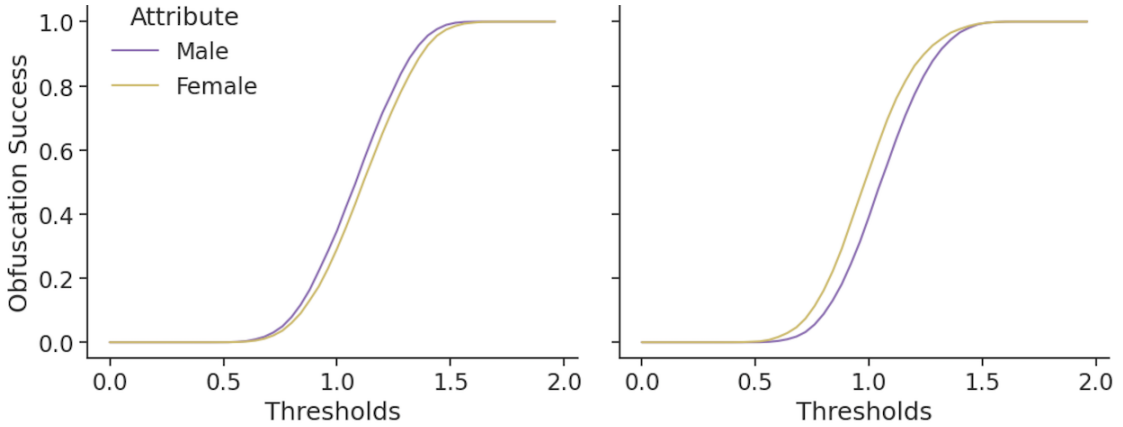


(b) Obfuscation success on FaceNet: Different (Left) vs. Same (Right) Sex

Figure A.3: Untargeted obfuscation success evaluated on the FaceNet metric embedding network in a white-box setting.



(a) Obfuscation success on OpenFace: Different (Left) vs. Same (Right) Race



(b) Obfuscation success on OpenFace: Different (Left) vs. Same (Right) Sex

Figure A.4: Untargeted obfuscation success evaluated on the OpenFace metric embedding network in a black-box setting.

Appendix B

Chapter 3 Appendix

B.1 Proof of Proposition

In this section, we provide the proof of proposition 3.8:

Proof. We will leverage triangle inequality to lower bound $\rho_{\mathrm{TV}}(\hat{\mathcal{D}}_S, \mathcal{D})$. That decomposition is

$$\rho_{\mathrm{TV}}(\mathcal{D}_S, \mathcal{D}) \leq \rho_{\mathrm{TV}}(\mathcal{D}_S, \hat{\mathcal{D}}_S) + \rho_{\mathrm{TV}}(\hat{\mathcal{D}}_S, \mathcal{D}) \tag{B.1}$$

The squared Hellinger distance lower bounds discrepancy. Hence, we have

$$\rho_{\mathrm{TV}}(\mathcal{D}_S, \mathcal{D}) = \sup_{z \in \mathcal{F}} |\mathcal{D}_S(z) - \mathcal{D}(z)| \tag{B.2}$$

$$\begin{aligned} &= \sup_{z \in \mathcal{F}} \Big| \beta \mathcal{D}_a(z) + (1-\beta)\mathcal{D}_b(z) \\ &\qquad - (\alpha \mathcal{D}_a(z) + (1-\alpha)\mathcal{D}_b(z)) \Big| \end{aligned} \tag{B.3}$$

$$= |\alpha - \beta| \rho_{\mathrm{TV}}(\mathcal{D}_a, \mathcal{D}_b) \tag{B.4}$$

$$\geq |\alpha - \beta| H^2(\mathcal{D}_a, \mathcal{D}_b) \tag{B.5}$$

We also need an upper bound on $\rho_{\mathrm{TV}}(\mathcal{D}_S, \hat{\mathcal{D}}_S)$. That upper-bound is sourced from Ashtiani et al. [156].

Theorem B.1 (PAC learnability of Gaussian Mixtures [156])**.** *Let $\epsilon, \delta > 0$. Let S be drawn i.i.d. from $\mathcal{Q}$, and let $\hat{\mathcal{Q}}$ be the distribution learned from S. With probability at least $1 - \delta$, given $O(kd^2 \log(k) \log(k/\delta)/(\epsilon)^4)$ samples in S, a realizable mixture of k-Gaussian mixtures is ϵ-learned, i.e*

$$\mathbb{P}[\rho_{\mathrm{TV}}(\hat{\mathcal{Q}}, \mathcal{Q}) \leq \epsilon] \geq 1 - \delta \tag{B.6}$$

Thus, combining the theorem and triangle inequality, we arrive at a high probability lower-bound on $\rho_{\mathrm{TV}}(\hat{\mathcal{D}}_S, \mathcal{D})$. When $|S| = O(d^2 \log(2) \log(2/\delta)/\epsilon^4)$, then

$$\mathbb{P}\left[\rho_{\mathrm{TV}}(\hat{\mathcal{D}}_S, \mathcal{D}) > |\alpha - \beta| H^2(\mathcal{D}_a, \mathcal{D}_b) - \epsilon\right] \geq 1 - \delta \tag{B.7}$$

Setting $\epsilon = \frac{|\alpha-\beta|}{2} H^2(\mathcal{D}_a, \mathcal{D}_b)$, we arrive at the proposition.

B.2 Identity Selection

figs. B.1 and B.2 contain examples of the CLIP retrieval tool's Graphical User Interface. We used the available API to retrieve examples from the dataset and select identities for different demographic groups (as explained in section 3.2.2). fig. B.1 is an example where the tool picks the examples of the expected identity from the LAION-5B dataset. The tool is largely successful in retrieving examples from the dataset. However, we noticed failed instances, like in fig. B.2, where a different identity, albeit similarly spelled, is picked.

table B.2 contains the names of 50 identities per demographic that were filtered by CLIP-based retrieval. We used GPT 3.5 to associate names from IMDB and LFW celebrity list into the eight demographic-sex groups. GPT 3.5 fails to correctly associate some names. The failures are also noted in table B.2. table B.3 contains the list of final identities chosen

for transformed face generation and face API evaluation. We manually verified these faces.

B.3 Transformed Faces

fig. B.3 and fig. B.4 contain examples of transformed faces for attributes that were not covered in fig. 3.2. As can be noticed in all these figures, the transformations produced by SEGA are not always perfect. Firstly, we noticed some of the transformations led to distorted faces (fig. B.5). As described in section 3.2.2, such failures were removed from our evaluation set with LLaVA. We also noticed cases in which SEGA's image edits were not faithful to the edit prompt. This can be observed in fig. B.6.

