Margin-distancing for safe model explanation

Tom Yan
Carnegie Mellon University

Chicheng Zhang
University of Arizona

Abstract

The growing use of machine learning models in consequential settings has highlighted an important and seemingly irreconcilable tension between transparency and vulnerability to gaming. While this has sparked sizable debate in legal literature, there has been comparatively less technical study of this contention. In this work, we propose a clean-cut formulation of this tension and a way to make the tradeoff between transparency and gaming. We identify the source of gaming as being points close to the *decision boundary* of the model. And we initiate an investigation on how to provide example-based explanations that are expansive and yet consistent with a version space that is sufficiently uncertain with respect to the boundary points' labels. Finally, we furnish our theoretical results with empirical investigations of this tradeoff on real-world datasets.

1 INTRODUCTION

With the increasing use of machine learning models in automating decision making, there is growing concern over the opacity of these models. Such concerns have given rise to laws, such as the European GDPR, which aim to provide a "Right to Explanation"(Wachter et al., 2017; Edwards and Veale, 2017; Selbst and Powles, 2018). However, one stumbling block to this solution is the tension between transparency and gaming: greater transparency into the model gives rise to gaming – individuals strategically misreporting their features to induce desired classification outcomes from the ML model.

As a result, some government agencies are still to this day reluctant about revealing details on the deployed algorithms. This has in turn lead to Freedom of Information requests, such as those submitted by civil interest groups in the Netherlands, calling for greater transparency (Wieringa, 2020), as well as organized movements such as the OpenSCHUFA project (OpenSCHUFA, 2019), through which citizens take matters in their own hands and try to crowd-source data in an effort to reverse-engineer the algorithms.

In this work, we formalize this tension in a natural, formal model, which to the best of our knowledge, is the *first formal model* capturing the tradeoff between transparency and gaming in machine learning.

The setting we will study is one where an organization uses model $h^* : \mathcal{X} \to \{-1, +1\}$ to perform classification over feature space $\mathcal{X}$ and provides transparency through model explanations. We focus on example-based explanations $\mathcal{E}_{h^*}$, which have been found to be one of the most intuitive types of explanations in a recent human study (Jeyakumar et al., 2020), and in particular on prototype-based explanations (e.g k-medoid or MMD-critic (Kim et al., 2016)).

In more detail, the explanation mechanism $\mathcal{E}_{h^*} : \mathcal{X} \to 2^{\mathcal{X}}$ will select a representative subset of $\mathcal{X}$ to label and explanations $\{(x, h^*(x)) \mid x \in \mathcal{E}_{h^*}(\mathcal{X})\}$ will be released. For example, for loan applications, such explanation could be in the form of past, anonymized (un)successful profiles.

Intuitively, the concern with releasing explanations is that applicants may use the knowledge of the hypothesis class $\mathcal{H} \ni h^*$ along with the explanations to construct the version space (VS), $\mathcal{H}_C = \{h \in \mathcal{H} \mid h(x) = h^*(x), \forall x \in \mathcal{E}_{h^*}(\mathcal{X})\}$, to infer h^*. If the explanation is "good" and allows for "simulatability" of h^* (Murdoch et al., 2019), then the few models in $\mathcal{H}_C$ would be constrained by the explanations to have very similar predictions on $\mathcal{X}$ as h^*. And so, even though the VS does not *directly* identify h^*, the VS allows one to estimate h^*'s prediction with high certainty. This we will be formalize soon.

To address this issue, we propose *margin-distancing* as a simple and general method that can make the

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

tradeoff between transparency and gaming. We show that with margin-distancing it need not be one or the other: it is possible to offer individuals *some idea* of how the model works while still preventing gaming.

Concretely, given classification models h^* and input example x, we use $f^* : \mathcal{X} \to \mathbb{R}$ to denote a function that outputs an underlying margin score, $h^*(x) = \text{sign}(f^*(x))$, where $\text{sign}(a) = +1$ for $a \geq 0$ and $\text{sign}(a) = -1$ otherwise. Margin-distancing selects a subset of $\mathcal{X}$ whose margin score $\left|f^*(x)\right|$ is greater than some threshold α. This is done to induce a sufficiently large $\mathcal{H}_C$ and, as a result, sufficiently low certainty on how h^* predicts to dissuade gaming.

This approach is compatible with any example-based explanations. We note that our approach is also applicable with local surrogate based methods with bounded fidelity region. Indeed, these methods may be viewed as example-based explanation methods that impart labels for all points within the fidelity regions.

Our Contributions:

(1) We formalize the tradeoff between transparency and gaming, and propose *margin-distancing* as a way of making this tradeoff.

(2) We prove that margin-distancing does *monotonically* decreases decision boundary certainty under a uniform prior over homogeneous linear models and spherical feature space. We also give a set of complementary negative results showing that monotonicity does not hold in general.

(3) We evaluate boundary points' certainty using sampling for general model classes. Our empirical studies suggest margin-distancing does reduce boundary certainty in a relatively monotonic fashion, and in some cases, completely monotonically, which would enable binary search as a computationally efficient means of finding the optimal amount of explanations to release.

2 RELATED WORKS

Transparency vs Gaming: To the best of our knowledge, there has been only one technical paper (Tsirtsis and Gomez-Rodriguez, 2020) that examines the tension between explanation and gaming. In this work, an organization focuses on releasing an optimal set of counterfactual explanations S to induce agents to change their reports in a way that maximizes the organization's utility; this work does not focus on examining the tradeoff explored in our paper. Moreover, the key assumption that differs from our setting is that all feature alteration is viewed as being causal. Lastly, in our work, we do not assume that agents can only change to points in S (if possible), but rather to any point $\hat{x}$ in the neighborhood of x.

Strategic ML: Similar to most of strategic classification literature (Hardt et al., 2016; Dong et al., 2018; Kleinberg and Raghavan, 2020; Chen et al., 2018b), we assume strategic behavior is gaming. However, different from most, past formulations, agents in our setting do not have *full knowledge* of h^* and have to best respond with only partial knowledge (explanations) of h^*.

In the interest of space, we have included further related works on topics including Improvement vs Gaming, Explanation Manipulation in Appendix D.

3 PROBLEM FORMULATION

Gaming: We assume all individuals desire to be classified the positive label (e.g "loan granted") by h^*. An individual with profile x may use the explanations of h^* to compute and misreport $\hat{x} \neq x$ so as to improve the chance of being classified as the positive label. As is standard in strategic classification, this act of misreporting is referred to as *gaming* (Hardt et al., 2016).

In face of gaming, the organization wishes to have its predictions be unaffected by the release of explanations $\mathcal{E}_{h^*}(\mathcal{X})$: $h^*(\hat{x}) = h^*(x)$, $\forall x \in \mathcal{X}$.

For our analysis, we first assume that applicants cannot report arbitrary profiles – otherwise everyone will simply report some $x \in \mathcal{E}_{h^*}(\mathcal{X})$ with a positive label. This assumption may also be motivated as follows: in strategic ML literature, individuals are typically assumed to have a cost function. This naturally induces a region beyond which it is too costly to change to. For modeling purposes, we assume that if an applicant has feature x, then $\hat{x} \in \mathcal{R}_r(x) := \{x' \mid \|x - x'\| < r, x' \in \mathcal{X}\}$, with $r > 0$ being the maximum extent of manipulation. Additionally, we assume that applicants are aware of the model class $\mathcal{H} \ni h^*$ used by the organization.

Next, since the explanations only allow one to conclude that $h^* \in \mathcal{H}_C$, we need to specify how individuals reason about whether to misreport x' or report x truthfully with only *partial knowledge* about h^*. To model this calculus, as is common in Economics, we assume that the individual is Bayesian and calculates the *increased* chance of obtaining positive label under x' instead of x through a prior distribution $\mathcal{U}$ that gets updated to posterior $\mathcal{U}(\mathcal{H}_C)$ (the restriction of $\mathcal{U}$ on the set $\mathcal{H}_C$) with knowledge of $\mathcal{E}_{h^*}(\mathcal{X})$:

$$\pi(x, x') = \Pr_{h \sim \mathcal{U}(\mathcal{H}_C)}(h(x') = 1) - \Pr_{h \sim \mathcal{U}(\mathcal{H}_C)}(h(x) = 1).$$

A natural choice for $\mathcal{U}$ is the uniform distribution, though it need not be so. We assume that the organi-

zation also knows $\mathcal{U}$.

Naturally, individuals will choose to misreport if there is a sufficiently high certainty of success, since they obtain positive utility for getting the positive label (i.e if h^* is s.t $h^*(\hat{x}) = 1$). However, in misreporting, they incur negative utility for the cost of manipulation: $x \to x'$. These two may be weighted linearly in rational agents or nonlinearly in behavioral agents due to risk-aversion (Kahneman and Tversky, 2013). Following the formal model of the rationality of crime as introduced by Becker (Becker, 1968), we abstract this away by assuming that there is some threshold κ such that if $\pi(x, x') \leq \kappa$, the individual is too risk-averse to misreport $\hat{x} = x'$: the cost of manipulation offsets the increased likelihood of obtaining positive utility through positive classification.

This brings us to our main insight: we only need $\mathcal{H}_C$ to be sufficiently ambiguous near the decision boundary because *only* individuals with points near the boundary can misreport in a way that flips h^*'s prediction.

Formally, define the set of *boundary points* to be all x's where such a label flip is possible: $\mathcal{N}_r(\mathcal{X}) := \{x \in \mathcal{X} \mid \exists x' \in \mathcal{R}_r(x) \wedge h^*(x') \neq h^*(x)\}$. Similarly, we define *boundary pairs* to be pairs (x, x') that are within a distance of r, but predicted differently by h^*; formally, $\mathcal{M}_r(\mathcal{X}) := \big\{(x, x') \in \mathcal{X}^2 \mid x' \in \mathcal{R}_r(x) \wedge h^*(x') \neq h^*(x)\big\}$. Observe that $\mathcal{M}_r(\mathcal{X}) \subset \mathcal{N}_r(\mathcal{X})^2$.

Margin-distancing: To make it difficult to infer the decision boundary through $\mathcal{H}_C$, it is natural to remove explanations that are close to the decision boundary. This gives rise to our approach of *margin-distancing*. We will designate some indicator function Λ_α for choosing explanations, which evaluates to 1 iff the examples' classification margin score is greater than cutoff α; formally, $\mathcal{E}_{h^*}(\mathcal{X}, \alpha) = \big\{x \in \mathcal{X} : \Lambda_\alpha(x) = 1\big\}$. Note that $\mathcal{H}_C$ is a function of α, since $\mathcal{H}_C$ is a function of the explanations, which are in turn a function of α. Intuitively, a big α that only retains explanations with large margins would decrease *boundary certainty*, which we define as $\max_{(x,x') \in \mathcal{M}_r(\mathcal{X})} \pi(x, x')$.

Policy Goals: Herein lies the tradeoff for the organization:

1) Provide explanation $\mathcal{E}_{h^*}(\mathcal{X}, \alpha)$ such that the boundary certainty is made sufficiently low: $\max_{(x,x') \in \mathcal{M}_r(\mathcal{X})} \pi(x, x') \leq \kappa$. This makes all individuals $x \in \mathcal{X}$ too risk-averse to misreport $\hat{x} \in \mathcal{R}_r(x)$ with $h^*(\hat{x}) \neq h^*(x)$, thus preventing gaming.

2) The explanation provided $\mathcal{E}_{h^*}(\mathcal{X}, \alpha)$ is as transparent as possible. That is, α is as small as possible to retain as many explanations from the full set of explanations as possible. Naturally, in our setting, we define transparency to be the amount of explanations that remain after margin-distancing.

The technical problem we study is:

> How can we search for the smallest threshold α possible such that $\max_{(x,x') \in \mathcal{M}_r(\mathcal{X})} \pi(x, x') \leq \kappa$, which is needed to prevent gaming?

Before we proceed, we obtain some intuition first through a qualitative visualization of $\mathcal{H}_C$ in a toy example, Figure 1. This figure helps to confirm that allowing explanations with small margins "boxes in" the version space too much, and makes models in $\mathcal{H}_C$ too similar to h^*. And so, removing explanations with small margin help enlarge $\mathcal{H}_C$ and decrease boundary-certainty.

Simple Example: Next, for a quantitative toy example, consider when $\mathcal{X} = [0, 1]$ and $\mathcal{H} = \{h_w(x) := \text{sign}(x - w) \mid w \in [0, 1]\}$ is the class of 1D thresholds. Let $\mathcal{U}$ be the uniform distribution over $\mathcal{H}$. We know then that $w^* \in [x^-, x^+]$, where x^- is the largest negative point in $\mathcal{E}_{h^*}(\mathcal{X})$ and x^+ the smallest positive point. Therefore, for $x \in (x^-, x^+)$ and some $x' \in \mathcal{R}_r(x) > x$, we have that $\pi(x, x') = \frac{\min\{x', x^+\} - x}{x^+ - x^-}$. In this case, it is evident that margin-distancing (i.e increasing x^+ and decreasing x^-) decreases boundary certainty $\pi(x, x')$.

In the section that follow, we study a more general hypothesis class and verify that the intuitive trend of removing information around the decision boundary does make it more difficult to infer the decision boundary, thus reducing boundary certainty.

4 HOMOGENEOUS LINEAR MODELS

We focus our theoretical study on the property of *monotonicity*, which if true, allows for binary search as an efficient way to compute the optimal α. In this section, we identify homogeneous linear models in $\mathbb{R}^d$, i.e. $\mathcal{H} = \big\{h_w \mid \|w\|_2 = 1\big\}$ (where $h_w := x \mapsto \text{sign}(\langle w, x \rangle)$), as one setting where margin-distancing monotonically leads to decreased boundary certainty.

For the results that follow, we also assume that individuals have uniform prior $\mathcal{U}$ over $\mathcal{H}$. We will also focus on when the feature space $\mathcal{X}$ is the origin-centered unit sphere, i.e., $\mathcal{X} = \big\{x \in \mathbb{R}^d \mid \|x\|_2 = 1\big\}$, which means that $r \leq 2$. Intuitively, this corresponds to a normalized dataset with profiles of "all kinds", which is not unreasonable for profiles of a general population. We handle more general settings in the following section.

For linear models, it is natural to take Λ to be a function of the margin of a point with respect to w^* (the param-

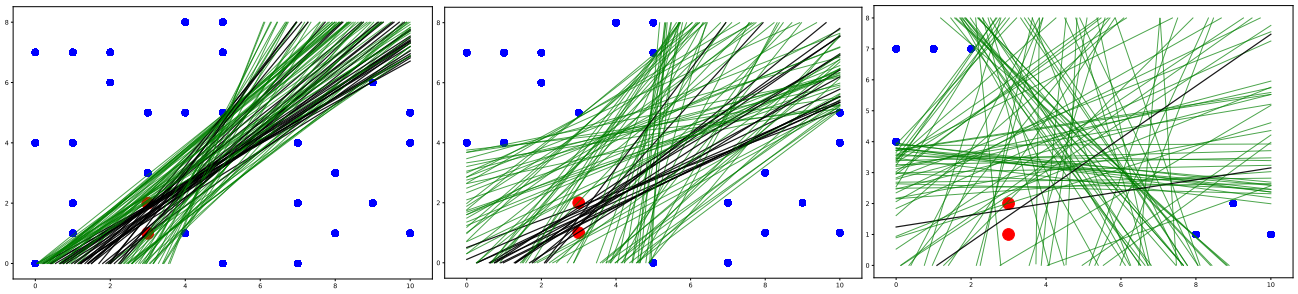

Figure 1: Visualization of $\mathcal{H}_C$ in a toy example where the amount of explanations (blue points) is varied ($80, 50, 20$ percent of all explanations is kept). In red is one randomly chosen boundary pair. 100 lines (green and black) are randomly sampled from $\mathcal{H}_C$; in black are lines that predict the pair like h^* (opposite labels), and green the same.

eter of h^*): $\Lambda_\alpha(x) = \mathbb{1}\{|\langle w^*, x\rangle| > \alpha\}$, for $\alpha \in [0, 1)$. Therefore, for every α, its associated set of explanations is $\mathcal{E}_{h^*}(\mathcal{X}, \alpha) = \left\{x \in \mathcal{X} : |\langle w^*, x\rangle| > \alpha\right\}$.

Under this "nice" setting, we first show that we can give a simple characterization of the version space in terms of α:

Lemma 1. *Fix $\alpha \in [0, 1)$. Recall that $\mathcal{H}_C = \{h \in \mathcal{H} \mid h(x') = h^*(x'),\ \forall x' \in \mathcal{E}_{h^*}(\mathcal{X}, \alpha)\}$ is the version space induced by explanation $\mathcal{E}_{h^*}(\mathcal{X}, \alpha)$. $\mathcal{H}_C$ can be equivalently written as:*

$$\mathcal{H}_C = \left\{h_w \mid \|w\|_2 = 1, w \cdot w^* \geq \sqrt{1-\alpha^2}\right\}.$$

For ease of the exposition of the next theorem, we reason in the spherical counterpart to α and r:

- Define ϕ to be the maximum angle between any $w \in \mathcal{H}_C$ and w^*. From Lemma 1, under explanation $\mathcal{E}_{h^*}(\mathcal{X}, \alpha)$, $\phi = \arccos(\sqrt{1-\alpha^2}) = \arcsin \alpha$. Intuitively, ϕ measures how large $\mathcal{H}_C$ is and shrinks with a bigger set of explanations.
- Define $\psi = 2\arcsin(\frac{r}{2}) = \arccos(1 - \frac{r^2}{2})$. The boundary region $\mathcal{M}_r(\mathcal{X})$ may then be described as the set of points $\{x \in \mathcal{X} \mid \langle w^*, x\rangle \in [-\sin\psi, \sin\psi)\}$. Intuitively, ψ measures how "thick" the boundary region is. Geometrically, this means that $\theta(x, w^*) \in [\pi/2 - \psi, \pi/2 + \psi]$ for x in the boundary region, where $\theta(x, w^*)$ denotes the angle between x and w^* the decision boundary: $\theta(u, v) = \arccos(\frac{\langle u, v\rangle}{\|u\|_2\|v\|_2}) \in [0, \pi]$.

Please refer to Figure 2 for an illustration of notation ϕ and ψ, which we note are both acute by definition, and refer to Table 1 for a summary of definitions.

Firstly, it is clear that increasing boundary thickness ψ leads to a larger $\mathcal{M}_r(\mathcal{X})$, therefore a higher $\max_{(x,x')\in\mathcal{M}_r(\mathcal{X})} \pi_\alpha(x, x')$. We derive an analytical form of $\max_{(x,x')\in\mathcal{M}_r(\mathcal{X})} \pi_\alpha(x, x')$ below that formalizes this.

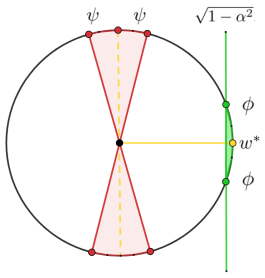

Figure 2: Visualization of the notation: $\mathcal{H}_C$ in green, boundary region in red and true model w^* in yellow.

Theorem 1. *We have:*

$$\max_{(x,x')\in\mathcal{M}_r(\mathcal{X})} \pi_\alpha(x, x') = \begin{cases} \frac{\int_0^{\psi/2} F(\theta)d\theta}{\int_0^{\phi} F(\theta)d\theta} & \psi \leq 2\phi \\ 1 & \psi > 2\phi, \end{cases}$$

where $F(\theta) = (1 - \frac{\cos^2\phi}{\cos^2\theta})^{d/2-1}$; therefore, it is strictly increasing for ψ in $[0, 2\phi]$.

Our next two theorems consider the margin-distancing effect in terms of α. For simplicity and to relate α's effect on $\mathcal{H}_C$ through explanations $\mathcal{E}_{h^*}(\mathcal{X}, \alpha)$, we subsequently abbreviate boundary certainty $\max_{(x,x')\in\mathcal{M}_r(\mathcal{X})} \pi_\alpha(x, x')$ as $\Pi(\alpha)$.

To recap, a higher threshold α, corresponding to more margin-distancing, leads to a smaller set of explanations $\mathcal{E}_{h^*}(\mathcal{X}, \alpha)$ (lowered transparency since more explanations are removed) and thus a bigger $\mathcal{H}_C$. This leads to lower boundary certainty $\Pi(\alpha)$, preventing gaming.

In the next result, we show that $\Pi(\alpha)$ is provably monotonically decreasing in α. Thus, this enables the use of binary search to efficiently find the optimal α. Indeed, it is not clear that decreasing the amount of explanations and enlarging the version space will always decrease $\Pi(\alpha)$. The reason is that enlarging $\mathcal{H}_C$ increases both models that agree with h^* on x, x'

r	max extent of manipulation
α	min distance from the margin
$\Pi(\alpha)$	boundary certainty, $\Pi(\alpha) = \max_{(x,x')\in\mathcal{M}_r(\mathcal{X})} \pi_\alpha(x, x')$
ϕ	max angle between $w \in \mathcal{H}_C$ and w^*; related to α by $\alpha = \sin\phi$
ψ	max angle: related to r by $\cos\psi = 1 - r^2/2$

Table 1: A table of notations that appears in Section 4.

(black lines in Figure 1) and models that do not (green lines). If *proportionally* more of them do predict like h^*, then the new $\pi_\alpha(x, x')$ will actually increase. We prove Theorem 2 that shows this is not so in this "nice" setting; the proof may be found in Appendix A.1.

Theorem 2. *$\Pi(\alpha)$ is decreasing in α, for $\alpha \in [0, 1)$, and is strictly decreasing in $[\sin(\psi/2), 1)$.*

Finally, in some cases, we may skip the search if we can analytically derive conditions on ϕ, ψ in which $\Pi(\alpha)$ is upper bounded. Next, we show that there exists some constant c such that $\lim_{\alpha\to 1}\Pi(\alpha) \le c\psi$. Thus, when ψ is small and α increases to 1, $\Pi(\alpha)$ decreases to a small value.

Theorem 3. *1. If $\alpha \ge 1 - \frac{1}{8d}$, then $\Pi(\alpha) \le 9\psi$.*

2. For any $C_1 \in (0, 1)$, there exists $C_2 > 0$ such that the following holds: if $\alpha \le 1 - \frac{1}{\sqrt{d}}$ and $\psi \ge \frac{C_2}{d^{1/4}}$, then $\Pi(\alpha) \ge 1 - C_1$.

A more refined version of this theorem and proofs of other theorems may be found in Appendix A.

5 GENERAL MODELS

For arbitrary feature spaces, it is unclear if it is possible to explicitly characterize $\mathcal{H}_C$ even for non-homogeneous linear models. Still, let us suppose we have devised some function Λ parameterized by threshold parameter α. Algorithmically, how do we search for the smallest α such that $\Pi(\alpha) < \kappa$ for a given κ?

First, we will need an approach to approximate $\Pi(\alpha)$ under a given threshold α. Indeed, there is generally no closed-form expression for $\Pi(\alpha)$, so we will assume access to an algorithm that can sample from the posterior distribution $\mathcal{U}(\mathcal{H}_C)$. Our approach is simply to draw samples $h_1, ..., h_n$ using the algorithm and evaluate: $\hat\rho(x') - \hat\rho(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{h_i(x') = 1\} - \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{h_i(x) = 1\}$.

To understand the sample complexity needed, we see that, $\hat\rho(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{h_i(x) = 1\} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{H_x^*(h_i) = 1\}$, where for a fixed x, $H_x^* : h \mapsto h(x)$ is its associated *dual* function.

Definition 1 (Dual Class)**.** *For any domain $\mathcal{X}$ and set of functions $\mathcal{H}$ whose image is $\{-1, +1\}$, the dual class of $\mathcal{H}$ is defined as $\mathcal{H}^* := \{H_x^* \mid x \in \mathcal{X}\}$.*

As introduced in (Assouad, 1983), $\mathrm{VC}(\mathcal{H}^*)$ is finite as long as $\mathrm{VC}(\mathcal{H})$ is finite. And so, with $O\left(\frac{\mathrm{VC}(H^*)+\log 1/\delta}{\epsilon^2}\right)$ random draws, we may obtain an $2\epsilon-$accurate estimation of $\hat\rho(x) - \hat\rho(x')$ for *all* boundary pairs x, x', due to uniform convergence. This gives us a 4ϵ-accurate estimation of $\Pi(\alpha)$. In the case of linear models, due to point-line duality, we know that $\mathrm{VC}(\mathcal{H}^*) = \mathrm{VC}(\mathcal{H}) = O(d)$, which informs us how many samples are needed to calculate a high fidelity approximation of $\pi_\alpha(x, x')$.

Search: Once we know how to approximate $\max \pi_\alpha(x, x')$ for a given α, if monotonicity does hold, then search for the optimal threshold may be efficiently done through binary search. Recall from Theorem 2 that, if a) the feature space is spherical, and b) the prior distribution over the hypothesis class is uniform, and c) the hypothesis class is homogeneous halfspaces, then $\Pi(\alpha)$ decreases monotonically to $O(\psi)$. To complement this result, we next show that removing one of a, b or c (and keeping the rest) breaks this pattern.

Our next two proposition show that, removing the spherical feature space condition, or removing the assumption of $\mathcal{U}$ being uniform, can cause boundary certainty to *increase* with increasing margin distancing parameter α in worst-case settings.

Proposition 1. *Suppose $d = 2$. We have uniform prior over homogeneous linear models $\mathcal{H} = \{w \in \mathbb{R}^d \mid \|w\| = 1\}$, there exists a feature space $\mathcal{X}$ and thresholds $0 < \alpha_2 < \alpha_1$ such that $\Pi(\alpha_2) < \Pi(\alpha_1)$.*

Proposition 2. *Suppose $\mathcal{X}$ is the d-dimensional unit sphere with $d \ge 3$. There exists a non-uniform distribution $\mathcal{U}$ over homogeneous linear models $\mathcal{H}$, such that there exists thresholds $0 < \alpha_2 < \alpha_1$ with $\Pi(\alpha_2) < \Pi(\alpha_1)$.*

Finally, we show that by removing the assumption that the hypothesis class is the set of homogeneous linear models, $\Pi(\alpha)$ can stay at a high value for all $\alpha \in (0, 1]$ and all $\psi \in (0, \pi]$. This is in sharp contrast with the homogeneous linear model class setting, in which $\lim_{\alpha\to 1}\Pi(\alpha) \le O(\psi)$ and could thus be made arbitrarily small with $\psi \to 0$.

Proposition 3. *There exists a class of non-homogeneous linear models, with spherical $\mathcal{X}$ such that $\Pi(\alpha)$ decreases monotonically (and strictly so at some*

point) with increasing α, and yet $\Pi(\alpha) \geq 1/3$ for all $\alpha \in [0, 1)$ and $\psi \in (0, \pi]$.

Thus, we have that in general monotonicity does not hold. However, our negative results are worst-case in nature. Next, we turn to experiments to examine the relationship between margin-distancing and boundary-certainty on real-world, non-worst case datasets.

6 EXPERIMENTS

In this section, we empirically chart the relationship between margin distancing (the amount of explanation omission) and boundary certainty. We experiment with linear and multi-layer Perceptron (MLP) models.

Explanation Methods: As mentioned in the formulation, we focus on example-based explanation methods that can return a subset of prototypical instances that serve as explanations. This leads us to use k-medoid and MMD-critic (Kim et al., 2016), and rules out other example-based explanation methods such as (Koh and Liang, 2017) that return a single (and not subset), most "influential" data point out of the training set. Note also, that counterfactual and contrastive-based explanations are ruled out by the need to margin-distance. Indeed, by construction, counterfactual/contrastive-based explanations are boundary points, whose release greatly increase the users' boundary certainty – in fact, $\max_{(x,x') \in \mathcal{M}_r(\mathcal{X})} \pi(x, x') = 1$. Thus, if manipulation (gaming) is to be prevented, the use and release of this type of explanations is a non-starter.

Our experimental procedure goes as follows:

1) The explanation method (e.g k-medoid) is used to compute the full set of explanations.

2) Then, we vary the degree of margin-distancing and remove explanations that are too close to the decision boundary. To measure the closeness of an explanation point with respect to the decision boundary, we look at its percentile in the distribution of all explanations' margin scores. This allows us to identify which points are in the top l percent of all explanations closest to the margin. We do this separately for positive and negative explanations as they have different distributions of margin scores.

3) To compute boundary certainty, we remove this top l percent closest explanations, compute models $\mathcal{H}_C$ consistent with the remaining explanations $\mathcal{E}_{h^*}(\mathcal{X}, \alpha)$ and compute $\pi(x, x')$ using $\mathcal{H}_C$.

4) To generate our plots, we vary l for l ranging from 0 to 75 (on the x-axis) and plot this against three metrics that capture boundary certainty (on the y-axis). The three metrics that summarize $\pi(x, x')$ for all boundary pairs (x, x') are: $\max_{(x,x') \in \mathcal{M}_r(\mathcal{X})} \pi(x, x')$ (worst boundary pair), average of top 5 percent of $\pi(x, x')$'s (somewhat worse case) and average of all $\pi(x, x')$.

6.1 Linear Models

Procedure: We train a linear model on the `Credit Card Default` dataset (Yeh and Lien, 2009) using Logistic Regression to obtain w^*. We focus on mutable features only that preclude features age and marital status. We take Λ to be margin distance $\langle w^*, x \rangle$. For these experiments, at a given r, we focus on and use w^* to find the set of all pairs of boundary points (x, x') that lead to a positive flip: $\left\{(x, x') : w^\star \cdot x < 0, w^\star \cdot x' \geq 0\right\}$. This is relatively cheap since by Cauchy-Schwarz, we only need to try all pairs of points whose margin score is $\leq r$, a much smaller set.

For a given set of explanations, we construct and sample from $\mathcal{H}_C$, which is a polytope. Sampling from polytopes is a well-studied problem and we use the state-of-the-art John's Walk (Chen et al., 2018a) with mixing time $O(d^2)$. We assume uniform $\mathcal{U}$ over $\mathcal{H}$. Thus, with these samples, we compute the empirical $\max_{(x,x') \in \mathcal{M}_r(\mathcal{X})} \hat{\pi}(x, x')$ with w's sampled uniformly from $\mathcal{H}_C$. We repeat this sampling 16 times for each set of explanations corresponding to a margin-distance percentile.

Monotonicity: We present our results in Figure 3. Qualitatively, we observe a generally smooth decreasing trend with increased distance of explanations from the margin and we observe some non-monotonicity under all three metrics, most prominently under the max metric. For all three metrics, we see that the trend levels out quickly. This suggests that trying smaller values of α (small amounts of explanation omission) can quickly decrease various measures of boundary certainty and this strategy is effective in this setting.

Quantitatively, we check if the trend is generally monotonic in an experiment that goes as follows. We pick 10 target boundary certainty values evenly spaced out from the attainable boundary certainties as found on the y-axis. Then, for each target value, we find the minimum percent of explanation points that need to be removed to bring the boundary certainty below the target; this optimal percentage is found simply by sweeping through all (percentage, certainty) pairs we have from left to right. Finally, we obtain the percentage that need to be removed as found by binary search and compute the difference between the percentage found by binary search against the optimal.

Under k-medoid explanations for linear model, we summarize the results by looking at the average of the difference and the max difference, which we report

as follows. For plots of the $\max_{(x,x')\in\mathcal{M}_r(\mathcal{X})}\pi(x,x')$: $r = 0.1$, 7, 35; $r = 0.2$, 11, 55; $r = 0.3$, 0, 0. For plots of average of top 5 percent of all $\pi(x,x')$: $r = 0.1$, 7, 35; $r = 0.2$, 10, 50; $r = 0.3$, 0, 0. For plots of average of all $\pi(x,x')$: $r = 0.1$, 6, 30; $r = 0.2$, 11, 55; $r = 0.3$, 0, 0. We record the full set of differences in tables in Appendix B.4.

As a synopsis, we observe that the difference is generally small for higher r's and larger for lower r's. The relatively jagged line means that binary search is likely to be quite far off. Here we wish to note that this problem may be alleviated by electing to try the smaller amounts of explanation omission instead of binary search, in the case that we find that the boundary certainties are close at the extremes. Indeed, the closeness would suggest that not much decrease in boundary certainty could be obtained by significantly increasing the percentage of explanation omission.

We also observe the result from varying the allowed extent of manipulation r. As expected, the larger the manipulation extent r, the higher the $\pi(x,x')$ that may be attainable.

6.2 Neural Network Models

Procedure: We train MLPs with one or two hidden layers on the `givemecredit`[1] dataset. We present the one layer MLP experiment results in the main body and the two layer in the appendix. We experiment with k-medoid and MMD-critic (Kim et al., 2016), whose results we present in the appendix. To measure of distance from margin, we take $\Lambda_\alpha(x)$ to be the model's confidence of a point: $\Lambda_\alpha(x) = \mathbb{1}\{|f^*(x)| \geq \alpha\}$, where $f^* : \mathcal{X} \to [-\frac{1}{2}, \frac{1}{2}]$ represents the MLP's predictive probability of class 1, offset by $-\frac{1}{2}$.