SEGA struggles to manipulate certain face attributes. fig. B.7 contains examples of a subset of these attributes where SEGA fails to capture the semantics of the edit prompt and either applies a wrong transformation or performs no operation. As these attributes failed to work a large proportion of the faces, we didn't consider them for final evaluation. Examples of such attribute include coloring eyebrows, changing the shape and size of nose, distance between the eyes etc.

B.4 Dataset Quality Validation

As explained in section 3.2.2, we use LLaVA-1.5 to detect and filter out distorted images. This ensures that any degradation in face recognition performance is solely attributed to the semantic shift induced by SEGA, rather than SEGA's potential failure to produce high-quality transformed images. We validated LLaVA's efficacy via a user survey, where annotators manually labeled the transformed images either as *distorted* or *not distorted*.

table B.1 presents the confusion matrix of LLaVA's detection performance. The data shows that of the 873 images LLaVA identified as *not distorted*, only 17 were labeled as distorted by annotators. Hence, the distorted image rate—defined as the proportion of

actual distorted images out of all images passed after LLaVA's filtration process—equals only $\frac{17}{(856+17)} = 1.95\%$.

Note that, upon review, we found that we had an arithmetic error in the main these text regarding the false positive and negative rates (where positive indicates the presence of a distortion). The correct numbers are: the false positive rate equals $\frac{272}{(272+856)} = 24.11\%$, the false negative rate equals $\frac{17}{(17+50)} = 25.37\%$. Importantly, these adjustments do not alter the chapter's fundamental conclusions: the distorted image rate, which quantifies the proportion of distorted images in the evaluation set, is only 1.95%. LLaVA generally removes SEGA's distortions from the evaluation set. Therefore, we are still able to faithfully measure the impact of semantic transformations on face recognition models.

		LLaVA Response	
		Not Distorted	**Distorted**
Survey Response	**Not Distorted**	856	272
	Distorted	17	50

Table B.1: Confusion Matrix of LLaVA's image distortion prediction against survey results on 1195 validation images.

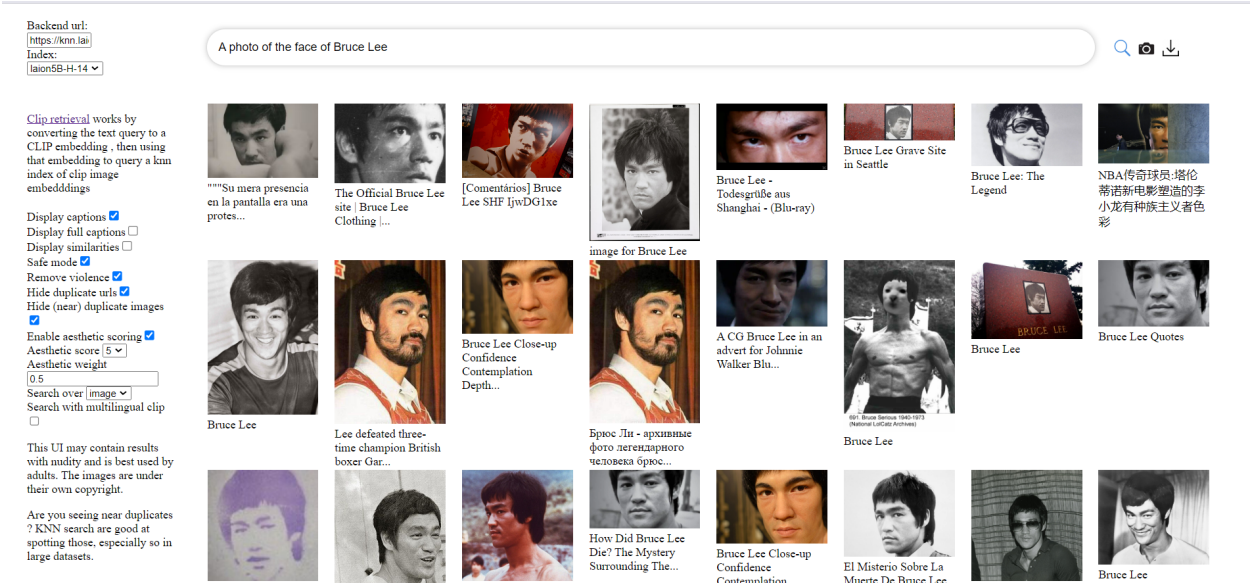


Figure B.1: Example of a successful retrieval with the CLIP-based retrieval tool

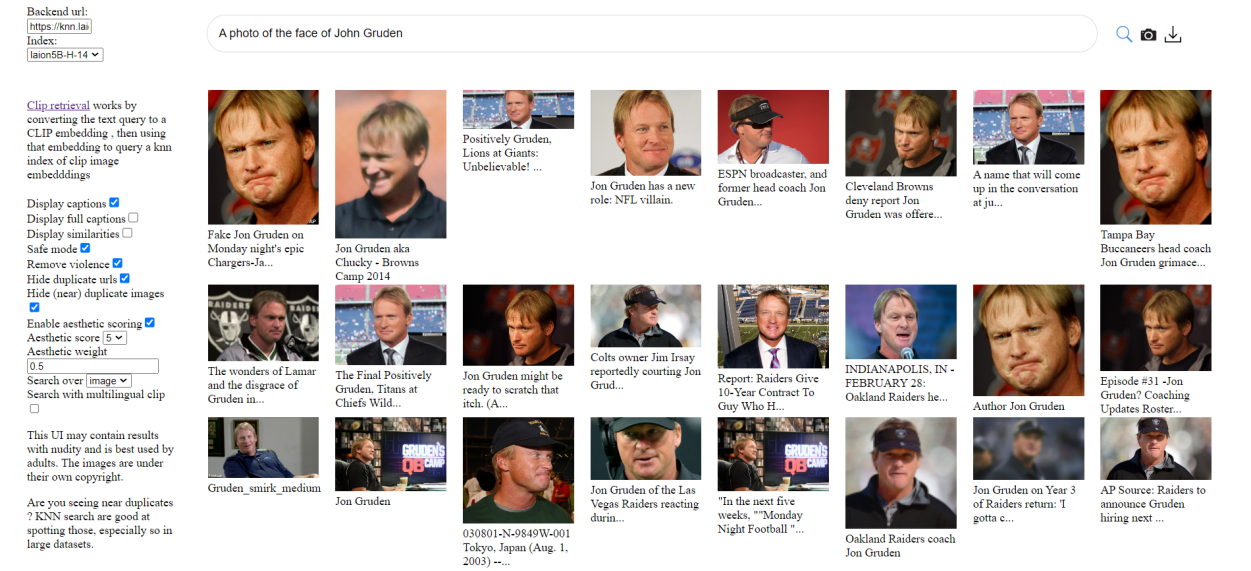


Figure B.2: Example of a failure with the CLIP-based retrieval tool. "John Gruden" is queried, yet images of Jon Gruden are displayed. There is a relatively famous John Gruden, but he is less well-known than Jon Gruden.