To the best of our knowledge, there is no known algorithm that provably sample uniformly from neural network version spaces. Indeed, this is an important problem described by recent works on the "Rashomon effect" (D'Amour et al., 2020; Semenova et al., 2019; Marx et al., 2020). We use the procedure in (D'Amour et al., 2020) used to probe the version space: randomly initialize the network with different seeds to obtain different models consistent with the explanations. For computational tractability, we sample 100 MLPs this way with 4 repetitions per margin-distance percentile.

Observations: Our first observation is that varying just the initialization is not an effective sampling procedure under the `givemecredit` dataset. We find small variation in the MLPs produced. To showcase this, we randomly sample 100 pairs of MLPs from the $\mathcal{H}_C$ we collected and calculate their label agreement on the boundary points, $\Pr_{h,h'\sim\mathcal{U}(\mathcal{H}_C),x\sim\mathrm{Unif}(\mathcal{M}_r(\mathcal{X}))}(h(x) = h'(x))$. The high average consistency of $\mathcal{H}_C$ is charted in green in Figure 5.

We also compute the three metrics in this setting (Figure 6), which interestingly are very high despite the overall low agreement with respect to h^* – defined as $\Pr_{h\sim\mathcal{U}(\mathcal{H}_C),x\sim\mathrm{Unif}(\mathcal{M}_r(\mathcal{X}))}(h(x) = h^*(x))$ (please see right figure in Figure 5). This seems to be due to a small fraction of points which most MLPs in $\mathcal{H}_C$ consistently agree with h^* on. The large values of $\max_{(x,x')\in\mathcal{M}_r(\mathcal{X})}\pi(x,x')$ in this case suggests the difficulty of preventing worst-case manipulation when the full set of hyperparameters used to train the network is known.

Indeed, as is noted in (Jagielski et al., 2020), it seems generally implausible for attackers to know the *exact* hyperparameters used to train the networks, which has been the assumption in the past model extraction works. And so, from hereon, we experiment with the natural, sampling procedure in the absence of such knowledge, which is just to randomly initialize the network and also the set of hyperparameters (ℓ_2 regularization constant, learning rate, momentum, batch size). These are randomly sampled from uniform distributions that contain the hyperparameters' true values. Verily, this leads to greater variation (please see the yellow barplots in Figure 5).

Since neural networks may require higher sample complexity, we also examine data augmentation techniques that one might consider to enhance the explanation set. In addition to 1) just the explanations, we consider 2) explanations plus random draws from Gaussian balls of radius 0.1 around the explanations 3) the full $\{x \mid x \in \mathcal{X}, \Lambda_\alpha(x) = 1\}$, which would correspond to "perfect" extrapolation of the feature space based off of $\mathcal{E}_{h^*}(\mathcal{X})$. The plots are given in Figure 4.

Comparing the effectiveness of the data augmentation, We observe small change in the π with mildly augmented data as in 1). However, the full knowledge of the $\{x \mid x \in \mathcal{X}, \Lambda_\alpha(x) = 1\}$ results in higher measures of boundary certainty. Indeed, this is to be expected since more labeled data naturally induces higher boundary certainty.

Monotonicity: In terms of the general trend for monotonicity, we again observe that margin-distancing does help to reduce all three metrics. Qualitatively, $\max_{(x,x')\in\mathcal{M}_r(\mathcal{X})}\pi(x,x')$ trend is non-monotonic and jagged at places, but smooths out with even a bit of averaging (the latter two metrics). In fact, we see that the average of top 5 percent of $\pi(x,x')$'s and average of all $\pi(x,x')$ metrics are monotonic. This is instructive in that it suggests that binary search could be used to efficiently search for the appropriate threshold.

[1] http://www.kaggle.com/c/GiveMeSomeCredit/

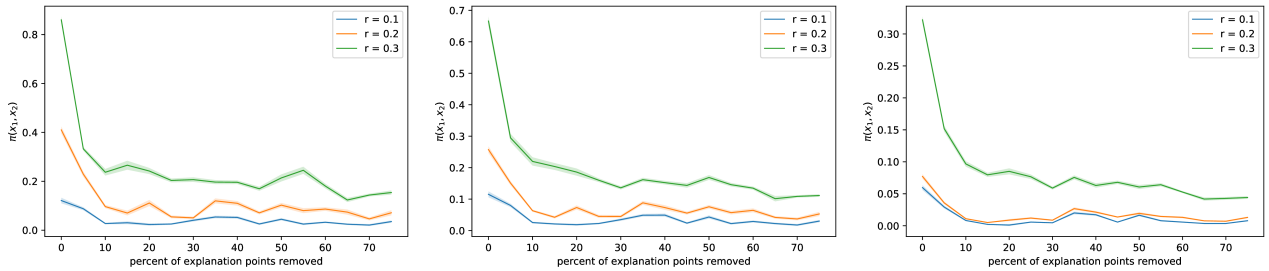

Figure 3: Plots of the $\max_{(x,x')\in\mathcal{M}_r(\mathcal{X})} \pi(x,x')$ (left), average of top 5 percent of all $\pi(x,x')$ (middle) and average of all $\pi(x,x')$ (right) under k-medoid explanations for linear models.

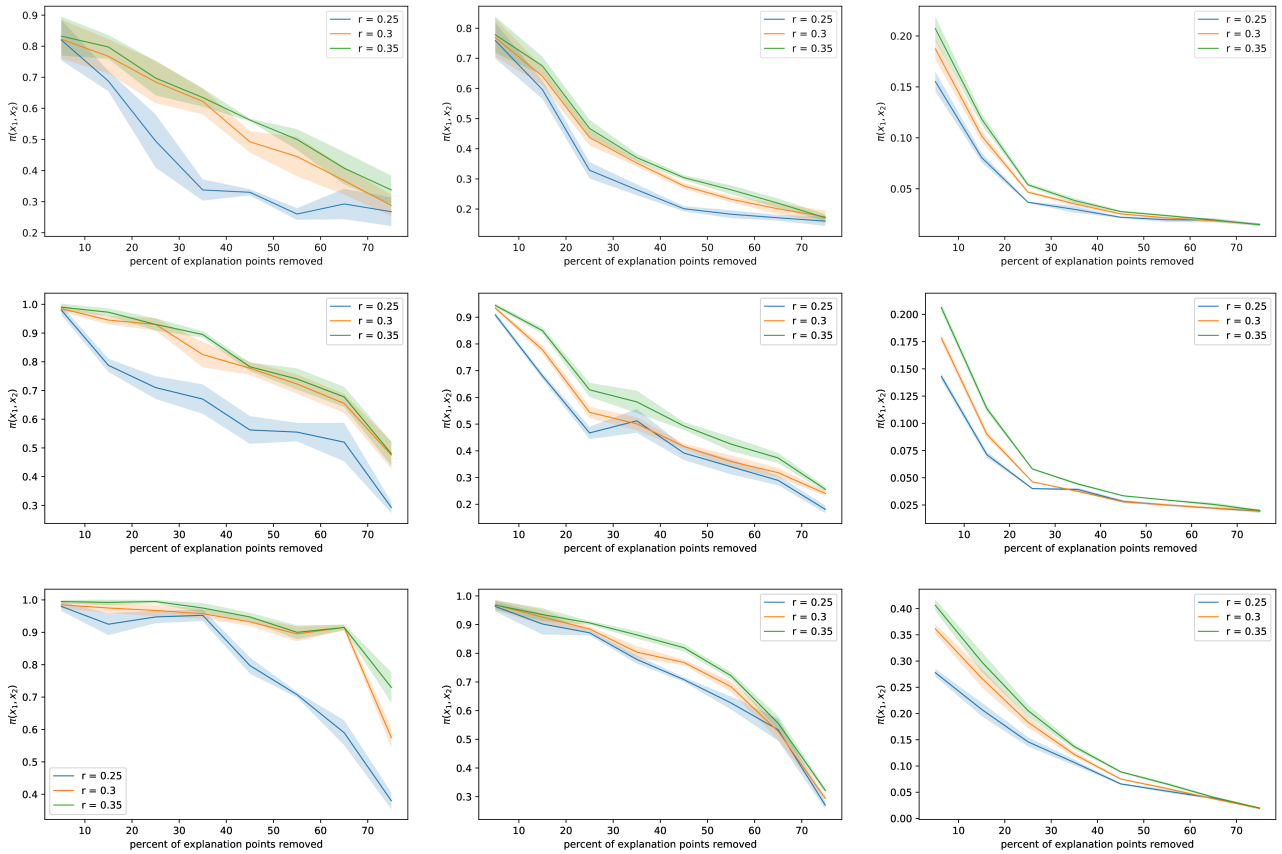

Figure 4: MLP results: k-medoid explanations (top), k-medoid explanations + random draws from small balls around the explanations (middle), full $\{x \mid x \in \mathcal{X}, \Lambda_\alpha(x) = 1\}$ (bottom). The three metrics are in column: max $\pi(x,x')$ (left), top 5 percent of all $\pi(x,x')$'s (middle), average $\pi(x,x')$ (right).

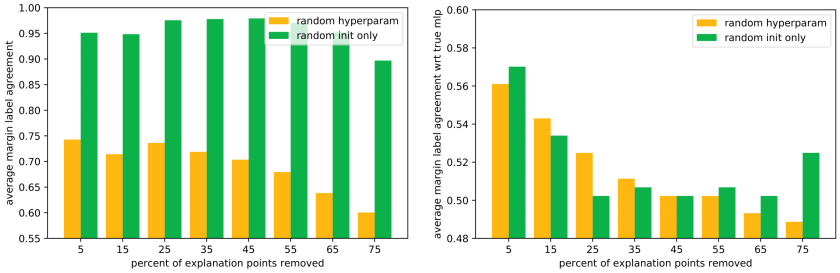

Figure 5: Boundary point label agreement within $\mathcal{H}_C$ (left), and boundary point label agreement of $\mathcal{H}_C$ with respect to h^* (right). This is estimated by sampling h from version space using random initializations of parameters (green) and hyperparameters (yellow), respectively.

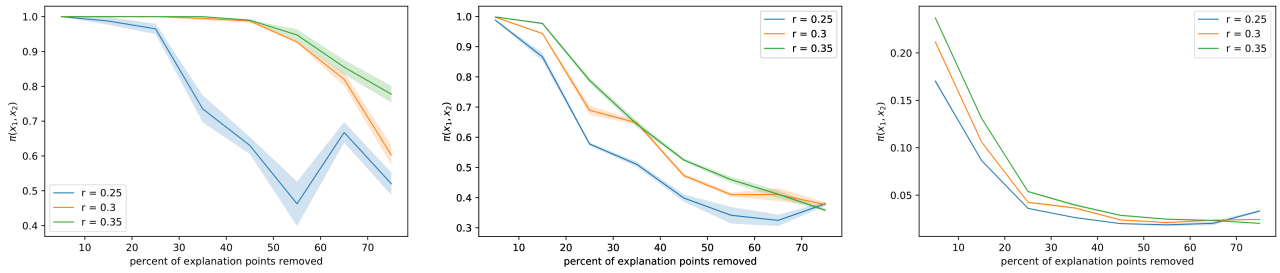

Figure 6: Plots of the $\max_{(x,x')\in\mathcal{M}_r(\mathcal{X})}\pi(x,x')$ (left), average of top 5 percent of all $\pi(x,x')$'s (middle) and average of all $\pi(x,x')$ (right) for the MLP case with random initialization only under k-medoid explanations.

Quantitatively, we verify if this trend is generally monotonic as before. We pick 10 target boundary certainty values evenly spaced out from the attainable boundary certainties as found on the y-axis. For each target value, we find the minimum percent of explanation points that need to be removed to bring the boundary certainty below the target and compare against the percentage found by binary search.

Under k-medoid explanations for MLP models, we again summarize the results by looking at the average of the difference and the max difference. Here, due to the much smoother curves (relative to those of the linear models) and the large discrepancy in boundary certainties at the two extremes, we find that under all three r's, binary search is able to match the optimal percentage needed to bring the boundary certainty below the target value.

6.3 Fair accessibility to explanations

A notable concern that may arise with margin distancing is that though omission of prototypical explanations is necessary, it may disproportionately affect individuals in regions close to the boundary. We plot the composition of the boundary region in the appendix under linear models logistic and SVM models. We observe that margin-distancing does disparately affect the release of explanations to different groups. Verily, this is another important factor that needs to be taken into account in the explanation release process.

7 CONCLUSION

In this paper, we propose margin-distancing as a way of making the tradeoff between transparency and gaming. We identify the source of the tension as boundary points. Our technical contribution is an "average-case" analysis of strategic manipulation with partial knowledge of the true model through model explanations. Altogether, this work puts the intersection between strategic ML and explainability on firmer theoretical foundation.

Our paper opens up several novel directions: 1) For what other settings can we prove monotonicity or upper bounds? Especially useful would be upper bounds on whether a certain threshold κ is achievable with at least l percent of all explanations. With this, one can avoid futile searches for non-realizable κ's. 2) How could we induce small boundary certainty for other types of explanations such as global explanations? 3) How else can we adapt explainability methods to account for gaming? For this, we believe our proposal of measuring $\mathcal{E}_{h^*}(\mathcal{X})$ quality in terms of the boundary certainty of $\mathcal{H}_C$ may still be helpful as a measure of how much a strategic agent can infer about h^* from $\mathcal{E}_{h^*}(\mathcal{X})$.

Acknowledgments. We thank the anonymous reviewers for helpful comments that improve the presentation of this paper. TY wishes to thank Ariel Procaccia, Yiling Chen and Chara Podimata for discussions.

References

U. Aïvodji, H. Arai, O. Fortineau, S. Gambs, S. Hara, and A. Tapp. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*, pages 161–170. PMLR, 2019.

C. Anders, P. Pasliev, A.-K. Dombrowski, K.-R. Müller, and P. Kessel. Fairwashing explanations with off-manifold detergent. In *International Conference on Machine Learning*, pages 314–323. PMLR, 2020.

P. Assouad. Densité et dimension. In *Annales de l'Institut Fourier*, volume 33, pages 233–282, 1983.

G. S. Becker. Crime and punishment: An economic approach. In *The economic dimensions of crime*, pages 13–68. Springer, 1968.

Y. Chen, R. Dwivedi, M. J. Wainwright, and B. Yu. Fast mcmc sampling algorithms on polytopes. *The Journal of Machine Learning Research*, 19(1):2146–2231, 2018a.

Y. Chen, C. Podimata, A. D. Procaccia, and N. Shah. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 9–26, 2018b.

A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

J. Dong, A. Roth, Z. Schutzman, B. Waggoner, and Z. S. Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

L. Edwards and M. Veale. Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.*, 16:18, 2017.

M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

M. Jagielski, N. Carlini, D. Berthelot, A. Kurakin, and N. Papernot. High accuracy and high fidelity extraction of neural networks. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pages 1345–1362, 2020.

J. V. Jeyakumar, J. Noor, Y.-H. Cheng, L. Garcia, and M. Srivastava. How can i explain this to you? an empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*, 33, 2020.

D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

B. Kim, O. Koyejo, R. Khanna, et al. Examples are not enough, learn to criticize! criticism for interpretability. In *NIPS*, pages 2280–2288, 2016.

J. Kleinberg and M. Raghavan. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–23, 2020.

P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

C. Marx, F. Calmon, and B. Ustun. Predictive multiplicity in classification. In *International Conference on Machine Learning*, pages 6765–6774. PMLR, 2020.

J. Miller, S. Milli, and M. Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pages 6917–6926. PMLR, 2020.

S. Milli, L. Schmidt, A. D. Dragan, and M. Hardt. Model reconstruction from model explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 1–9, 2019.

T. M. Mitchell. Version spaces: A candidate elimination approach to rule learning. In *Proceedings of the 5th international joint conference on Artificial intelligence-Volume 1*, pages 305–310, 1977.

W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

OpenSCHUFA. Openschufa project. 2019. URL `https://openschufa.de/`.

A. Selbst and J. Powles. "meaningful information" and the right to explanation. In *Conference on Fairness, Accountability and Transparency*, pages 48–48. PMLR, 2018.

L. Semenova, C. Rudin, and R. Parr. A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *arXiv preprint arXiv:1908.01755*, 2019.

D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. Stealing machine learning models via prediction apis. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pages 601–618, 2016.

S. Tsirtsis and M. Gomez-Rodriguez. Decisions, counterfactual explanations and strategic behavior. *arXiv preprint arXiv:2002.04333*, 2020.

S. Wachter, B. Mittelstadt, and L. Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017.

M. Wieringa. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 1–18, 2020.

I.-C. Yeh and C.-h. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

A Proofs

A.1 Section 4 Proofs

Recall that in Section 4, $\mathcal{X}$ is the origin-centered unit sphere in $\mathbb{R}^d$, and $\mathcal{H}$ is the set of homogeneous linear classisfiers in $\mathbb{R}^d$, and $\mathcal{U}$ denotes the uniform distribution over $\mathcal{H}$.

In the proofs that follow, we will mainly work in terms of polar angles ϕ and ψ. Recall $\phi = \arcsin \alpha$ is defined to be the maximum angle between any $w \in \mathcal{H}_C$ and w^*, and $\psi = 2\arcsin(\frac{r}{2})$ measures the thickness of the boundary region $\mathcal{M}_r(\mathcal{X})$.

Now, we prove a characterization of the boundary region in terms of ψ.

Fact 1. $\mathcal{M}_r(\mathcal{X}) = \big\{x \in \mathcal{X} \mid \langle w^*, x\rangle \in [-\sin\psi, \sin\psi)\big\}$.

Proof. Recall our definition that $\mathcal{M}_r(\mathcal{X}) := \{x \in \mathcal{X} \mid \exists x' \in \mathcal{R}_r(x) \wedge h^*(x') \neq h^*(x)\}$, where $h^*(x) = \text{sign}(\langle w^*, x\rangle)$. Thus, it suffices to show that

$$\langle w^*, x\rangle \in [-\sin\psi, \sin\psi) \iff \exists x' \in \mathcal{R}_r(x) \,.\, \text{sign}(\langle w^*, x'\rangle) \neq \text{sign}(\langle w^*, x\rangle).$$

We show the implications in both directions.

($\Rightarrow$): Suppose we are given x such that $\langle w^*, x\rangle \in [-\sin\psi, \sin\psi)$. Then x can be represented as $x = \beta w^* + \sqrt{1-\beta^2}x_\perp$, for some $\beta \in [-\sin\psi, \sin\psi)$, and $x_\perp$ is a unit vector perpendicular to w^*. Observe that $x - x_\perp = \beta w^* + (\sqrt{1-\beta^2} - 1)x_\perp$, and therefore,

$$\|x - x_\perp\|_2 = \sqrt{\beta^2 + (\sqrt{1-\beta^2} - 1)^2} = \sqrt{2(1 - \sqrt{1-\beta^2})}.$$

We now consider two cases of β:

1. If $\beta \in [-\sin\psi, 0)$, we consider $x' = x_\perp$. First observe that $x' \in \mathcal{R}_r(x)$. Indeed,

   $$\|x - x'\|_2 = \sqrt{2(1 - \sqrt{1-\beta^2})} \leq \sqrt{2(1-\cos\psi)} = r.$$

   Meanwhile, $\text{sign}(\langle w^*, x'\rangle) = \text{sign}(0) = 1 \neq -1 = \text{sign}(\beta) = \text{sign}(\langle w^*, x\rangle)$, which establishes the claim.

2. If $\beta \in [0, \sin\psi)$, we first observe that $\|x - x_\perp\| = \sqrt{2(1 - \sqrt{1-\beta^2})} < \sqrt{2(1-\cos\psi)} = r$. Therefore, there exists a small enough $\gamma > 0$, such that $x' = -\gamma w^* + \sqrt{1-\gamma^2}x_\perp$ is close enough to $x_\perp$, and hence lie in $\mathcal{R}_r(x)$. Now, $\text{sign}(\langle w^*, x'\rangle) = \text{sign}(-1) = -1 \neq 1 = \text{sign}(\beta) = \text{sign}(\langle w^*, x\rangle)$, which establishes the claim.

($\Leftarrow$): Assume toward contradiction that $\langle w^*, x\rangle \in [-1, -\sin\psi) \cup [\sin\psi, +1]$. Without loss of generality (due to spherical symmetry) suppose that $w^* = (1, 0, \ldots, 0)$ and $x = (\sin\theta, \cos\theta, 0, \ldots, 0)$ with $\theta \in [-\frac{\pi}{2}, -\psi) \cup [\psi, \frac{\pi}{2}]$.

Consider any $z \in \mathcal{X} \cap \mathcal{R}_r(x)$. We have:

$$\sum_{i=1}^{d} z_i^2 = 1,$$

$$(z_1 - \sin\theta)^2 + (z_2 - \cos\theta)^2 + \sum_{i=3}^{d} z_i^2 \leq r^2,$$

holding simultaneously. Combining the above two equations, we get

$$\sin\theta z_1 + \cos\theta z_2 \geq 1 - \frac{r^2}{2} = \cos\psi.$$

We now consider two cases of θ:

1. $\theta \in [\psi, \frac{\pi}{2}]$. In this case, $\cos\theta \leq \cos\psi$. And so, $\sin\theta z_1 \geq \cos\psi - \cos\theta z_2 \geq 0$. Therefore, for all $z \in \mathcal{R}_r(x)$, $\sin\theta \cdot z_1 \geq 0$ and hence $z_1 \geq 0$. In this case, $\text{sign}(\langle w^*, x\rangle) = \text{sign}(\sin\theta) = 1 = \text{sign}(z_1) = \text{sign}(\langle w^*, z\rangle)$.
2. $\theta \in [-\frac{\pi}{2}, -\psi)$. In this case, $\cos\theta < \cos\psi$. And so, $\sin\theta z_1 \geq \cos\psi - \cos\theta z_2 > 0$. Therefore, for all $z \in \mathcal{R}_r(x)$, $\sin\theta \cdot z_1 > 0$ and hence $z_1 < 0$. In conclusion, $\text{sign}(\langle w^*, x\rangle) = \text{sign}(\sin\theta) = -1 = \text{sign}(z_1) = \text{sign}(\langle w^*, z\rangle)$.

In either case, $\text{sign}(\langle w^*, x\rangle) = \text{sign}(\langle w^*, z\rangle)$ holds for all $z \in \mathcal{R}_r(x)$, which contradicts the assumption that $\exists x' \in \mathcal{R}_r(x)$. $\text{sign}(\langle w^*, x'\rangle) \neq \text{sign}(\langle w^*, x\rangle)$. This concludes the proof. □

Recall that we define $\Lambda_\alpha(x) = \mathbb{1}(|\langle w^*, x\rangle| > \alpha)$ and assume a uniform prior over homogeneous linear model class $\mathcal{H}$ and that $\mathcal{X}$ is the origin-centered unit sphere in $\mathbb{R}^d$. With this, we show that the trend of monotonicity exists in this "nice" setting and we can also develop direct upper bounds on Π.

To do this, we first begin by characterizing the version space,

Lemma 2 (Restatement of Lemma 1)**.** *Fix $\alpha \in [0, 1)$. Recall that $\mathcal{H}_C = \{h \in \mathcal{H} \mid h(x') = h^*(x'),\ \forall x' \in \mathcal{E}_{h^*}(\mathcal{X}, \alpha)\}$ is the version space induced by explanation $\mathcal{E}_{h^*}(\mathcal{X}, \alpha)$. $\mathcal{H}_C$ can be equivalently written as:*

$$\mathcal{H}_C = \left\{h_w \mid \|w\|_2 = 1, w \cdot w^* \geq \sqrt{1-\alpha^2}\right\}.$$

Proof. First observe that $w^* \in \mathcal{E}_{h^*}(\mathcal{X}, \alpha)$. We will show

$$(\forall x \in \mathcal{E}_{h^*}(\mathcal{X}, \alpha) \,.\, \text{sign}(\langle w, x\rangle) = \text{sign}(\langle w^*, x\rangle)) \iff \langle w, w^*\rangle \geq \sqrt{1-\alpha^2}.$$

We show the implications in both directions:

(⇒) First, since $w^* \in \mathcal{E}_{h^*}(\mathcal{X}, \alpha)$, we must have $\langle w, w^*\rangle \geq 0$.

Assume towards contradiction that $\langle w, w^*\rangle < \sqrt{1-\alpha^2}$, then w can be represented as $w = \sqrt{1-\beta^2}w^\star + \beta w_\perp$, where $\beta > \alpha$ and $w_\perp$ is a unit vector perpendicular to $w^\star$. We now show that there is an $x_0 \in \mathcal{E}_{h^*}(\mathcal{X}, \alpha)$ such that $\text{sign}(\langle w^*, x_0\rangle) \neq \text{sign}(\langle w, x_0\rangle)$, which will reach contradiction.

Choose $\gamma \in (\alpha, \beta)$, and define $x_0 = \gamma w^\star - \sqrt{1-\gamma^2}w_\perp$. It can be readily checked that $\langle w^*, x_0\rangle = \gamma > \alpha$, so $x_0 \in \mathcal{E}_{h^*}(\mathcal{X}, \alpha)$. Meanwhile, because $\gamma < \beta$,

$$\langle w, x_0\rangle = \sqrt{1-\beta^2}\gamma - \beta\sqrt{1-\gamma^2} = \sqrt{1-\beta^2}\gamma\left(1 - \frac{\beta}{\gamma}\cdot\frac{\sqrt{1-\gamma^2}}{\sqrt{1-\beta^2}}\right) < 0,$$

implying $\text{sign}(\langle w, x_0\rangle) = -1 \neq 1 = \text{sign}(\langle w^*, x_0\rangle)$.

(⇐) If $\langle w, w^*\rangle \geq \sqrt{1-\alpha^2}$, then w can be represented as $w = \sqrt{1-\beta^2}w^\star + \beta w_\perp$, where $\beta \leq \alpha$ and $w_\perp$ is a unit vector perpendicular to $w^\star$.

Now consider any $x \in \mathcal{E}_{h^*}(\mathcal{X}, \alpha)$; we would like to show that $\text{sign}(\langle w^*, x_0\rangle) = \text{sign}(\langle w, x_0\rangle)$. First, since $x \in \mathcal{E}_{h^*}(\mathcal{X}, \alpha)$, x can be represented as $x = \xi w^\star + \sqrt{1-\xi^2}x_\perp$, where $\xi \in [-1, -\alpha) \cup (\alpha, +1]$ and $x_\perp$ is a unit vector perpendicular to $w^\star$.

Without loss of generality, assume that $\xi \in (\alpha, +1]$; the case of $\xi \in [-1, \alpha)$ is symmetric. In this case, we have $\text{sign}(\langle w^\star, x\rangle) = 1$. Meanwhile,

$$\begin{aligned}
\langle w, x\rangle =& \langle\sqrt{1-\beta^2}w^\star + \beta w_\perp, \xi w^\star + \sqrt{1-\xi^2}x_\perp\rangle \\
=& \sqrt{1-\beta^2}\xi + \beta\sqrt{1-\xi^2}\langle w_\perp, x_\perp\rangle \\
\geq& \sqrt{1-\beta^2}\xi - \beta\sqrt{1-\xi^2} \\
=& \sqrt{1-\beta^2}\xi\left(1 - \frac{\beta}{\xi}\cdot\frac{\sqrt{1-\xi^2}}{\sqrt{1-\beta^2}}\right) > 0,
\end{aligned}$$

where the first inequality is by Cauchy-Schwarz; the second inequality uses the observation that $\beta \leq \alpha < \xi$. The above implies that $\text{sign}(\langle w, x\rangle) = 1 = \text{sign}(\langle w^\star, x\rangle)$. □

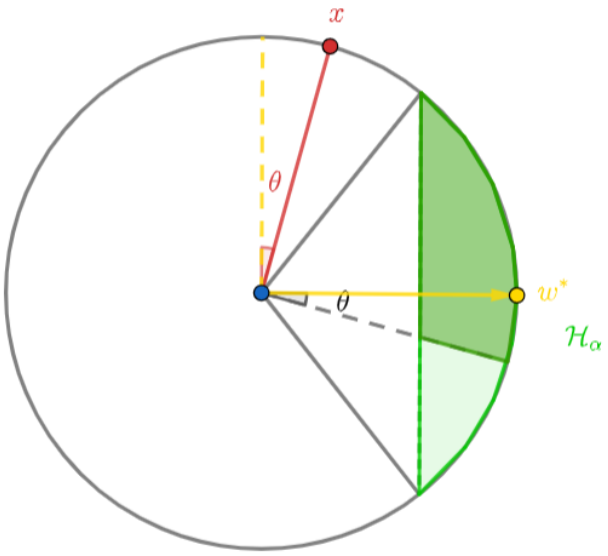

Figure 7: An illustration of $\Pr_{h_w \sim \mathcal{U}(\mathcal{H}_C)}(\langle w, x' \rangle \geq 0)$ in the proof of Theorem 1. Suppose x (red dot) has angle $\frac{\pi}{2} - \theta$ with w^*, and we project $\mathcal{U}(\mathcal{H}_C)$ to the 2-dimensional plane spanned by w^* and x; $\mathcal{U}(\mathcal{H}_C)$ (after projection) is supported on the green circle segment (the union of the dark and light green regions), whereas the subset $\{h_w \in \mathcal{H}_C : \langle w, x' \rangle \geq 0\}$ corresponds to the dark green region.

It is clear that increasing margin thickness ψ leads to a strictly bigger margin region, and a higher $\max_{(x,x') \in \mathcal{M}_r(\mathcal{X})} \pi_\alpha(x, x')$. We derive an analytical form of this.

Theorem 4 (Restatement of Theorem 1)**.** $\max_{(x,x') \in \mathcal{M}_r(\mathcal{X})} \pi_\alpha(x, x')$ *can be written as:*

$$\max_{(x,x') \in \mathcal{M}_r(\mathcal{X})} \pi_\alpha(x, x') = \begin{cases} \frac{\int_0^{\psi/2} F(\theta) d\theta}{\int_0^{\phi} F(\theta) d\theta} & \psi \leq 2\phi \\ 1 & \psi > 2\phi, \end{cases}$$

where $F(\theta) = (1 - \frac{\cos^2 \phi}{\cos^2 \theta})^{(d-2)/2}$*; therefore, it is strictly increasing for* ψ *in* $[0, 2\phi]$.

Proof. Denote by $F_+(\theta) = (1 - \frac{\cos^2 \phi}{\cos^2 \theta})_+^{(d-2)/2}$, where $(z)_+ := \max(z, 0)$. Note that $F_+(\theta) = 0$ if $\theta \notin [-\phi, \phi]$.

To show the theorem statement, note that $\int_{-\pi}^{\pi} F_+(\theta) d\theta = 2 \int_0^{\phi} F(\theta) d\theta$; it therefore suffices to show that,

$$\max_{(x,x') \in \mathcal{M}_r(\mathcal{X})} \pi_\alpha(x, x') = \frac{\int_{-\psi/2}^{\psi/2} F_+(\theta) d\theta}{\int_{-\pi}^{\pi} F_+(\theta) d\theta}$$

We show the left hand side is both at most and at least the right hand side, respectively. Without loss of generality, let $w^* = (1, 0, \ldots, 0)$.

1. LHS $\geq$ RHS: We choose $x' = (\sin \frac{\psi}{2}, \cos \frac{\psi}{2}, 0, \ldots, 0)$, $x = (-\sin \frac{\psi}{2}, \cos \frac{\psi}{2}, 0, \ldots, 0)$. It can be seen that $\|x - x'\|_2 = 2 \sin \frac{\psi}{2} = r$, and $\langle w^*, x' \rangle > 0$, $\langle w^*, x \rangle < 0$, and therefore (x, x') is indeed a boundary pair (i.e. in $\mathcal{M}_r(\mathcal{X})$).

   In addition, for $w = (w_1, w_2)$, denote by $\phi(w) \in (-\pi, \pi]$ its polar angle with respect to $(1, 0)$ (so that $\phi((1, 0)) = 0$).