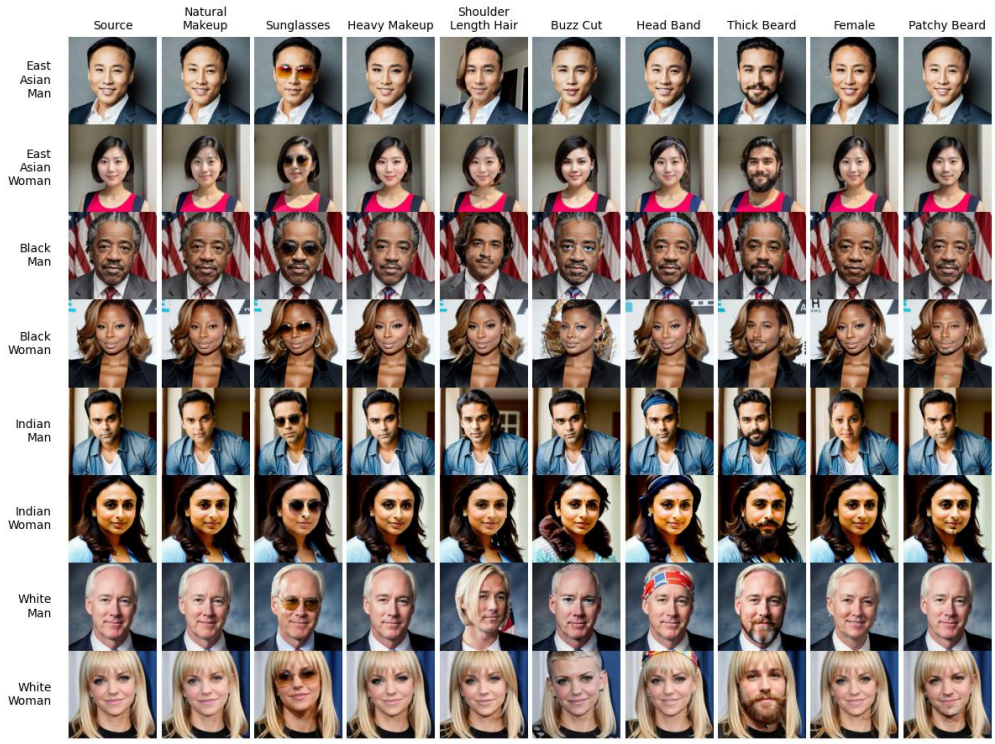


Figure B.3: Additional examples of transformed faces.

Figure B.4: Further additional examples of transformed faces.

Figure B.5: Distorted faces picked by LLaVA: The first row includes examples where the distortions are pronounced, occurring in areas of the face where the edit attribute doesn't belong to. The second row includes examples picked by LLaVA with relatively minimal distortions (second column) or none at all (last column).

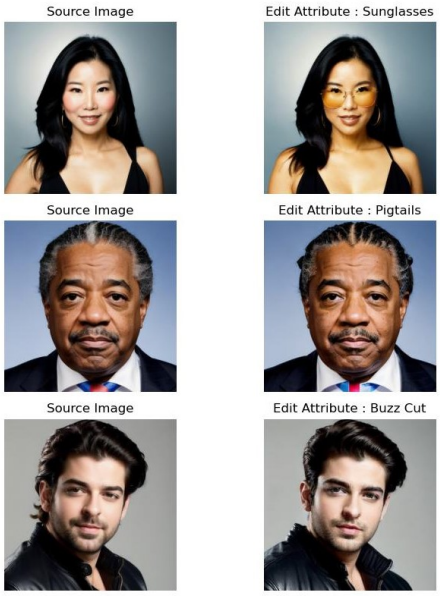


Figure B.6: Examples where SEGA's transformations are not perfect. First row: SEGA applies other edits apart from the requested edit (Change in skin tone). Second row : Edit hasn't been fully applied (Hairstyle changes, but not to text-specified hairstyle). Third row: A different edit has been applied (Beard trimmed and almost no change to hairstyle)

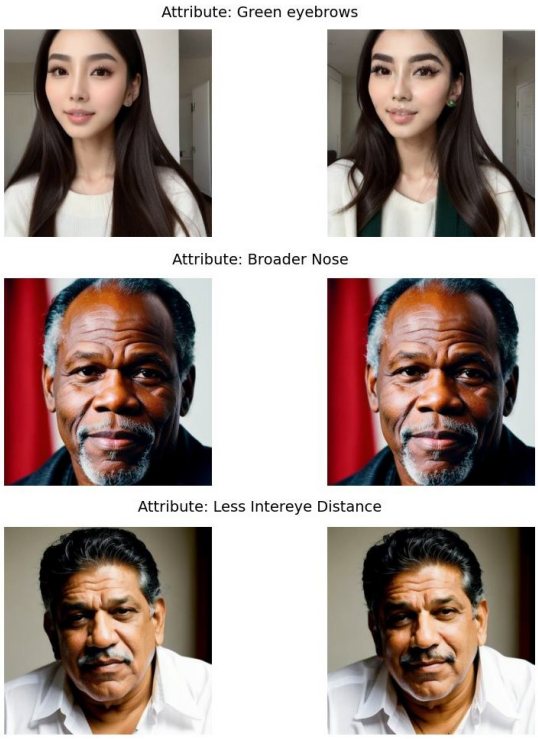


Figure B.7: SEGA and Diffusion models fail to semantically manipulate certain attributes.

Asian Woman	Asian Man	Indian Man	Indian Woman	Black Man	Black Woman	White Man	White Woman
Amr Moussa*	Ai Weiwei	Aamir Khan	Adah Sharma	Al Sharpton	Andra Day	Adam Sandler	Alanis Morissette
Amy Tan	Alberto Fujimori	Abhay Deol	Aditi Rao Hydari	Bill Cosby	Angela Bassett	Armie Hammer	Alison Brie
Angelababy	Avan Jogia*	Abhishek Bachchan	Alia Bhatt	Blair Underwood	Celia Cruz	Ashton Kutcher	Amy Poehler
Aretha Franklin*	BD Wong	Adil Hussain	Amrita Rao	Chris Brown	Ciara	Benedict Cumberbatch	Anna Faris
Awkwafina	Benedict Wong	Ajay Devgn	Anushka Sharma	Chris Rock	Danai Gurira	Blake Shelton	Arielle Kebbel
Beilei Li	Bob Menendez*	Akshaye Khanna	Anushka Shetty	Chris Tucker	Danielle Brooks	Bob Geldof	Bianca Jagger
Chloe Zhao	Bolo Yeung	Ali Zafar	Asin Thottumkal	Damon Dash	Dionne Warwick	Bon Jovi^	Carmen Electra
Chun Li	Bruce Lee	Amitabh Bachchan	Billy Bob Thornton*	Damon Wayans Jr	Erykah Badu	Brian Regan	Celine Dion
Coretta Scott King*	Charles Melton	Anil Kapoor	Bipasha Basu	Danny Glover	Eva Marcille	Clay Aiken	Claire Forlani
Devon Aoki	Cy Young	Arjun Kapoor	Daisy Shah	Deon Cole	Gabourey Sidibe	Darrell Issa	Claudia Cardinale
Diane Lane*	Daniel Dae Kim	Arjun Rampal	Deepika Padukone	Djimon Hounsou	Garcelle Beauvais	David Byrne	Doris Roberts
Elaine Chao	Dawn Staley*	Arshad Warsi	Divya Bharti	Donald Glover	Genevieve Nnaji	Denis Leary	Ellie Kemper
Elizabeth Dole*	Donnie Yen	Ayushmann Khurrana	Esha Gupta	Drake	Gladys Knight	Dennis Quaid	Evan Rachel Wood
Elizabeth Pena*	Fan Ho	BJ Habibie*	Evelyn Sharma	Eminem*	Herbie Hancock*	Derek Hough	Hilary Duff
Evo Morales*	Hyun Bin	Boman Irani	Huma Qureshi	Eric Benet	Janet Jackson	Geoffrey Rush	Hilary Swank
Faye Wong	IM Pei	Brian Cashman	Jacqueline Fernandez	Floyd Mayweather	Jennifer Hudson	Gilbert Gottfried	Ivana Trump
Gemma Chan	Iko Uwais	Craig David*	Juhi Chawla	Hank Aaron	Jessica Simpson*	Harvey Weinstein	JK Rowling
Gong Li	Jack Ma	Farhan Akhtar	Kalpana Chawla	Idris Elba	Kandi Burruss	James Cameron	Jaime Pressly
Gwen Stefani*	Jackie Chan	Girish Karnad	Kareena Kapoor	James Brown	Kenya Moore	Jason Aldean	Jane Fonda
Jamie Chung	Jerry Jones*	Hrithik Roshan	Karisma Kapoor	James Earl Jones	Kerry Washington	Jeff Bridges	Jennifer Lawrence
Jayne Meadows*	Jet Li	John Abraham	Katrina Kaif	Jamie Foxx	Laverne Cox	Jim Carrey	Jennifer Lopez
Jeri Ryan*	Jo Jung-suk	Johnny Hallyday*	Konkona Sen Sharma	Jesse Jackson	Leslie Jones	John Cornyn	Jordin Sparks
Joan Chen	John Williams*	Kamal Haasan	Kriti Sanon	Jon Stewart	Macy Gray	John Edwards	Kate McKinnon
Karen Mok	Johnny Chan	Kapil Sharma	Lara Dutta	Jussie Smollett	Mae Jemison	John Gruden	Kate Walsh
Kathy Bates*	Johnny Yong Bosch*	Leander Paes	Lisa Leslie*	Justin Gatlin	Mary J Blige	Johnny Carson	Katharine Hepburn
Keiko Sofia Fujimori	Justine Henin*	Madhavan	Lisa Murkowski*	Kevin Hart	Maya Rudolph*	Justin Bieber	Katie Couric
Kim Chiu	Kechun Li	Manoj Bajpayee	Mallika Sherawat	Kirk Franklin	Meagan Good	Keanu Reeves	Katie Holmes
Kim Jong-Il*	Ken Watanabe	Naseeruddin Shah	Manisha Koirala	Kobe Bryant	Mel B	Kelsey Grammer	Kelly Clarkson
Li-Xin Zhao*	Kim Soo-hyun	Nawazuddin Siddiqui	Nandita Das	Kofi Annan	Natalie Cole	Kris Kristofferson	Kelly Ripa
Li Na	Kris Wu	Neil Nitin Mukesh	Nargis Fakhri	Lil Nas X	NeNe Leakes	Lenny Kravitz	Kim Richards
Lisa Ling	Lee Byung-hun	Om Puri	Parineeti Chopra	Ludacris	Niecy Nash	Mark Wahlberg	Kourtney Kardashian
Liu Wen	Liu Xiaobo	Pankaj Tripathi	Prachi Desai	Marlon Wayans	Octavia Spencer	Matthew Perry	Lauren Graham
Lucy Liu	Manish Dayal*	Prabhu Deva	Preity G Zinta	Martin Lawrence	Rihanna	Michael Bloomberg	Lena Dunham
Maggie Cheung	Niu Tie	Prakash Raj	Priyanka Chopra^	Michael Jai White	Serena Williams	Michael Dell	Lindsay Lohan
Megawati Sukarnoputri	Prince Naruhito	Rahul Dev	Priyanka Chopra Jonas^	Michael Jordan	Sharon Leal	Michael Sheen	Mae West
Meijuan Xi	Rick Yune	Rajeev Khandelwal	Radhika Apte	Mike Tyson	Sophie Okonedo	Mick McCarthy	Melissa McCarthy
Meng Wanzhou	Ronny Chieng	Rajesh Khanna	Rani Mukerji	Nelson Mandela	Tamar Braxton	Neil Young	Michelle Bachelet
Michelle Kwan	Ross Butler*	Randeep Hooda	Rekha	Nick Cannon	Teyana Taylor	Nick Frost	Molly Sims
Michelle Yeoh	Ru Thing	Ranveer Singh	Richa Chadha	Pedro Martinez	Tia Mowry	Nick Jonas	Nancy Pelosi
Nastia Liukin*	Takeshi Kaneshiro	Rishi Kapoor	Shabana Azmi	Ray J	Tiffany Haddish	Richard Lugar	Naya Rivera
Nora Ephron*	Takuma Sato	Ronit Roy	Shilpa Shetty Kundra	Robert Mugabe	Tika Sumpter	Ricky Martin	Newt Gingrich
Patsy Mink	Thaksin Shinawatra	Sidharth Malhotra	Shraddha Kapoor	Ruben Studdard	Tyra Banks	Ridley Scott	Patricia Arquette
Paul Brandt*	Tony Clement*	Sonu Sood	Shruti Haasan	Samuel L Jackson	Uzo Aduba	Robbie Williams	Reba McEntire
Queen Latifah*	Tony Jaa	Suniel Shetty	Sonakshi Sinha	Sidney Poitier	Viola Davis	Robin Gibb	Ronda Rousey
Rafidah Aziz*	Tony Leung Chiu-wai	Tiger Shroff	Sonal Chauhan	Tavis Smiley	Vivica A Fox^	Ryan Phillippe	Shailene Woodley
Rosalind Chao	Xiaomi Le	Tusshar Kapoor	Sushmita Sen	Taye Diggs	Vivica Fox^	Ryan Seacrest	Sharon Osbourne
Shanshan Li	Xin Zhao	Varun Dhawan	Taapsee Pannu	Thabo Mbeki	Wanda Sykes	Sebastian Maniscalco	Steven Tyler*
Susan Collins*	Yao Ming	Vidyut Jammwal	Twinkle Khanna	Tyler Perry	Wendy Williams	Sting	Tamara Mowry*
Tara VanDerveer*	Yat-sen Sun	Vir Das	Vidya Balan	Venus Williams*	Yvette Nicole Brown	Vicente Fernandez	Taylor Swift
Yu Li	Zhang Yimou	Vivek Oberoi	Yami Gautam	William Harrison	Zoe Saldana	Vince Gill	Valerie Harper