   In this case, by Claim 1 given below (see also Figure 7 for an illustration), we have:

$$\begin{aligned} \Pr_{h_w \sim \mathcal{U}(\mathcal{H}_C)}(\langle w, x' \rangle \geq 0) &= \Pr_{h_w \sim \mathcal{U}(\mathcal{H}_C)}(\phi(w) \in [-\psi/2, \pi/2]) \\ &= \frac{\int_{-\psi/2}^{\pi/2} F_+(\theta) d\theta}{\int_{-\pi}^{\pi} F_+(\theta) d\theta}, \end{aligned}$$

and

$$\begin{aligned}\Pr_{h_w\sim\mathcal{U}(\mathcal{H}_C)}(\langle w,x\rangle\geq 0) =& \Pr_{h_w\sim\mathcal{U}(\mathcal{H}_C)}(\phi(w)\in[\psi/2,\pi/2])\\ =&\frac{\int_{\psi/2}^{\pi/2}F_+(\theta)d\theta}{\int_{-\pi}^{\pi}F_+(\theta)d\theta},\end{aligned}$$

and therefore,

$$\max_{(x,x')\in\mathcal{M}_r(\mathcal{X})}\pi_\alpha(x,x')\geq \Pr_{h_w\sim\mathcal{U}(\mathcal{H}_C)}(\langle w,x'\rangle\geq 0)-\Pr_{h_w\sim\mathcal{U}(\mathcal{H}_C)}(\langle w,x\rangle\geq 0)=\frac{\int_{-\psi/2}^{\psi/2}F_+(\theta)d\theta}{\int_{-\pi}^{\pi}F_+(\theta)d\theta}$$

2. LHS $\leq$ RHS: First, for every $z\in\mathbb{R}^d$, denote by $\theta(w^*,z)=\arccos(\frac{\langle w^*,z\rangle}{\|w^*\|\|z\|})\in[0,\pi]$ the angle between z and w^*.

$$\Pr_{h_w\sim\mathcal{U}(\mathcal{H}_C)}(\langle w,z\rangle\geq 0)=\frac{\int_{\theta(w,z)-\frac{\pi}{2}}^{\frac{\pi}{2}}F_+(\theta)d\theta}{\int_{-\pi}^{\pi}F_+(\theta)d\theta}$$

To see this, without loss of generality, let $z=(z_1,z_2,0,\ldots,0)$. Then, by Claim 1 (given below), we have

$$\Pr_{h_w\sim\mathcal{U}(\mathcal{H}_C)}(\langle w,z\rangle\geq 0)=\Pr_{h_w\sim\mathcal{U}(\mathcal{H}_C)}\left(\phi((z_1,z_2))\in[\theta(w,z)-\frac{\pi}{2},\frac{\pi}{2}]\right)=\frac{\int_{\theta(w,z)-\frac{\pi}{2}}^{\frac{\pi}{2}}F_+(\theta)d\theta}{\int_{-\pi}^{\pi}F_+(\theta)d\theta}.$$

Therefore, for every $(x,x')\in\mathcal{M}_r(\mathcal{X})$,

$$\begin{aligned}&\Pr_{h_w\sim\mathcal{U}(\mathcal{H}_C)}(\langle w,x'\rangle\geq 0)-\Pr_{h_w\sim\mathcal{U}(\mathcal{H}_C)}(\langle w,x\rangle\geq 0)\\ =&\frac{\int_{\theta(w,x')-\frac{\pi}{2}}^{\theta(w,x)-\frac{\pi}{2}}F_+(\theta)d\theta}{\int_{-\pi}^{\pi}F_+(\theta)d\theta}\\ \leq&\frac{\max\left\{\int_a^b F_+(\theta)d\theta : b-a\leq\psi\right\}}{\int_{-\pi}^{\pi}F_+(\theta)d\theta},\end{aligned}$$

where the inequality follows by observing $\theta(w,x)-\theta(w,x')\leq\theta(x,x')\leq\psi$, which follows from $2\sin\frac{\theta(x,x')}{2}=\|x-x'\|\leq r=2\sin\frac{\psi}{2}$ and that $\psi/2$ is acute by definition, which means that $\theta(x,x')/2\leq\psi/2$ and $\psi/2$ are both acute. It suffices to show that for every a,b such that $b-a\leq\psi$,

$$\int_a^b F_+(\theta)d\theta\leq\int_{-\psi/2}^{\psi/2}F_+(\theta)d\theta. \tag{1}$$

As $F_+(\theta)\geq 0$ for any $\theta\in\mathbb{R}$, the max must be achieved at $b-a=\psi$ and so it suffices to show $\forall c$,

$$\int_{c-\psi/2}^{c+\psi/2}F_+(\theta)d\theta\leq\int_{-\psi/2}^{\psi/2}F_+(\theta)d\theta.$$

Let $F(c)=\int_{c-\psi/2}^{c+\psi/2}F_+(\theta)d\theta$; it can be seen that $F'(a)=F_+(c+\psi/2)-F_+(c-\psi/2)$. Therefore,

$$F'(c)\begin{cases}\geq 0 & c\leq-\psi/2\\ \geq 0 & -\psi/2\leq c\leq 0\\ \leq 0 & 0\leq c\leq\psi/2\\ \leq 0 & c\geq\psi/2,\end{cases}$$

and hence $\max_{c\in\mathbb{R}}F(c)=F(0)=\int_{-\psi/2}^{+\psi/2}F_+(\theta)d\theta$, which concludes the proof of Equation (1), and concludes that LHS $\leq$ RHS. □

Fact 2. *The probability density function of the uniform distribution over unit sphere projected onto the first two dimensions is*

$$p(w_1, w_2) = \frac{d-2}{2\pi}(1 - w_1^2 - w_2^2)^{\frac{d-4}{2}}.$$

Claim 1. *In the notation of the proof of Theorem 1 above, for every $a < b$ such that $[a, b] \subset (-\pi, \pi]$,*

$$\Pr_{h_w \sim \mathcal{U}(\mathcal{H}_C)} \left(\phi((w_1, w_2)) \in [a, b]\right) = \frac{\int_a^b F_+(\theta)d\theta}{\int_{-\pi}^{\pi} F_+(\theta)d\theta}$$

Proof. Recall Lemma 1 that characterizes $\mathcal{H}_C$ (see also Figure 2), we have:

$$\Pr_{h_w \sim \mathcal{U}(\mathcal{H}_C)} \left(\phi((w_1, w_2)) \in [a, b]\right) = \frac{\Pr_{h_w \sim \mathcal{U}} \left(w_1 \geq \sqrt{1-\alpha^2}, \phi((w_1, w_2)) \in [a, b]\right)}{\Pr_{h_w \sim \mathcal{U}} \left(w_1 \geq \sqrt{1-\alpha^2}\right)}$$

From Fact 2 above, we can express the numerator and the denominator in integral form. For the denominator, by changing of variables to the polar coordinates,

$$\begin{aligned}
&\Pr_{h_w \sim \mathcal{U}} \left(w_1 \geq \sqrt{1-\alpha^2}\right) \\
=&\int_{-\phi}^{\phi} \left(\int_{\frac{\cos\phi}{\cos\theta}}^{1} \frac{d-2}{2\pi}(1-r^2)^{\frac{d-4}{2}} r dr \right) d\theta \\
=&\frac{1}{2\pi} \int_{-\phi}^{\phi} \left(1 - \frac{\cos^2\phi}{\cos^2\theta}\right)^{\frac{d-2}{2}} d\theta \\
=&\frac{1}{2\pi} \int_{-\pi}^{\pi} F_+(\theta)d\theta.
\end{aligned}$$

For the numerator,

$$\begin{aligned}
&\Pr_{h_w \sim \mathcal{U}} \left(w_1 \geq \sqrt{1-\alpha^2}, \phi((w_1, w_2)) \in [a, b]\right) \\
=&\int_{\max(-\phi, a)}^{\min(\phi, b)} \left(\int_{\frac{\cos\phi}{\cos\theta}}^{1} \frac{d-2}{2\pi}(1-r^2)^{\frac{d-4}{2}} r dr \right) d\theta \\
=&\frac{1}{2\pi} \int_{\max(-\phi, a)}^{\min(\phi, b)} \left(1 - \frac{\cos^2\phi}{\cos^2\theta}\right)^{\frac{d-2}{2}} d\theta \\
=&\frac{1}{2\pi} \int_{a}^{b} F_+(\theta)d\theta.
\end{aligned}$$

The lemma follows by combining two equalities above. □

Theorem 5 (Restatement of Theorem 2)**.** $\Pi(\alpha)$ *is decreasing in α, for $\alpha \in [0, 1)$, and is strictly decreasing in* $[\sin(\psi/2), 1)$.

Proof. Consider $\Pi(\alpha)$ for $\alpha = \sin\phi \in [\sin(\psi/2), 1]$, which, from the proof of Theorem 1, has the following form:

$$\begin{aligned}
\Pi(\alpha) =& \Pr_{h_w \sim \mathcal{U}(\mathcal{H}_C)} \left(\phi(w) \in [-\psi/2, \psi/2]\right) \\
=&\frac{\Pr_{h_w \sim \mathcal{U}} \left(w_1 \geq \sqrt{1-\alpha^2}, \phi((w_1, w_2)) \in [-\psi/2, \psi/2]\right)}{\Pr_{h_w \sim \mathcal{U}} \left(w_1 \geq \sqrt{1-\alpha^2}\right)} \\
=&\frac{\int_{\sqrt{1-\alpha^2}}^{1} \left(\int_0^{w_1 \tan\psi} p(w_1, w_2) dw_2 \right) dw_1}{\int_{\sqrt{1-\alpha^2}}^{1} \left(\int_0^{\sqrt{1-w_1^2}} p(w_1, w_2) dw_2 \right) dw_1},
\end{aligned}$$

where $p(w_1, w_2) = \frac{d-2}{2\pi}(1 - w_1^2 - w_2^2)^{(d-4)/2}$ is the pdf of (w_1, w_2) when $h_w \sim \mathcal{U}$ (Fact 2).

Consider $f(w_1) = \int_0^{w_1 \tan\psi} p(w_1, w_2)dw_2$, and $g(w_1) = \int_0^{\sqrt{1-w_1^2}} p(w_1, w_2)dw_2$, and $F(t) = \frac{\int_t^1 f(w_1)dw_1}{\int_t^1 g(w_1)dw_1}$. ;with this, $\Pi(\alpha) = F(\sqrt{1-\alpha^2})$. It suffices to show that $F(t)$ is monotonically increasing, i.e. $F'(t) \geq 0$ for all t.

To show this, first observe that $\frac{f(w_1)}{g(w_1)}$ is monotonically increasing: indeed,

$$\frac{f(w_1)}{g(w_1)} = \frac{\int_0^{\frac{w_1\tan\psi}{\sqrt{1-w_1^2}}} (1-v^2)^{\frac{d-4}{2}} dv}{\int_0^1 (1-v^2)^{\frac{d-4}{2}} dv},$$

which is increasing in w_1. As a consequence,

$$\int_t^1 f(w_1)dw_1 = \int_t^1 g(w_1) \cdot (\frac{f(w_1)}{g(w_1)})dw_1 \geq \frac{f(t)}{g(t)} \cdot \int_t^1 g(w_1)dw_1 \tag{2}$$

Therefore,

$$F'(t) = \frac{-f(t)\int_t^1 g(w_1)dw_1 + g(t)\int_t^1 f(w_1)dw_1}{(\int_t^1 g(w_1)dw_1)^2} \geq 0,$$

where the last inequality is from Equation (2). □

Below, we derive bounds on $\Pi(\alpha)$ given specific assumptions on ϕ and ψ.

Theorem 6 (Refined version of Theorem 3)**.** *We have the following:*

1. *If* $\cos\phi \leq \frac{1}{2d^{1/4}}$*, then* $\Pi(\alpha) \leq 6 \cdot \left(\psi(1 + d^{\frac{1}{2}}\cos\phi)\right)$.
2. *For any* $c_1, c_2 > 0$*, there exists* $c_3 > 0$ *such that the following holds: given any* $\phi \in \left[c_1, \frac{\pi}{2}\right)$*, and*

$$\psi \geq c_3 \max\left(\cos\phi, \frac{1}{d^{\frac{1}{2}}\cos\phi}\sqrt{\ln\frac{4}{c_2} + \ln\left(1 + \frac{1}{d^{\frac{1}{2}}\cos\phi}\right)}\right), \tag{3}$$

   then $\Pi(\alpha) \geq 1 - c_2$.

Before presenting the proof of Theorem 6, we first show how it concludes the proof of Theorem 3.

Proof of Theorem 3. We show the two items respectively.

1. Recall that $\alpha = \sin\phi$. If $\alpha \geq 1 - \frac{1}{8d}$, then $\cos^2\phi = 1 - \alpha^2 \leq \frac{1}{4d}$, implying that $\cos\phi \leq \frac{1}{2\sqrt{d}}$. As $\frac{1}{2\sqrt{d}} \leq \frac{1}{2d^{1/4}}$, the conditions of item 1 of Theorem 6 is satisfied. As a result,

$$\Pi(\alpha) \leq 6 \cdot \left(\psi(1 + d^{\frac{1}{2}}\cos\phi)\right) \leq 9\psi.$$

2. Let $C_1 \in (0,1)$. Choose $\phi' := \arccos(\frac{1}{d^{1/4}})$. Note that $\phi' \geq \phi$, since $1 - \cos^2\phi = \alpha^2 = (1 - \frac{1}{\sqrt{d}})^2 \leq 1 - \frac{1}{\sqrt{d}} = 1 - \cos^2\phi'$. Denote by $\alpha := \sin\phi$ and $\alpha' := \sin\phi'$; we have $\alpha' \geq \alpha$.
   In addition, as $\phi' = \arccos(\frac{1}{d^{1/4}})$, there exists some numerical constant $c_1 > 0$ such that $\phi' \geq c_1$. Now, by item 2 of Theorem 6, there exists some $c_3 > 0$, such that when $\psi \geq \frac{c_3\sqrt{\ln\frac{8}{C_1}}}{d^{1/4}} \geq c_3 \max\left(\frac{1}{d^{1/4}}, \frac{\sqrt{\ln\frac{4}{C_1} + \ln\left(1+\frac{1}{d^{1/4}}\right)}}{d^{1/4}}\right)$, $\Pi(\alpha') \geq 1 - C_1$. Now, as $\Pi(\cdot)$ is monotonically decreasing in α, $\Pi(\alpha) \geq \Pi(\alpha') \geq 1 - C_1$. Therefore, the theorem statement holds with $C_2 = c_3\sqrt{\ln\frac{8}{C_1}}$. □

We now present the proof of Theorem 6.

Proof. Recall that

$$\Pi(\alpha) = \begin{cases} \frac{\int_0^{\psi/2} F(\theta)d\theta}{\int_0^{\phi} F(\theta)d\theta}, & \arcsin\alpha = \phi \geq \psi/2 \\ 1, & \arcsin\alpha = \phi < \psi/2. \end{cases}$$

1. First we note that $\cos\phi \leq \frac{1}{2d^{1/4}}$ implies that $\phi \geq \frac{\pi}{3}$.

   If $\phi \leq \psi/2$, then $\psi \geq \frac{2}{3}\pi$. Therefore, $\Pi(\alpha) = 1 \leq 6\psi \leq 6 \cdot \left(\psi(1 + d^{\frac{1}{2}}\cos\phi)\right)$ holds.

   For the rest of the proof, we focus on the case of $\phi > \psi/2$. In this case, $\Pi(\alpha)$ equals the integral ratio $\frac{\int_0^{\psi/2} F(\theta)d\theta}{\int_0^{\phi} F(\theta)d\theta}$. With foresight, define $\theta' = \min\left(\frac{\phi}{2}, \arctan(\frac{1}{d^{\frac{1}{2}}\cos\phi}), \arccos(d^{\frac{1}{4}}\cos\phi)\right)$. As we will see below, this is a "critical threshold" of the integral $\int_0^{\phi} F(\theta)d\theta$, in the sense that the contribution of $[\theta', \psi]$ to the integral is negligible.

   By our assumption that $\cos\phi \leq \frac{1}{2d^{\frac{1}{4}}}$, $\arccos(d^{\frac{1}{4}}\cos\phi) \geq \frac{\pi}{3}$. In addition, $\arctan(\frac{1}{d^{\frac{1}{2}}\cos\phi}) \geq \min\left(\frac{\pi}{4}, \frac{1}{2d^{\frac{1}{2}}\cos\phi}\right)$ by Lemma 6 given after the proof. Moreover, recall that $\phi \geq \frac{\pi}{3}$. Combining the above bounds, $\theta' \geq \min\left(\frac{\pi}{6}, \frac{1}{2d^{\frac{1}{2}}\cos\phi}\right)$.

   We now upper bound $\Pi(\alpha)$. First we upper bound the numerator:

$$\int_0^{\psi/2} F(\theta)d\theta \leq \psi/2 \cdot F(0) = \frac{\psi}{2}(1-\cos^2\phi)^{\frac{d-2}{2}} \leq \frac{\psi}{2}\exp\left(-\frac{d-2}{2}\cos^2\phi\right).$$

   We next lower bound the denominator. As $\theta' \leq \frac{\phi}{2} \leq \frac{\pi}{4}$ (since by definition, $\phi/2 \leq \pi/2$), this implies that $\cos^2\theta' \geq \frac{1}{2}$ and hence $\phi \geq \pi/3 \Rightarrow \frac{\cos^2\phi}{\cos^2\theta'} \in [0, \frac{1}{2}]$. Therefore,

$$\int_0^{\phi} F(\theta)d\theta \geq \int_0^{\theta'} F(\theta)d\theta \geq \theta' F(\theta') = \theta'\left(1 - \frac{\cos^2\phi}{\cos^2\theta'}\right)^{\frac{d-2}{2}} \geq \theta'\exp\left(-\frac{d-2}{2}\left(\frac{\cos^2\phi}{\cos^2\theta'} + \frac{\cos^4\phi}{\cos^4\theta'}\right)\right),$$

   where the last inequality uses the elementary fact that $1 - x \geq \exp(-x - x^2)$ for $x \in [0, \frac{1}{2}]$.