Table B.2: Names chosen from CLIP retrieval for the eight demographics. As the process of assigning sex and gender was done using GPT 3.5, some of the names appear in the wrong groups. * indicates a name assigned to a wrong group and ^ indicates names referring to same individual

Asian Woman	Asian Man	Indian Man	Indian Woman	Black Man	Black Woman	White Man	White Woman
Rosalind Chao	Ai Weiwei	Suniel Shetty	Twinkle Khanna	Kevin Hart	Wendy Williams	Johnny Carson	Katharine Hepburn
Patsy Mink	BD Wong	Arshad Warsi	Juhi Chawla	Kofi Annan	Vivica A Fox	Kelsey Grammer	Patricia Arquette
Angelababy	Bruce Lee	Neil Nitin Mukesh	Rekha	Thabo Mbeki	Eva Marcille	Nick Jonas	Mae West
Michelle Kwan	Hyun Bin	Rajesh Khanna	Priyanka Chopra	Robert Mugabe	Tika Sumpter	Michael Sheen	Hilary Duff
Lisa Ling	Kim Soo-hyun	Om Puri	Aditi Rao Hydari	Djimon Hounsou	Gladys Knight	Clay Aiken	Taylor Swift
Shanshan Li	Kris Wu	Abhishek Bachchan	Bipasha Basu	Kirk Franklin	Sharon Leal	John Cornyn	Molly Sims
Gemma Chan	Benedict Wong	Abhay Deol	Mallika Sherawat	Danny Glover	Tamar Braxton	Mick McCarthy	Jennifer Lawrence
Lucy Liu	Ronny Chieng	John Abraham	Kareena Kapoor	Al Sharpton	Octavia Spencer	Ryan Seacrest	Anna Faris
Kim Chiu	Iko Uwais	Ajay Devgn	Preity G Zinta	Nelson Mandela	Gabourey Sidibe	Harvey Weinstein	Arielle Kebbel
Karen Mok	Alberto Fujimori	Kamal Haasan	Rani Mukerji	Idris Elba	Jennifer Hudson	Armie Hammer	Claire Forlani

Table B.3: Final 10 names for each demographic group

Appendix C

Chapter 4 Appendix

C.1 Proof of Main Theorem

This section contains our proof of theorem 4.8, with which we formulate general categorywise risk generalization bounds. Our technique also puts categorywise risk generalization bounds in terms of sample complexities for ERM learning. Recall the definition of categorywise sample complexity from definition 4.6. Let us begin by assuming a category-wise sample complexity exists for each category. Such a formulation allows use of any sort of concentration bound, including those based upon Rademacher complexity and VC dimension. Once we have a general concentration bound on categorywise risk, we leverage the frequency of each category to yield an overall generalization bound.

For our proof, we will also need the Relative Chernoff bound:

Lemma C.1 (Relative Chernoff Bound)**.** *Let $X_1, \ldots, X_m$ be i.i.d. random variables taking values in $\{0, 1\}$ and let X denote their sum. Furthermore, let $\epsilon \in (0, 1)$, then*

$$\mathbb{P}\left[X - (1-\epsilon)\,\mathbb{E}\left[X\right]\right] \leq \exp\left\{-\frac{1}{2}\epsilon^2\,\mathbb{E}\left[X\right]\right\} \tag{C.1}$$

The proof of theorem 4.8 follows:

Proof. Let $S \triangleq \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ be a sample of N labeled examples drawn i.i.d. from $\mathcal{D}$. Further, let $\mathcal{H}$ denote a fixed hypothesis class and assume $m_{\mathcal{H},\mathsf{g},\hat{y}}(\epsilon, \delta)$ exists for each category.

Lastly, let $N_{(\mathsf{g},\hat{y})}$ be the number of samples in S belonging to category $(\mathsf{g}, \hat{y})$.

Recall that γ is the lowest frequency amongst categories $(\mathsf{g}, \hat{y}) \subseteq \mathsf{G} \times \mathcal{Y}$: Let us assume $N \geq \frac{8 \log\left(\frac{2|\mathsf{G}||\mathcal{Y}|}{\delta}\right)}{\gamma}$. This assumption allows us to invoke lemma C.1, which tells that with probability $1 - \frac{\delta}{2}$, for every category $(\mathsf{g}, \hat{y})$:

$$N_{(\mathsf{g},\hat{y})} \geq \frac{\gamma N}{2} \tag{C.2}$$

Next, we show that, with high probability, having a large number of examples in each category $(\mathsf{g}, \hat{y})$, $N_{(\mathsf{g},\hat{y})}$ yields an accurate estimate of $\ell_{0\text{-}1}$ on each category $(\mathsf{g}, \hat{y})$. For this purpose, we define

$$\delta' \triangleq \frac{\delta}{2\,|\mathsf{G}|\,|\mathcal{Y}|} \tag{C.3}$$

Given the existence of $m_{\mathcal{H},\mathsf{g},\hat{y}}(\epsilon, \delta)$, we know for any hypothesis class $\mathcal{H}$, and any prediction value $v \in \mathcal{Y}$ and any $\mathsf{g} \in \mathsf{G}$, with probability at least $1-\delta'$, for a random sample of N examples from category $(\mathsf{g}, \hat{y})$ where

$$N_{(\mathsf{g},\hat{y})} = m_{\mathcal{H},\mathsf{g},\hat{y}}\left(\epsilon, \delta'\right) \tag{C.4}$$

$$= m_{\mathcal{H},\mathsf{g},\hat{y}}\left(\epsilon, \frac{\delta}{2\,|\mathsf{G}|\,|\mathcal{Y}|}\right) \tag{C.5}$$

then we have for each $\mathrm{h} \in \mathcal{H}$:

$$|\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h}, \mathsf{g}, \hat{y}) - \mathcal{L}_{S,\ell}(\mathrm{h}, \mathsf{g}, \hat{y})| \leq \epsilon \tag{C.6}$$