   Combining the upper and lower bounds, we get that the integral ratio is bounded by:

$$\frac{\int_0^{\psi/2} F(\theta)d\theta}{\int_0^{\phi} F(\theta)d\theta} \leq \frac{\psi}{2\theta'}\exp\left(\frac{d-2}{2}\left(\cos^2\phi\tan^2\theta' + \frac{\cos^4\phi}{\cos^4\theta'}\right)\right)$$

   From our choice of θ', it can be easily seen that: (1) $\cos^2\phi\tan^2\theta' \leq \cos^2\phi \cdot \frac{1}{d\cos^2\phi} \leq \frac{1}{d}$, and (2) $\frac{\cos^4\phi}{\cos^4\theta'} \leq \frac{\cos^4\phi}{(d^{\frac{1}{4}}\cos\phi)^4} \leq \frac{1}{d}$. This implies that the exponential term is at most $\exp\left(\frac{d-2}{2} \cdot \frac{2}{d}\right) \leq e$.

   In conclusion, we have that:

$$\frac{\int_0^{\psi/2} F(\theta)d\theta}{\int_0^{\phi} F(\theta)d\theta} \leq \frac{e}{2} \cdot \frac{\psi}{\theta'} \leq 6 \cdot \left(\psi(1 + d^{\frac{1}{2}}\cos\phi)\right),$$

   where in the last inequality we recall that $\theta' \geq \min\left(\frac{\pi}{6}, \frac{1}{2d^{\frac{1}{2}}\cos\phi}\right)$, and use that for $A, B > 0$,$\max(A, B) \leq A + B$.

2. Fix $c_1, c_2 > 0$, and let $\phi \geq c_1$.

   If $\phi \leq \psi/2$, then $\Pi(\alpha) = 1 \geq 1 - c_2$ holds.

   For the rest of the proof, we focus on the case of $\phi > \psi/2$. As $\phi \geq c_1 > 0$, $\cos\phi \leq \cos c_1 < 1$. Therefore there exists some small constant $c_5 > 0$ such that $\cos\phi \leq 1 - 2c_5$; meanwhile there exists some small enough constant $c_4 < \frac{1}{4}$ such that $\cos^2(c_4\psi) \geq 1 - c_5$ since $c_4\psi \leq \pi/4$; as a consequence, $\cos^2\phi / \cos^2(c_4\psi) \leq \frac{1-2c_5}{1-c_5} \leq 1 - c_5$. In summary, there exist some small enough constants $c_4, c_5 > 0$ (independent of ϕ), such that $c_4 < \frac{1}{4}$ and $\frac{\cos^2\phi}{\cos^2(c_4\psi)} \leq 1 - c_5$.

   By Lemma 5 (deferred after the proof), there exists some constant $c_6 > 0$ (independent of ϕ) such that

$$1 - \frac{\cos^2\phi}{\cos^2(c_4\psi)} \geq \exp\left(-\left(\frac{\cos\phi}{\cos(c_4\psi)}\right)^2 - c_6\left(\frac{\cos\phi}{\cos(c_4\psi)}\right)^4\right). \tag{4}$$

   Therefore,

$$\begin{aligned}
\frac{\int_0^{\psi/2} F(\theta)d\theta}{\int_{\psi/2}^{\phi} F(\theta)d\theta} \geq & \frac{\int_0^{c_4\psi} F(\theta)d\theta}{\int_{\psi/2}^{\phi} F(\theta)d\theta} \\
\geq & \frac{c_4\psi \cdot F(c_4\psi)}{\phi \cdot F(\psi/2)} \\
\geq & \frac{2c_4\psi}{\pi} \cdot \frac{\left(1 - \frac{\cos^2\phi}{\cos^2(c_2\psi)}\right)^{(d-2)/2}}{\left(1 - \frac{\cos^2\phi}{\cos^2(\psi/2)}\right)^{(d-2)/2}} \\
\geq & \frac{2c_4\psi}{\pi} \cdot \frac{\exp\left(-\frac{d-2}{2}\left(\frac{\cos^2\phi}{\cos^2(c_4\psi)} + c_6\left(\frac{\cos^2\phi}{\cos^2(c_4\psi)}\right)^2\right)\right)}{\exp\left(-\frac{d-2}{2}\frac{\cos^2\phi}{\cos^2(\psi/2)}\right)} \\
= & \frac{2c_4\psi}{\pi} \cdot \exp\left(\frac{d-2}{2}\cos^2\phi\left(\frac{1}{\cos^2(\psi/2)} - \frac{1}{\cos^2(c_4\psi)} - c_6\frac{\cos^2\phi}{\cos^4(c_4\psi)}\right)\right),
\end{aligned}$$

   where the first inequality is because $c_4 \leq \frac{1}{4}$; the second inequality is because $F(\theta)$ is monotonically decreasing for $\theta \geq 0$; the third inequality follows from the definition of $F(\theta)$, and $\phi \leq \frac{\pi}{2}$; the fourth inequality is from Equation (4) as well as using $1 - x \leq \exp(-x)$ to upper bound the denominator; the equality is by algebra.

   Observe:

$$\begin{aligned}
\frac{1}{\cos^2(\psi/2)} - \frac{1}{\cos^2(c_4\psi)} = & \frac{\cos^2(c_4\psi) - \cos^2(\psi/2)}{\cos^2(c_4\psi) \cdot \cos^2(\psi/2)} \\
= & \frac{\sin^2(\psi/2) - \sin^2(c_4\psi)}{\cos^2(c_4\psi) \cdot \cos^2(\psi/2)} \\
= & \frac{(\sin(\psi/2) + \sin(c_4\psi))(\sin(\psi/2) - \sin(c_4\psi))}{\cos^2(c_4\psi) \cdot \cos^2\psi} \\
\geq & \frac{\frac{\psi}{2\pi} \cdot \cos(\psi/2)\frac{\psi}{4}}{\cos^2(c_4\psi) \cdot \cos^2\psi} \\
\geq & \frac{\psi^2}{8\pi}.
\end{aligned}$$

   where the first inequality uses, $\sin(\psi/2) \geq \frac{\psi}{2\pi}$, and the Lagrange mean value theorem and the choice of c_4, such that $c_4 \leq \frac{1}{4}$ so that $\sin(\psi/2) - \sin(c_4\psi) = (\psi/2 - c_4\psi)\cos\xi$ for some $\xi \in [c_4\psi, \psi/2]$, which in turn is $\geq \frac{\psi}{4}\cos(\psi/2)$; the second inequality uses that $\cos(c_4\psi) \geq \cos(\psi/2)$, and $\cos\gamma \leq 1$ for any γ.

   With foresight, we will choose $c_3 \geq 16\sqrt{c_6}$, and defer the exact setting of c_3 to the next paragraph. By the assumption of lower bound on ψ (Equation (3)), We have $\psi \geq 16\sqrt{c_6}\cos\phi$, and therefore $\frac{\psi^2}{8\pi} \geq 8c_6\cos^2\phi$. In

addition, recall that $c_4 \leq \frac{1}{4}$, $c_6 \frac{\cos^2\phi}{\cos^4(c_4\psi)} \leq c_6 \cdot \frac{\cos^2\phi}{\cos^4(\frac{\pi}{8})} \leq 4c_6\cos^2\phi$. Hence,

$$\frac{1}{\cos^2\psi} - \frac{1}{\cos^2(c_2\psi)} - c_4\frac{\cos^2\phi}{\cos^4(c_2\psi)} \geq \frac{\psi^2}{8\pi}\cdot(1-\frac{1}{2}) \geq \frac{\psi^2}{16\pi}.$$

We would also like to set $c_3 > 0$ such that

$$\exp\left(\frac{d-2}{2}\cos^2\phi\cdot\frac{\psi^2}{16\pi}\right) \geq \frac{\pi}{c_2c_4\psi}, \tag{5}$$

because this would imply that

$$\frac{\int_0^{\psi/2}F(\theta)d\theta}{\int_{\psi/2}^{\phi}F(\theta)d\theta} \geq \frac{2c_4\psi}{\pi}\cdot\exp\left(\frac{d-2}{2}\cos^2\phi\cdot\frac{\psi^2}{16\pi}\right) \geq \frac{2}{c_2},$$

which in turn implies

$$\frac{\int_0^{\psi/2}F(\theta)d\theta}{\int_0^{\phi}F(\theta)d\theta} = \frac{1}{1+\frac{\int_{\psi/2}^{\phi}F(\theta)d\theta}{\int_0^{\psi/2}F(\theta)d\theta}} = \frac{1}{1+c_2/2} \geq 1-c_2.$$

We analyze a sufficient condition for Equation (5) to hold:

$$\begin{aligned}
&\exp\left(\frac{d-2}{2}\cos^2\phi\cdot\frac{\psi^2}{16\pi}\right) \geq \frac{\pi}{c_2c_4\psi}\\
\Leftarrow\ &\frac{d-2}{2}\cos^2\phi\cdot\frac{\psi^2}{16\pi} \geq \ln\left(\frac{\pi}{c_2c_4}\cdot\frac{1}{\psi}\right)\\
\Leftarrow\ &\psi^2 \geq \frac{96\pi}{d\cos^2\phi}\ln\left(\frac{2\pi}{c_2c_4}\frac{1}{\psi^2}\right)\\
\Leftarrow\ &\psi^2 \geq \frac{192\pi}{d\cos^2\phi}\left(\ln\frac{8\pi}{c_2c_4} + \ln\left(1+\frac{96\pi}{d\cos^2\phi}\right)\right)\\
\Leftarrow\ &\psi \geq \sqrt{\frac{192\pi}{d\cos^2\phi}\left(\ln\frac{8\pi}{c_2c_4} + \ln\left(1+\frac{96\pi}{d\cos^2\phi}\right)\right)}
\end{aligned}$$

Therefore, choosing $c_3 = \max\left(16\sqrt{c_6}, 2, 1+\frac{\ln(96\pi)+\ln(\frac{2\pi}{c_4})}{\ln\frac{4}{c_2}}\right)$ (which is independent of ϕ), and by algebra, it satisfies $c_3\frac{1}{d^{\frac{1}{2}}\cos\phi}\sqrt{\ln\frac{4}{c_2}+\ln\left(1+\frac{1}{d^{\frac{1}{2}}\cos\phi}\right)} \geq \sqrt{\frac{192\pi}{d\cos^2\phi}\left(\ln\frac{8\pi}{c_2c_4}+\ln\left(1+\frac{96\pi}{d\cos^2\phi}\right)\right)}$, we have that Equation (5) is satisfied, and therefore $\Pi(\alpha) = \frac{\int_0^{\psi/2}F(\theta)d\theta}{\int_0^{\phi}F(\theta)d\theta} \geq 1-c_2$. □

Lemma 3. *For $a, b > 0$, $\zeta \in (0,1)$, if $a \geq 2b\left(\ln\frac{4}{\zeta} + \ln(1+\frac{1}{b})\right)$, then $a \geq b\ln\frac{1}{\zeta a}$.*

Proof. If $a \geq 2b\left(\ln\frac{4}{\zeta}+\ln(1+\frac{1}{b})\right) = 2b\left(\ln\frac{1}{\zeta}+\ln(4+\frac{4}{b})\right)$, then $a \geq 2b\ln\frac{1}{\zeta}$ and $a \geq 2b\ln(\max(e, \frac{1}{2b}))$ hold simultaneously.

The latter condition implies that $\frac{1}{a} \leq \frac{\frac{1}{2b}}{\ln(\max(e,\frac{1}{2b}))}$. By Lemma 4, this gives $\frac{1}{a}\ln\frac{1}{a} \leq \frac{1}{2b}$, in other words, $a \geq 2b\ln\frac{1}{a}$.

Now combine this with $a \geq 2b\ln\frac{1}{\zeta}$ by taking average on both sides, we get $a \geq \frac{1}{2}(2b\ln\frac{1}{\zeta}+2b\ln\frac{1}{a}) = b\ln\frac{1}{a\zeta}$. The lemma follows. □

Lemma 4. *For $y > 0$, and $x \leq \frac{y}{\ln(\max(e,y))}$, then $x\ln x \leq y$.*

Proof. Define $x_0 := \frac{y}{\ln(\max(e,y))}$. We first verify that $x_0 \ln x_0 \le y$.

1. If $y \le e$, then $x_0 = y$; in this case, $x_0 \ln x_0 = y \ln y \le y$ holds.
2. Otherwise, $y > e$. In this case, $x_0 = \frac{y}{\ln y} \le y$. Therefore, $x_0 \ln x_0 \le x_0 \ln y = y$.

Now, given $x \le x_0$, we consider two cases of x:

1. If $x \le \frac{1}{e}$, then $x \ln x < 0 < y$ holds.
2. Otherwise, $x > \frac{1}{e}$, and since $f(x) = x \ln x$ is monotonically increasing in $(\frac{1}{e}, +\infty)$, we have that $x \ln x \le x_0 \ln x_0 \le y$.

In summary, if $x \le x_0$, we must have $x \ln x \le y$. □

Lemma 5. *For any $c_5 > 0$, there exists $c_6 > 0$ such that*

$$1 - x \ge \exp(-x - c_6 x^2), \quad \forall x \in [0, 1 - c_5].$$

Proof. It suffices to choose $c_6 > 0$ such that

$$-\ln(1-x) \le x + c_6 x^2, \quad \forall x \in [0, 1 - c_5].$$

By Taylor's expansion,

$$\begin{aligned} -\ln(1-x) =& x + \sum_{i=2}^{\infty} \frac{x^i}{i} \\ \le& x + \frac{x^2}{2} \left(\sum_{i=0}^{\infty} x^i \right) \\ \le& x + \frac{x^2}{2(1-x)}, \end{aligned}$$

therefore, it suffices to choose $c_6 = \frac{1}{2c_5}$ such that the above is at most $x + c_6 x^2$ for all $x \in [0, 1 - c_5]$. □

Lemma 6. *For $x \ge 0$,* $\arctan(x) \ge \min(\frac{\pi}{4}, \frac{x}{2})$.

Proof. We consider two cases:

1. If $x \ge 1$, $\arctan(x) \ge \frac{\pi}{4} \ge \min(\frac{\pi}{4}, \frac{x}{2})$.
2. If $x < 1$, by mean value theorem, there exists some $\xi \in [0, x]$, such that $\arctan(x) = 0 + x \cdot (\arctan(z))' \big|_{z=\xi} = \frac{x}{1+\xi^2} \ge \frac{x}{2} \ge \min(\frac{\pi}{4}, \frac{x}{2})$.

The lemma follows by combining the two cases. □

A.2 Section 5 Proofs

In this section, we provide complementary negative results to the positive results obtained under the assumptions that: 1) $\mathcal{X}$ is a sphere; and 2) $\mathcal{U}$ is the uniform distribution over $\mathcal{H}$, the class of homogeneous linear models. We show that removing one of the two conditions, i.e either allowing for non-spherical features (Proposition 1) or allowing $\mathcal{U}$ to be non-uniform over $\mathcal{H}$ (Proposition 2), leads to non-monotonicity.

Proposition 1. *Suppose $d = 2$. We have uniform prior over homogeneous linear models $\mathcal{H} = \{h_w \mid w \in \mathbb{R}^d, \|w\| = 1\}$, there exists a feature space $\mathcal{X}$ and thresholds $0 < \alpha_2 < \alpha_1$ such that $\Pi(\alpha_2) < \Pi(\alpha_1)$.*

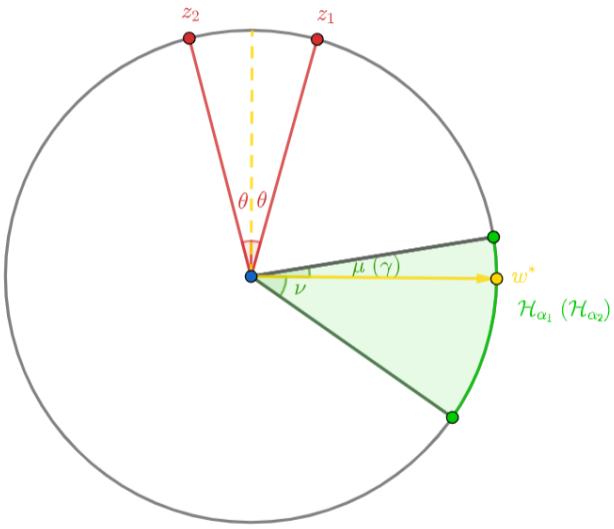

Figure 8: An illustration of $\mathcal{H}_{\alpha_1}$ and $\mathcal{H}_{\alpha_2}$ in the proof of Proposition 1.

Proof. Define $\mathcal{X} = \{x^1, x^2, x^3, z^1, z^2\}$, with the choices of x^1, x^2, x^3, z^1, z^2 specified shortly.