Set $N^\dagger$ such that it upper-bounds all $N_{(\mathsf{g},\hat{y})}$ terms:

$$N^\dagger = \max_{\mathsf{g}\in\mathsf{G},\hat{y}\in\mathcal{Y}} m_{\mathcal{H},\mathsf{g},\hat{y}}\left(\epsilon, \frac{\delta}{2\left|\mathsf{G}\right|\left|\mathcal{Y}\right|}\right) \tag{C.7}$$

Then, if for all categories, $(\mathsf{g},\hat{y})$, we have $N_{(\mathsf{g},\hat{y})} \geq N$, by union bound, we have the following with probability at least $1 - \left|\mathsf{G}\right|\left|\mathcal{Y}\right|\delta' = 1 - \frac{\delta}{2}$:

$$\begin{aligned} &\forall \mathrm{h} \in \mathcal{H}, \forall \mathsf{g} \in \mathsf{G}, \forall v \in \mathcal{Y} : \\ &\left|\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h},\mathsf{g},\hat{y}) - \mathcal{L}_{S,\ell}(\mathrm{h},\mathsf{g},\hat{y})\right| \leq \epsilon \end{aligned} \tag{C.8}$$

Let us now choose the following overall sample size :

$$N \triangleq \max_{\mathsf{g}\in\mathsf{G},\hat{y}\in\mathcal{Y}} \frac{2}{\gamma} m_{\mathcal{H},\mathsf{g},\hat{y}}\left(\epsilon, \frac{\delta}{2\left|\mathsf{G}\right|\left|\mathcal{Y}\right|}\right) \tag{C.9}$$

Recall our use of lemma C.1 at the beginning of this proof. Given a sample of size at least N, we know with probability at least $1 - \frac{\delta}{2}$, for every category $(\mathsf{g},\hat{y}) \subseteq \mathsf{G}\times\mathcal{Y}$:

$$N_{(\mathsf{g},\hat{y})} \geq \frac{\gamma N}{2} = N^\dagger \tag{C.10}$$

Combined with our knowledge from (C.8), the assumption in (C.9) allows us to invoke the union bound, such that we have, with probability at least $1-\delta$:

$$\begin{aligned} &\forall \mathrm{h} \in \mathcal{H}, \forall \mathsf{g} \in \mathsf{G}, \forall v \in \mathcal{Y} : \\ &\left|\mathcal{L}_{\mathcal{D},\ell}(\mathrm{h},\mathsf{g},\hat{y}) - \mathcal{L}_{S,\ell}(\mathrm{h},\mathsf{g},\hat{y})\right| \leq \epsilon \end{aligned} \tag{C.11}$$

Thereby concluding the proof.

C.2 Two-sided Rademacher Complexity Bounds

In this section, we discuss the motivation behind the one-sided Rademacher complexity based generalization bounds of Shalev-Shwartz and Ben-David [138], and we show how to transform these bounds into two-sided generalization bounds. Notation and proof techniques in this section largely follow that of Shalev-Shwartz and Ben-David. Using the Rademacher random variable, Bartlett et al. [130] were able to develop a notion of expressiveness for a given hypothesis class. This measure of expressiveness is known as the Rademacher Complexity:

Definition C.2 (Rademacher Complexity)**.** *The Rademacher complexity of a hypothesis class $\mathcal{H}$ with respect to a sample S is*

$$\mathcal{R}\left(\mathcal{H} \circ S\right) \triangleq \frac{1}{N} \mathop{\mathbb{E}}_{\boldsymbol{\sigma} \sim \{\pm 1\}} \left[\sup_{\mathrm{h} \in \mathcal{H}} \sum_{i=1}^{N} \sigma_i \mathrm{h}(\boldsymbol{z}_i) \right] \tag{C.12}$$

where sample $S \triangleq \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$ is a set of of N i.i.d. samples drawn from distribution $\mathcal{D}$. and $\mathcal{H} \circ S$ denotes the set of all possible evaluations a hypothesis $\mathrm{h} \in \mathcal{H}$ can achieve on a sample S, i.e.

$$\mathcal{H} \circ S \triangleq \{(\mathrm{h}(\boldsymbol{z}_1), \ldots, \mathrm{h}(\boldsymbol{z}_N)) : \mathrm{h} \in \mathcal{H}\} \tag{C.13}$$

The Rademacher complexity (definition C.2) is a useful notion of hypothesis class expressiveness because it can be used to bound the *representativeness* of a sample S with respect to $\ell \circ \mathcal{H}$ as the largest gap between the true risk of a hypothesis h and its empirical risk.

Definition C.3. *The* representativeness *of S, a sample drawn from distribution $\mathcal{D}$, with respect to $\mathcal{H}$, denoted $\mathrm{Rep}_{\mathcal{D}}(\ell \circ \mathcal{H}, S)$, is the largest gap between the true risk of a hypothesis* h *in hypothesis class $\mathcal{H}$ and its empirical risk.*

$$\mathrm{Rep}_{\mathcal{D}}(\ell \circ \mathcal{H}, S) \triangleq \sup_{\mathrm{h} \in \mathcal{H}} \left(L_{\mathcal{D}}(\mathrm{h}) - L_S(\mathrm{h})\right) \tag{C.14}$$

The following lemma, from Shalev-Shwartz and Ben-David describes the relationship between expected representativeness and expected Rademacher complexity:

Lemma C.4 (Lemma 26.2 from Shalev-Shwartz and Ben-David [138])**.**

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}} \mathrm{Rep}_{\mathcal{D}}(\ell \circ \mathcal{H}, S) \leq 2 \mathop{\mathbb{E}}_{S\sim\mathcal{D}} \mathcal{R}_S\left(\ell \circ \mathcal{H}\right) \tag{C.15}$$

Inspired by this bound, it is possible to construct generalization bounds based on empirical Rademacher complexity:

Theorem C.5 (Theorem 26.5 from Shalev-Shwartz and Ben-David [138])**.** *Let $S \sim \mathcal{D}^N$ be an i.i.d. sample of size N. Assume that for all $(\boldsymbol{x}, y) \in S$ and $\mathrm{h} \in \mathcal{H}$, we have that $|\ell(\mathrm{h}(\boldsymbol{x}), y)| \leq c$. Then with probability of at least $1-\delta$, for all $\mathrm{h} \in \mathcal{H}$:*

$$L_{\mathcal{D}}(\mathrm{h}) - L_S(\mathrm{h}) \leq 2 \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^N} \mathcal{R}_{S'}(\ell \circ \mathcal{H}) + c\sqrt{\frac{2\ln\left(2/\delta\right)}{N}} \tag{C.16}$$

$$L_{\mathcal{D}}(\mathrm{h}) - L_S(\mathrm{h}) \leq 2\mathcal{R}_S(\ell \circ \mathcal{H}) + 4c\sqrt{\frac{2\ln\left(4/\delta\right)}{N}} \tag{C.17}$$

Note that theorem C.5 contains only one-sided bounds on $L_{\mathcal{D}}(\mathrm{h}) - L_S(\mathrm{h})$, yet we would like two-sided bounds, i.e. bounds on the probability $|L_{\mathcal{D}}(\mathrm{h}) - L_S(\mathrm{h})|$ is small. Let us now pursue such bounds $|L_{\mathcal{D}}(\mathrm{h}) - L_S(\mathrm{h})|$ using the Rademacher complexity. Our proof follows the style of the proof of lemma 26.2 from Shalev-Shwartz and Ben-David [138].

Lemma C.6 (Two-sided version of lemma C.4)**.**

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^N} \sup_{\mathrm{h}\in\mathcal{H}} |L_{\mathcal{D}}(\mathrm{h}) - L_S(\mathrm{h})| \leq 2 \mathop{\mathbb{E}}_{S\sim\mathcal{D}^N} \mathcal{R}_S\left(\ell \circ \mathcal{H}\right) \tag{C.18}$$

Proof. Let us now provide a proof of lemma C.6:

$$\sup_{\mathrm{h}\in\mathcal{H}} |L_{\mathcal{D}}(\mathrm{h}) - L_S(\mathrm{h})| = \sup_{\mathrm{h}\in\mathcal{H}} \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^N} |L_{S'}(\mathrm{h}) - L_S(\mathrm{h})| \tag{C.19}$$

Because expectation of supremum is larger than supremum of expectation, we have:

$$\sup_{\mathrm{h}\in\mathcal{H}} |L_{\mathcal{D}}(\mathrm{h}) - L_S(\mathrm{h})| \leq \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^N} \sup_{\mathrm{h}\in\mathcal{H}} |L_{S'}(\mathrm{h}) - L_S(\mathrm{h})| \tag{C.20}$$

Taking expectation over S on both sides, we have

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^N} \left[\sup_{\mathrm{h}\in\mathcal{H}} |L_{\mathcal{D}}(\mathrm{h}) - L_S(\mathrm{h})|\right] \leq \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^N} \left[\sup_{\mathrm{h}\in\mathcal{H}} |L_{S'}(\mathrm{h}) - L_S(\mathrm{h})|\right] \tag{C.21}$$

$$= \frac{1}{N} \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^N} \left[\sup_{\mathrm{h}\in\mathcal{H}} \left|\sum_{i=1}^{N} \ell(\mathrm{h}(\boldsymbol{x}'_i), y'_i) - \ell(\mathrm{h}(\boldsymbol{x}_i), y_i)\right|\right] \tag{C.22}$$

For each j, $(\boldsymbol{x}'_j, y'_j)$ and $(\boldsymbol{x}_j, y_j)$ are i.i.d. variables, which allows us to make the following statement:

$$\underbrace{\frac{1}{N} \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^N} \left[\sup_{\mathrm{h}\in\mathcal{H}} \left|\ell(\mathrm{h}(\boldsymbol{x}'_j), y'_j) - \ell(\mathrm{h}(\boldsymbol{x}_j), y_j) + \sum_{i\neq j} \ell(\mathrm{h}(\boldsymbol{x}'_i), y'_i) - \ell(\mathrm{h}(\boldsymbol{x}_i), y_i)\right|\right]}_{A} \tag{C.23}$$

$$= \underbrace{\frac{1}{N} \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^N} \left[\sup_{\mathrm{h}\in\mathcal{H}} \left|\ell(\mathrm{h}(\boldsymbol{x}_j), y_j) - \ell(\mathrm{h}(\boldsymbol{x}'_j), y'_j) + \sum_{i\neq j} \ell(\mathrm{h}(\boldsymbol{x}'_i), y'_i) - \ell(\mathrm{h}(\boldsymbol{x}_i), y_i)\right|\right]}_{B} \tag{C.24}$$

Let σ_j be a Rademacher random variable (i.e. a draw from the uniform distribution over $\{+1, -1\}$). Then we have:

$$\frac{1}{N} \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^N,\sigma_j} \left[\sup_{\mathrm{h}\in\mathcal{H}} \left|\sigma_j(\ell(\mathrm{h}(\boldsymbol{x}'_j), y'_j) - \ell(\mathrm{h}(\boldsymbol{x}_j), y_j)) + \sum_{i\neq j} \ell(\mathrm{h}(\boldsymbol{x}'_i), y'_i) - \ell(\mathrm{h}(\boldsymbol{x}_i), y_i)\right|\right] \tag{C.25}$$

$$= \frac{1}{2}(A + B)$$

$$= \frac{1}{N} \mathop{\mathbb{E}}_{S,S' \sim \mathcal{D}^N} \left[\sup_{\mathrm{h} \in \mathcal{H}} |\ell(\mathrm{h}(\boldsymbol{x}_i'), y_i') - \ell(\mathrm{h}(\boldsymbol{x}_i), y_i)| \right] \tag{C.26}$$