Let $w^\star = (1, 0)$, and therefore $h^\star((x_1, x_2)) = \text{sign}(x_1)$. Let $\theta \in (0, \frac{\pi}{4})$ be an angle. Define $z^1 = (\frac{r}{2}\sin\theta, \frac{r}{2}\cos\theta)$, $z^2 = (-\frac{r}{2}\sin\theta, \frac{r}{2}\cos\theta)$; it can be readily seen that $\|z^1 - z^2\| \le r$ and $\text{sign}(h^\star(z^1)) = +1 \neq -1 = \text{sign}(h^\star(z^2))$; therefore $(z^1, z^2) \in \mathcal{M}_r(\mathcal{X})$. As we will see shortly, this is the only pair in $\mathcal{M}_r(\mathcal{X})$ up to reordering.

Let α_1', α_2' be such that $0 < r < \alpha_2' < \alpha_1'$, and angles γ, μ, ν be such that $\gamma < \mu < \theta < \nu$, and $\theta + \nu < \frac{\pi}{2}$. Define $x^1 = (\alpha_1', -\alpha_1' \cot\mu)$, $x^2 = (\alpha_1', \alpha_1' \cot\nu)$, and $x^3 = (\alpha_2', -\alpha_2' \cot\gamma)$. It can be seen that $h^*(x^1) = h^*(x^2) = h^*(x^3) = +1$; in addition, note that all of $\|x^1 - z^2\|$, $\|x^2 - z^2\|$, $\|x^3 - z^2\|$ are $> r$, ensuring that $\mathcal{M}_r(\mathcal{X}) = \{(z^1, z^2), (z^2, z^1)\}$.

Let $\alpha_2 = \alpha_2'/2$ and $\alpha_1 = (\alpha_1' + \alpha_2')/2$. Observe that $\{x \in \mathcal{X} : \Lambda_{\alpha_1}(x) = 1\} = \{x^1, x^2\}$, and $\{x \in \mathcal{X} : \Lambda_{\alpha_2}(x) = 1\} = \{x^1, x^2, x^3\}$.

Numerical Example. For concreteness, we can take $\alpha_1' = 10$, $\alpha_2' = 5$, $\alpha_1 = 7.5$, $\alpha_2 = 2.5$, $r = 1$, $\gamma = \frac{\pi}{16}$, $\mu = \frac{\pi}{12}$, $\theta = \frac{\pi}{8}$, and $\nu = \frac{\pi}{4}$, which satisfy all requirements above.

Given $w = (w_1, w_2) \in \mathbb{R}^2$, denote by $\phi(w) \in (-\pi, \pi]$ its polar angle with respect to $(1, 0)$ (so that $\phi((1, 0)) = 0$).

We now calculate $\Pi(\alpha_1)$. First, observe that

$$\mathcal{H}_{\alpha_1} = \Big\{h \in \mathcal{H} : h(x^1) = 1, h(x^2) = 1\Big\} = \big\{h_w : \|w\|_2 = 1, \phi(w) \in [-\nu, \mu]\big\}$$

Therefore,

$$\begin{aligned}
\Pi(\alpha_1) &= \max_{(x,x') \in \mathcal{M}_r(\mathcal{X})} \Big(\mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_1})}(\langle w, x\rangle \ge 0) - \mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_1})}(\langle w, x'\rangle \ge 0)\Big) \\
&= \Big|\mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_1})}(\langle w, z^1\rangle \ge 0) - \mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_1})}(\langle w, z^2\rangle \ge 0)\Big| \\
&= \left|\frac{\mu + \theta}{\mu + \nu} - 0\right| = \frac{\mu + \theta}{\mu + \nu}.
\end{aligned}$$

We now calculate $\Pi(\alpha_2)$. First observe that

$$\mathcal{H}_{\alpha_2} = \Big\{h \in \mathcal{H} : h(x^1) = 1, h(x^2) = 1, h(x^3) = 1\Big\} = \big\{h_w : \|w\|_2 = 1, \phi(w) \in [-\nu, \gamma]\big\}$$

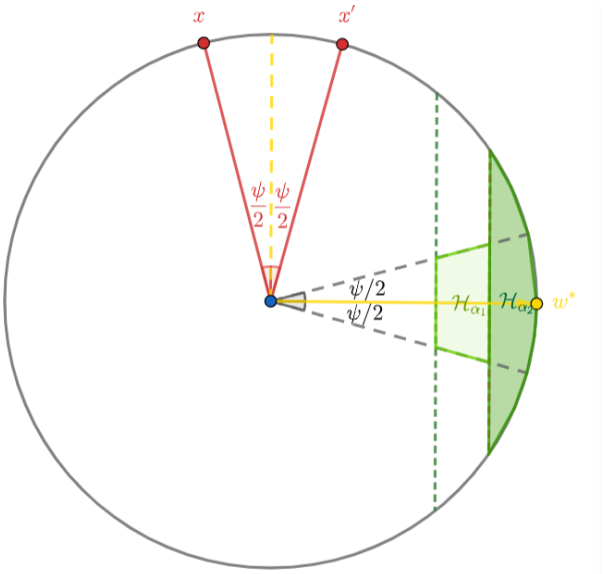

Figure 9: In the proof of Proposition 2, a projection of $\mathcal{U}$ onto the 2-dimensional plane spanned by w^*, x and x'; it is uniform when restricted to $\mathcal{H}_{\alpha_2}$ (the dark green region), and is concentrated in $\{h_w \in \mathcal{H}_{\alpha_1} \backslash \mathcal{H}_{\alpha_2} : -1 = \text{sign}(w \cdot x) \neq \text{sign}(w \cdot x') = +1\}$ (the light green region) when restricted to $\mathcal{H}_{\alpha_1} \setminus \mathcal{H}_{\alpha_2}$.

Therefore,

$$\begin{aligned}
\Pi(\alpha_2) &= \max_{(x,x') \in \mathcal{M}_r(\mathcal{X})} \left(\mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_2})}(\langle w, x \rangle \geq 0) - \mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_2})}(\langle w, x' \rangle \geq 0) \right) \\
&= \left| \mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_2})}(\langle w, z^1 \rangle \geq 0) - \mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_2})}(\langle w, z^2 \rangle \geq 0) \right| \\
&= \left| \frac{\gamma + \theta}{\gamma + \nu} - 0 \right| = \frac{\gamma + \theta}{\gamma + \nu}.
\end{aligned}$$

In conclusion,

$$\Pi(\alpha_1) = \frac{\mu + \theta}{\mu + \nu} \geq \frac{\gamma + \theta}{\gamma + \nu} = \Pi(\alpha_2). \qquad \square$$

Proposition 2. *Suppose $\mathcal{X}$ is the d-dimensional unit sphere with $d \geq 3$. There exists a non-uniform distribution $\mathcal{U}$ over homogeneous linear models $\mathcal{H}$, such that there exists thresholds $0 < \alpha_2 < \alpha_1$ with $\Pi(\alpha_2) < \Pi(\alpha_1)$.*

Proof. WLOG, we assume that $w^* = (1, 0, \ldots, 0)$. Define $x = (-\sin(\psi/2), \cos(\psi/2), 0, \ldots, 0)$ and $x' = (\sin(\psi/2), \cos(\psi/2), 0, \ldots, 0)$ which will be used later. It can be seen that x, x' and w^* are on the same 2-dimensional plane.

Let α_2, α_1 be such that $0 < \alpha_2 < \alpha_1 < 1$ and with $\phi_1 = \arcsin \alpha_1$ and $\phi_2 = \arcsin \alpha_2$, $\phi_1 > \phi_2 > \psi/2$. We know from Lemma 1 that

$$\mathcal{H}_{\alpha_2} = \left\{ h_w : \|w\|_2 = 1, \langle w, w^* \rangle \geq \sqrt{1 - \alpha_2^2} \right\} \subset \mathcal{H}_{\alpha_1} = \left\{ h_w : \|w\|_2 = 1, \langle w, w^* \rangle \geq \sqrt{1 - \alpha_1^2} \right\},$$

and that $\mathcal{H}_{\alpha_1} \backslash \mathcal{H}_{\alpha_2} = \left\{ h_w : \|w\|_2 = 1, \langle w, w^* \rangle \in [\sqrt{1 - \alpha_1^2}, \sqrt{1 - \alpha_2^2}) \right\}$.

We define the density of the non-uniform prior $\mathcal{U}$ as follows. Let $\mathcal{U}$ be uniform when restricted to $\mathcal{H}_{\alpha_2}$. And let $\mathcal{U}$ have positive density that is uniform over $\{h_w : w \in \mathcal{H}_{\alpha_1} \backslash \mathcal{H}_{\alpha_2}, -1 = \text{sign}(w \cdot x) \neq \text{sign}(w \cdot x') = +1\}$; note that this is an non-empty set as it comprises of all w's whose projection onto w^* has value in $[\sqrt{1 - \alpha_1^2}, \sqrt{1 - \alpha_2^2}]$ and has polar angle wrt w^* in $[-\psi/2, \psi/2]$. Finally, let $\mathcal{U}$ have zero density over all other parts of $w \in \mathcal{H}_{\alpha_1} \backslash \mathcal{H}_{\alpha_2}$. The density of $\mathcal{U}$ outside $\mathcal{H}_{\alpha_1}$ can be chosen arbitrarily. See Figure 9 for an illustration.

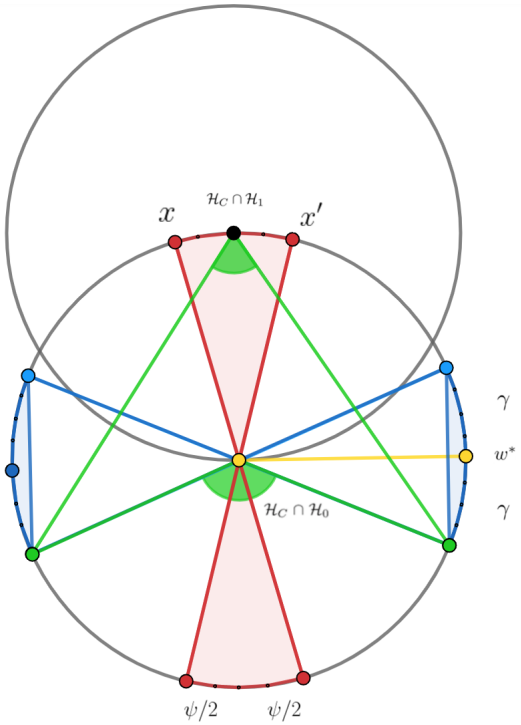

Figure 10: The construction in Proposition 3. In blue are the explanations, in green are the decision boundaries of models in the version space, in red is the margin region and in yellow is w^*.

By the definition of x, x', and the fact that $\mathcal{U}$ is uniform when restricted to $\mathcal{H}_{\alpha_2}$, from the proof of Theorem 1, $(x, x') \in \arg\max_{(x,x') \in \mathcal{M}_r(\mathcal{X})} \pi_{\alpha_2}(x, x')$; in other words, $\pi_{\alpha_2}(x, x') = \Pi(\alpha_2)$.

With this, we know that since $\phi_0 > \psi$, $\Pi(\alpha_2) = \pi_{\alpha_2}(x, x') < 1$. Then,

$$\begin{aligned}
\Pi(\alpha_1) \geq \pi_{\alpha_1}(x, x') &= \mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_1})}(\langle w, x' \rangle \geq 0) - \mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_1})}(\langle w, x \rangle \geq 0) \\
&= \Big(\mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_2})}(\langle w, x' \rangle \geq 0) - \mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_2})}(\langle w, x \rangle \geq 0)\Big) \cdot \mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_1})}(w \in \mathcal{H}_{\alpha_2}) + \\
&\quad \Big(\mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_1} \setminus \mathcal{H}_{\alpha_2})}(\langle w, x' \rangle \geq 0) - \mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_1} \setminus \mathcal{H}_{\alpha_2})}(\langle w, x \rangle \geq 0)\Big) \cdot \mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_1})}(w \in \mathcal{H}_{\alpha_1} \setminus \mathcal{H}_{\alpha_2}) \\
&= \pi_{\alpha_2}(x, x') \cdot \mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_1})}(w \in \mathcal{H}_{\alpha_2}) + \mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_1})}(w \in \mathcal{H}_{\alpha_1} \setminus \mathcal{H}_{\alpha_2}) \\
&> \pi_{\alpha_2}(x, x') = \Pi(\alpha_2),
\end{aligned}$$

where the first inequality is from the definition of $\Pi(\alpha_1)$; the first equality is by the definition of $\pi(\alpha_1)$; the second equality is by the total law of probability; the third equality is by the construction that $\mathcal{U}$ has zero density in $\{h_w : w \in \mathcal{H}_{\alpha_1} \setminus \mathcal{H}_{\alpha_2}, \text{sign}(w \cdot x) = +1 \vee \text{sign}(w \cdot x') = -1\}$, so that $\mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_1} \setminus \mathcal{H}_{\alpha_2})}(\langle w, x' \rangle \geq 0) = 1$ and $\mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_1} \setminus \mathcal{H}_{\alpha_2})}(\langle w, x \rangle \geq 0) = 0$, along with the definition of $\pi_{\alpha_2}(x, x')$; the last inequality is strict because $\mathbb{P}_{h_w \sim \mathcal{U}(\mathcal{H}_{\alpha_1})}(w \in \mathcal{H}_{\alpha_1} \setminus \mathcal{H}_{\alpha_2}) > 0$ and that $\Pi(\alpha_2) = \pi_{\alpha_2}(x, x') < 1$. □

Lastly, fixing assumptions 1 and 2, one may also wonder if it is possible to achieve any threshold κ in the more general, non-homogeneous linear models. We saw that this is not so asymptotically in the homogeneous case (Theorem 3). Here, we demonstrate that this does not hold in general.

Proposition 3. *There exists a class of 2-dimensional non-homogeneous linear models, with spherical $\mathcal{X}$ such that $\Pi(\alpha)$ decreases monotonically (and strictly so at some point) with increasing α, and yet $\Pi(\alpha) \geq 1/3$ for all $\alpha \in [0, 1)$ and $\psi \in (0, \pi]$.*

Proof. Let the hypothesis class of interest be $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_0$, where

$$\mathcal{H}_0 = \{x \mapsto \text{sign}(w_1 x_1 + w_2 x_2) : \|w\|_2 = 1\}$$

is its homogeneous part, and

$$\mathcal{H}_1 = \left\{x \mapsto \text{sign}(w_1 x_1 + w_2(x_2 - 1)) : \|w\|_2 = 1\right\}$$

is its non-homogeneous part.

We will take same setting as before $\mathcal{X}$ is a unit circle centered at $(0,0)$ and $\mathcal{E}_{h^*}(\mathcal{X}, \alpha) = \{x \in \mathcal{X} \mid \Lambda_\alpha(x) = 1\}$. We assume an uniform prior $\mathcal{U}$ over $\mathcal{H}$, i.e. drawing $i \sim \text{Bernoulli}(\frac{1}{2})$, and chooses a classifier uniformly at random from $\mathcal{H}_i$ induces $\mathcal{U}$.

Let $h^*(x) = x \mapsto \text{sign}(x_1)$, which is a member of $\mathcal{H}$. We consider a boundary pair $(x, x') \in \mathcal{M}_r(\mathcal{X})$ where $\|x' - x\|_2 \le r$, $h^*(x') = +1 \ne -1 = h^*(x)$.

Given $w = (w_1, w_2) \in \mathbb{R}^2$, denote by $\phi(w) \in (-\pi, \pi]$ its polar angle with respect to $(1,0)$ (so that $\phi((1,0)) = 0$).

Given a value of $\alpha \in [0, 1)$, the induced explanation set

$$\mathcal{E}_{h^*}(\mathcal{X}, \alpha) = \left\{x \in \mathcal{X} : \phi(x) \in [-\pi, -\pi + \gamma) \cup (-\gamma, \gamma) \cup (\pi - \gamma, \pi]\right\},$$

with $\gamma = \arccos \alpha \in (0, \frac{\pi}{2}]$.

We will examine the structure of version space $\mathcal{H}_C$ and count how much of it predicts (x, x') differently. Please refer to Figure 10 for an illustration. We will look at $\mathcal{H}_C \cap \mathcal{H}_1$ and $\mathcal{H}_C \cap \mathcal{H}_0$ respectively.