Let $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_N]^\mathsf{T}$ be a vector consisting of N i.i.d. samples from the Rademacher distribution. Repeating this procedure for all j, we have

$$\begin{aligned} &\frac{1}{N} \mathop{\mathbb{E}}_{S,S' \sim \mathcal{D}^N} \left[\sup_{\mathrm{h} \in \mathcal{H}} \left| \sum_{i=1}^{N} \ell(\mathrm{h}(\boldsymbol{x}_i'), y_i') - \ell(\mathrm{h}(\boldsymbol{x}_i), y_i) \right| \right] \\ &= \frac{1}{N} \mathop{\mathbb{E}}_{S,S' \sim \mathcal{D}^N, \boldsymbol{\sigma}} \left[\sup_{\mathrm{h} \in \mathcal{H}} \left| \sum_{i=1}^{N} \sigma_i (\ell(\mathrm{h}(\boldsymbol{x}_i'), y_i') - \ell(\mathrm{h}(\boldsymbol{x}_i), y_i) \right| \right] \end{aligned} \tag{C.27}$$

$$= \frac{1}{N} \mathop{\mathbb{E}}_{S,S' \sim \mathcal{D}^N, \boldsymbol{\sigma}} \left[\sup_{\mathrm{h} \in \mathcal{H}} \left(\sum_{i=1}^{N} \sigma_i \ell(\mathrm{h}(\boldsymbol{x}_i'), y_i') \right) - \inf_{\mathrm{h} \in \mathcal{H}} \left(\sum_{i=1}^{N} \sigma_i \ell(\mathrm{h}(\boldsymbol{x}_i), y_i) \right) \right] \tag{C.28}$$

We know the distribution of $\sum_{i=1}^{N} \sigma_i \ell(\mathrm{h}(\boldsymbol{x}_i), y_i)$ to be symmetric, hence we have

$$= \frac{1}{N} \mathop{\mathbb{E}}_{S,S' \sim \mathcal{D}^N, \boldsymbol{\sigma}} \left[\sup_{\mathrm{h} \in \mathcal{H}} \left(\sum_{i=1}^{N} \sigma_i \ell(\mathrm{h}(\boldsymbol{x}_i'), y_i') \right) + \sup_{\mathrm{h} \in \mathcal{H}} \left(\sum_{i=1}^{N} \sigma_i \ell(\mathrm{h}(\boldsymbol{x}_i), y_i) \right) \right] \tag{C.29}$$

$$= \frac{2}{N} \mathop{\mathbb{E}}_{S \sim \mathcal{D}^N, \boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{H}} \left(\sum_{i=1}^{N} \sigma_i \ell(\mathrm{h}(\boldsymbol{x}_i), y_i) \right) \right] \tag{C.30}$$

$$= 2 \mathop{\mathbb{E}}_{S \sim \mathcal{D}^N, \boldsymbol{\sigma}} [\mathcal{R}_S(\ell \circ \mathcal{H})] \tag{C.31}$$

Upon replacing lemma C.4 with lemma C.6, theorem C.7 naturally follows from the proof of theorem C.5. Hence, we arrive at the following two-sided bound:

Theorem C.7 (Two-sided version of Theorem 26.5 from Shalev-Shwartz and Ben-David [138])**.** *Let $S \sim \mathcal{D}^N$ be a sample of size N. Assume that for all $(\boldsymbol{x}, y) \in S$ and $\mathrm{h} \in \mathcal{H}$, we have that $|\ell(\mathrm{h}(\boldsymbol{x}), y)| \leq c$. Then with probability of at least $1 - \delta$, for all $\mathrm{h} \in \mathcal{H}$:*

$$|L_{\mathcal{D}}(\mathrm{h}) - L_S(\mathrm{h})| \leq 2 \mathop{\mathbb{E}}_{S' \sim \mathcal{D}^N} \mathcal{R}_{S'}(\ell \circ \mathcal{H}) + c \sqrt{\frac{2 \ln (2/\delta)}{N}} \tag{C.32}$$

$$|L_{\mathcal{D}}(\mathrm{h}) - L_S(\mathrm{h})| \leq 2 \mathcal{R}_S(\ell \circ \mathcal{H} \circ S) + 4c \sqrt{\frac{2 \ln (4/\delta)}{N}} \tag{C.33}$$

C.3 Bounds, Theorems, and Algebra

Theorem C.8 (Simplified version of theorem 2 from Cortes et al. [157]). *Let K be a kernel function and that $K(\boldsymbol{x}, \boldsymbol{x}) \leq B^2$ for all $\boldsymbol{x} \in \mathcal{X}$. Then for any sample S of size N, the Rademacher complexity of the hypothesis class $\mathcal{H}'_K$ can be bounded as follows*

$$\mathcal{R}_S(\mathcal{H}_K) \leq \sqrt{\frac{23eB^2}{22N}} \tag{C.34}$$

Lemma C.9 (RBF Kernel SVM Sample Complexity Bound). *Let $\mathcal{H}_K$ be an RBF kernel SVM predictor class. Let $v \in \mathcal{Y}$ be a prediction value. Further, let risk be bounded between 0 and 1. Then for any distribution $\mathcal{D}$ over $\mathcal{X} \times \{\pm 1\}$ and any $\epsilon, \delta \in (0, 1)$, if S is a random sample of at least N samples where*

$$N \geq \frac{23B^2 + 64 \log(4/\delta)}{\epsilon^2} \tag{C.35}$$

examples drawn i.i.d. according to $\mathcal{D}$, then with probability at least $1 - \delta$:

$$\forall \mathrm{h} \in \mathcal{H} : |\mathcal{L}_{\mathcal{D}}(\mathrm{h}) - \mathcal{L}_S(\mathrm{h})| \leq \epsilon \tag{C.36}$$

Proof. Fix $\mathrm{h}, \mathsf{g}, \hat{y}$ and S.

Invoking theorem C.7 over a bounded loss function, and solving for N:

$$|\mathcal{L}_{\mathcal{D}}(\mathrm{h}) - \mathcal{L}_S(\mathrm{h})| \leq 2\mathcal{R}_S(\hat{\ell} \circ \mathcal{H}_K) + 4c\sqrt{\frac{2 \log(4/\delta)}{N}} \tag{C.37}$$

Because risk is bounded between 0 and 1, we set $c = 1$, to yield

$$|\mathcal{L}_{\mathcal{D}}(\mathrm{h}) - \mathcal{L}_S(\mathrm{h})| \leq 23\frac{B^2}{N} + 64\frac{\log(4/\delta)}{N} \tag{C.38}$$

To guarantee this is smaller the right hand side of the inequality is less than ϵ, we need

$$N \geq \frac{23B^2 + 64 \log(4/\delta)}{\epsilon^2} \tag{C.39}$$

Thus arriving at the sample complexity bound.

C.3.1 VC Dimension Definition

Definition C.10 (VC dimension)**.** *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a binary hypothesis class. A hypothesis class $\mathcal{H}$ shatters a set $\mathcal{V} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \mathcal{X}$ if*

$$|\{\mathrm{h}(\boldsymbol{x}_1), \ldots, \mathrm{h}(\boldsymbol{x}_N)\} : \mathrm{h} \in \mathcal{H}| = 2^N \tag{C.40}$$

The VC dimension of $\mathcal{H}$ is the maximal size of $\mathcal{V}$ which is shattered by $\mathcal{H}$.