Part 1: $\mathcal{H}_C \cap \mathcal{H}_1$. For any $h \in \mathcal{H}_C \cap \mathcal{H}_1$, it always holds that $h(x) = +1$ and $h(x') = -1$ as long as $\gamma > 0$. This is because if the explanation is nonempty, then it includes points $(-1, 0)$ and $(1, 0)$, which enforces that any $h \in \mathcal{H}_C \cap \mathcal{H}_1$ must be a subset of $h \in \mathcal{H}_1$ with polar angle in interval $[-\pi/4, \pi/4]$ and all such h's predict (x, x') differently. More specifically,

$$\mathcal{H}_C \cap \mathcal{H}_1 = \left\{x \mapsto \text{sign}(w_1 x_1 + w_2(x_2 - 1)) : \|w\|_2 = 1, \phi(w) \in \left[-(\frac{\pi}{4} - \frac{\gamma}{2}), \frac{\pi}{4} - \frac{\gamma}{2}\right]\right\},$$

whose total arc length of $\frac{\pi}{2} - \gamma$. To summarize,

$$\mathbb{P}_{h \sim \mathcal{U}}\left(h \in \mathcal{H}_C \cap \mathcal{H}_1\right) = \frac{1}{2} \cdot \frac{\frac{\pi}{2} - \gamma}{2\pi} = \frac{\frac{\pi}{2} - \gamma}{4\pi},$$

and

$$\mathbb{P}_{h \sim \mathcal{U}(\mathcal{H}_C \cap \mathcal{H}_1)}(h(x') = +1) - \mathbb{P}_{h \sim \mathcal{U}(\mathcal{H}_C \cap \mathcal{H}_1)}(h(x) = +1) = 1.$$

Part 2: $\mathcal{H}_C \cap \mathcal{H}_0$. As we showed in Lemma 1,

$$\mathcal{H}_0 = \left\{x \mapsto \text{sign}(w_1 x_1 + w_2 x_2) : \|w\|_2 = 1, \phi(w) \in \left[-(\frac{\pi}{2} - \gamma), \frac{\pi}{2} - \gamma\right]\right\},$$

whose total arc length is $\pi - 2\gamma$.

In addition, by Theorem 1 with $d = 2$ with $\phi = \frac{\pi}{2} - \gamma$, we have

$$\max_{(x,x') \in \mathcal{M}_r(\mathcal{X})} \left(\mathbb{P}_{h \sim \mathcal{U}(\mathcal{H}_C \cap \mathcal{H}_0)}(h(x') = +1) - \mathbb{P}_{h \sim \mathcal{U}(\mathcal{H}_C \cap \mathcal{H}_0)}(h(x) = +1)\right) = \begin{cases} \frac{\psi}{2(\frac{\pi}{2} - \gamma)} & \psi \le 2(\frac{\pi}{2} - \gamma), \\ 1 & \psi > 2(\frac{\pi}{2} - \gamma). \end{cases}$$

To summarize,

$$\mathbb{P}_{h \sim \mathcal{U}}\left(h \in \mathcal{H}_C \cap \mathcal{H}_0\right) = \frac{2(\frac{\pi}{2} - \gamma)}{4\pi}$$

which is twice $\mathbb{P}_{h \sim \mathcal{U}}\left(h \in \mathcal{H}_C \cap \mathcal{H}_1\right)$ and,

$$\max_{(x,x') \in \mathcal{M}_r(\mathcal{X})} \left(\mathbb{P}_{h \sim \mathcal{U}(\mathcal{H}_C \cap \mathcal{H}_0)}(h(x') = +1) - \mathbb{P}_{h \sim \mathcal{U}(\mathcal{H}_C \cap \mathcal{H}_0)}(h(x) = +1)\right) = \min\left(1, \frac{\psi}{2(\frac{\pi}{2} - \gamma)}\right)$$

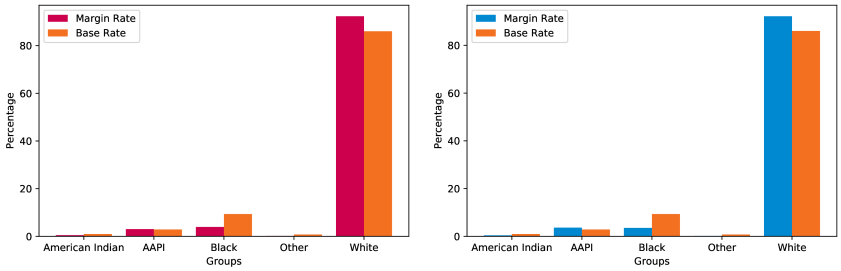

Figure 11: Racial composition of margin points under LR (left) and SVM (right).

Combining the two parts, observe that $\mathbb{P}_{h\sim\mathcal{U}(\mathcal{H}_C)}(h \in \mathcal{H}_C \cap \mathcal{H}_0) = \frac{2}{3}$, and by the law of total probability,

$$\begin{aligned}
\Pi(\alpha) =& \max_{(x,x')\in\mathcal{M}_r(\mathcal{X})} \pi_\alpha(x, x') \\
=& \max_{(x,x')\in\mathcal{M}_r(\mathcal{X})} \Big(\mathbb{P}_{h\sim\mathcal{U}(\mathcal{H}_C)}(h(x) = +1) - \mathbb{P}_{h\sim\mathcal{U}(\mathcal{H}_C)}(h(x') = +1)\Big) \\
=& \max_{(x,x')\in\mathcal{M}_r(\mathcal{X})} \Bigg(\mathbb{P}_{h\sim\mathcal{U}(\mathcal{H}_C)}(h \in \mathcal{H}_C \cap \mathcal{H}_0) \cdot \Big(\mathbb{P}_{h\sim\mathcal{U}(\mathcal{H}_C\cap\mathcal{H}_0)}(h(x') = +1) - \mathbb{P}_{h\sim\mathcal{U}(\mathcal{H}_C\cap\mathcal{H}_0)}(h(x) = +1)\Big) \\
& + \mathbb{P}_{h\sim\mathcal{U}(\mathcal{H}_C)}(h \in \mathcal{H}_C \cap \mathcal{H}_1) \cdot \Big(\mathbb{P}_{h\sim\mathcal{U}(\mathcal{H}_C\cap\mathcal{H}_1)}(h(x') = +1) - \mathbb{P}_{h\sim\mathcal{U}(\mathcal{H}_C\cap\mathcal{H}_1)}(h(x) = +1)\Big)\Bigg) \\
=& \frac{2}{3} \cdot \min\left(1, \frac{\psi}{2(\frac{\pi}{2} - \gamma)}\right) + \frac{1}{3} \\
\geq& \frac{1}{3}
\end{aligned}$$

through which we see that $\Pi(\alpha)$ is increasing in γ and strictly so for when $\pi/2 - \gamma > \psi/2$. In other words, $\Pi(\alpha)$ is identically 1 for $\alpha \in [0, \sin(\psi/2)]$, and is strictly decreasing in α for $\alpha \in [\sin(\psi/2), 1)$. □

B Additional Experiments

B.1 Fair accessibility to explanations

A notable concern that may arise with margin distancing is that omission of prototypical explanations is necessary for regions close to the margin. Thus, this could disproportionately affect individuals in those regions, since they will not have their representative explanation be in the explanation set. We plot the composition of margin set in Figure 11 with a threshold of 0.03 for both logistic and SVM models and note that there is some disproportionate effect. Verily, this is another important factor that needs to be taken into account in the explanation generation process.

B.2 MMD Explanations

We include results on the trend of the three metrics under MMD-Critic explanations to further empirically trace how the boundary certainty varies with explanation omission. Similar to the MLP results under k-medoid, we see that in Figure 12 the trend is almost monotonic everywhere. One difference however, is that the boundary certainty does not drop off as fast as in the k-medoid setting. This suggests that the search strategy of trying small omission percentages may work with some explanation methods such as the k-medoid, but will not with others like MMD-Critic.

B.3 Effects of Larger Models

We include results on the trend of the three metrics for a two hidden-layer MLP to showcase the effects of larger models. In Figure 13, we see similar trends under both explanations, but with higher values across the board in

Target Certainty	Binary Search	Optimal	Difference
0.036	45	10	35
0.046	45	10	35
0.055	10	10	0
0.065	10	10	0
0.075	10	10	0
0.084	10	10	0
0.094	5	5	0
0.103	5	5	0
0.113	5	5	0
0.122	5	5	0

Table 2: Difference table with the max metric and at $r = 0.1$

Target Certainty	Binary Search	Optimal	Difference
0.071	70	15	55
0.11	45	10	35
0.15	10	10	0
0.18	10	10	0
0.22	10	10	0
0.26	5	5	0
0.30	5	5	0
0.33	5	5	0
0.37	5	5	0
0.41	5	5	0

Table 3: Difference table with the max metric and at $r = 0.2$

comparison with the one-layer case. Again, as in the one-layer MLP case, under MMD-critic explanations, the drop in the metrics are slower than the drop under k-medoid explanations.

B.4 Monotonicity Tables

We present tables charting the differences between the percentage of explanations omitted calculated through binary search and the optimal percentage of explanation calculated through a left-to-right linear search, for ten, equally spaced out values of target boundary certainty corresponding to Figure 3 in Tables 2 through 10.

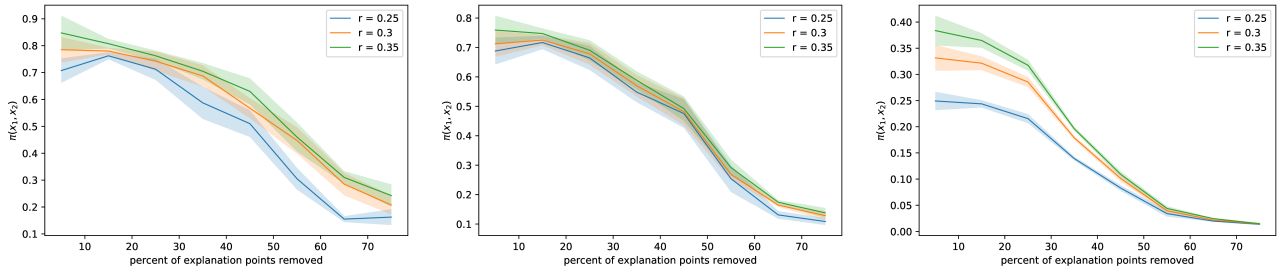

Figure 12: MLPs results with MMD-Critic explanations: max (left), top 5 percentile average (middle), average $\pi(x, x')$ (right). We observe similar trends as in the k-medoid case with one difference being that the drop off rate is slower in the MMD-Critic case.

Target Certainty	Binary Search	Optimal	Difference
0.16	65	65	0
0.23	25	25	0
0.31	10	10	0
0.39	5	5	0
0.47	5	5	0
0.55	5	5	0
0.63	5	5	0
0.7	5	5	0
0.78	5	5	0
0.86	5	5	0

Table 4: Difference table with the max metric and at $r = 0.3$

Target Certainty	Binary Search	Optimal	Difference
0.03	45	10	35
0.04	45	10	35
0.05	10	10	0
0.06	10	10	0
0.07	10	10	0
0.08	10	10	0
0.09	5	5	0
0.1	5	5	0
0.11	5	5	0
0.12	5	5	0

Table 5: Difference table with the top 5 percentile average and at $r = 0.1$

Target Certainty	Binary Search	Optimal	Difference
0.05	65	15	50
0.07	40	10	30
0.1	10	10	0
0.12	10	10	0
0.14	10	10	0
0.17	5	5	0
0.19	5	5	0
0.21	5	5	0
0.24	5	5	0
0.26	5	5	0

Table 6: Difference table with the top 5 percentile average and at $r = 0.2$

Target Certainty	Binary Search	Optimal	Difference
0.11	65	65	0
0.17	25	25	0
0.23	10	10	0
0.3	10	10	0
0.36	5	5	0
0.42	5	5	0
0.48	5	5	0
0.54	5	5	0
0.6	5	5	0
0.66	5	5	0

Table 7: Difference table with the top 5 percentile average and at $r = 0.3$

Target Certainty	Binary Search	Optimal	Difference
0.008	45	15	30
0.014	45	10	35
0.019	40	10	30
0.025	10	10	0
0.031	5	5	0
0.037	5	5	0
0.042	5	5	0
0.048	5	5	0
0.054	5	5	0
0.06	5	5	0

Table 8: Difference table with the average and at $r = 0.1$

Target Certainty	Binary Search	Optimal	Difference
0.013	65	10	55
0.02	45	10	35
0.027	10	10	0
0.034	10	10	0
0.041	5	5	0
0.049	5	5	0
0.056	5	5	0
0.063	5	5	0
0.07	5	5	0
0.077	5	5	0

Table 9: Difference table with the average and at $r = 0.2$

Target Certainty	Binary Search	Optimal	Difference
0.044	65	65	0
0.075	40	30	10
0.106	10	10	0
0.137	10	10	0
0.168	5	5	0
0.199	5	5	0
0.229	5	5	0
0.26	5	5	0
0.291	5	5	0
0.322	5	5	0

Table 10: Difference table with the average and at $r = 0.3$

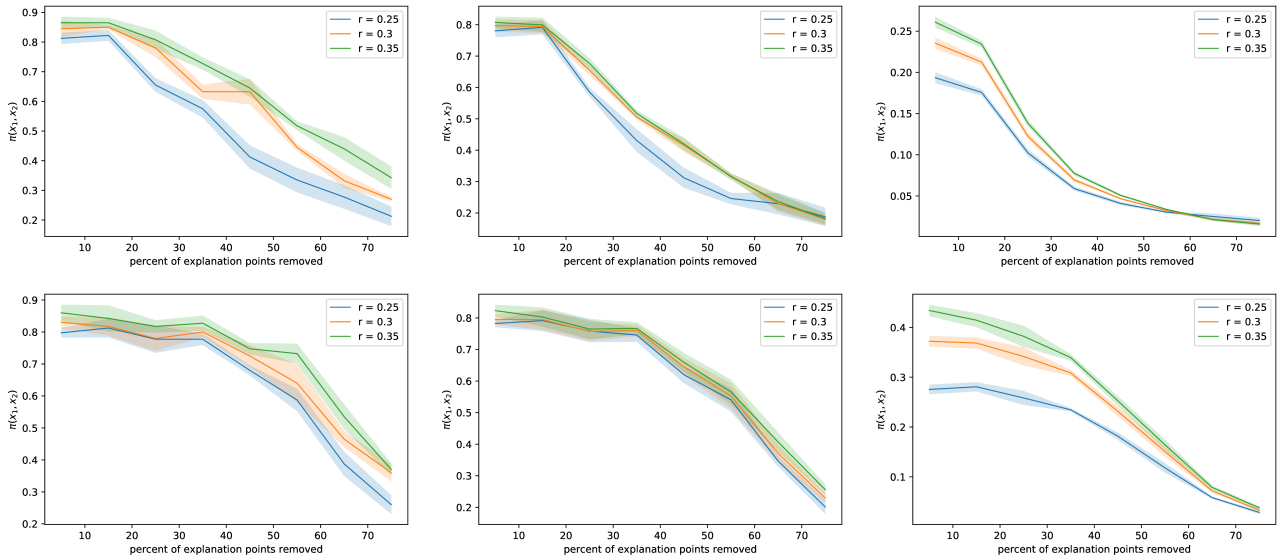

Figure 13: Two layer MLP results: under k-medoid explanations (top), under MMD explanations (bottom). The three metrics are in column: max (left), top 5 percentile average (middle), average $\pi(x, x')$ (right).

C Additional Modeling Discussion

One objection with our modeling assumption could be that if it is the case that most of the $\mathcal{X}$ is in $\mathcal{M}_r(\mathcal{X})$, then margin-distancing could remove most of the representative-based explanations $\mathcal{E}_{h^*}(\mathcal{X})$. We assume this is not the case and that $\mathcal{M}_r(\mathcal{X})$ is only a small fraction of $\mathcal{X}$.

Indeed, this assumes that the feature collection and modeling is done well and that most points are not within r of another point with the opposite label.

D Additional Related Works

Improvement vs Gaming: A crucial point about feature alteration is whether to think of it as causal (beneficial) or gaming (Miller et al., 2020). In our setting, the organization first offers individuals transparency into how the model "works" and predicts based on the reported features. We assume individuals are not aware of the underlying causal model. Hence, we view misreporting in the first stage as gaming.

Explanation Manipulation: There has been work focusing on how organizations may manipulate an unfair model's explanation to make it look more fair than it actually is Aïvodji et al. (2019); Anders et al. (2020); Slack et al. (2020). By contrast, we study how to provide explanations that are informative and cover as much of $\mathcal{X}$ as possible while protecting boundary points' label information.

Security of ML models: Our work is also related to model extraction literature Tramèr et al. (2016); Milli et al. (2019) that assumes one can query an API for model prediction/gradient-based explanation on any point. We view our work as a study on how to "limit" the API so as to prevent a new type of attack – individual-level gaming, which need not require the full model extraction in order to carry out the attack Jagielski et al. (2020).

Model Multiplicity: The set of models consistent with labelled data is also referred to as version space Mitchell (1977). Our paper thus pertains to a recent line of work highlighting the existence of the "Rashomon effect" Semenova et al. (2019); D'Amour et al. (2020) or model multiplicity Marx et al. (2020). These papers do not focus on strategic manipulation, but study or raise the importance of developing sampling algorithms that can explore the version space.