Estimating the Arc Length of the Optimal ROC Curve and Lower Bounding the Maximal AUC

Song Liu*
School of Mathematics
University of Bristol
Bristol, BS8 1UG, UK
song.liu@bristol.ac.uk

Abstract

In this paper, we show the arc length of the optimal ROC curve is an f-divergence. By leveraging this result, we express the arc length using a variational objective and estimate it accurately using positive and negative samples. We show this estimator has a non-parametric convergence rate $O_p(n^{-\beta/4})$ ($\beta \in (0, 1]$ depends on the smoothness). Using the same technique, we show the surface area sandwiched between the optimal ROC curve and the diagonal can be expressed via a similar variational objective. These new insights lead to a novel two-step classification procedure that maximizes an approximate lower bound of the maximal AUC. Experiments on CIFAR-10 datasets show the proposed two-step procedure achieves good AUC performance in imbalanced binary classification tasks while being less computationally demanding than the classic AUC maximizer **in the offline setting**.

1 Introduction

The study of Receiver operating characteristic (ROC) curves has a long history in medicine [24], psychology [13] and radiology [14, 11]. In machine learning, ROC curves have been primarily used to analyze the performance of different classification algorithms [6, 7]. Indeed, the Area Under the Curve (AUC) encodes the ranking accuracy of a classifier, which makes it a preferable performance metric for imbalanced class classification [6, 2]. In recent years, ROC curves have also been used in applications comparing two distributions and achieves promising results. Examples include analyzing the mode collapsing issue of Generative Adversarial nets (GAN) [23], and diagnosing the performance of an amortized Markov Chain Monte Carlo [16].

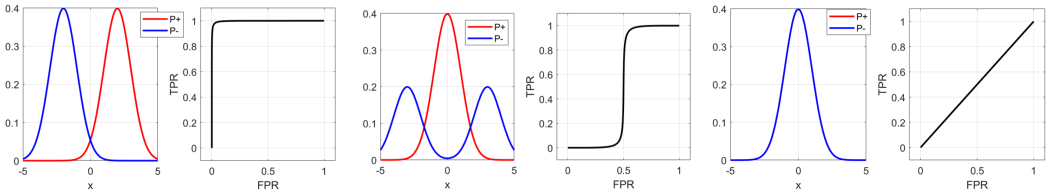
In applications that require computing statistical discrepancy between distributions (e.g. GAN [12] or Variational Inference (VI) [1]), f-divergences are a widely used discrepancy measures. The family of f-divergences includes Kullback-Leibler divergence [21] and Total Variational distance. It has been shown that f-divergences in general can be expressed via variational objectives and can be efficiently approximated from empirical samples [27, 28].

One might question that as the ROC curves are used as performance metrics in many two sample applications, are they in any way related to f-divergences? For example, can AUC be an f-divergence between positive and negative data distributions given some classification score function? An earlier investigation proves that the answer is no when the score function is the likelihood ratio [29]. Nonetheless, this result inspired us to look for f-divergence from other geometries of the ROC curve.

In this paper we show that, when using the likelihood ratio score, a novel f-divergence arises from computing the *arc length* of the corresponding ROC curve. By leveraging this result, we can

*Webpage: https://allmodelsarewrong.net

36th Conference on Neural Information Processing Systems (NeurIPS 2022).



(a) different distributions, long ROC (b) different distributions, long ROC (c) same distribution, short ROC

Figure 1: ROC curves generated for one dimensional datasets using the identity classification score function $t(x) = x$. Notice that the arc length of ROC curves seem to be a good indication on how different the positive (red) and negative (blue) data distributions are.

express the arc length using a variational objective and approximate it using only samples from two distributions. We show this arc length estimator is also a consistent estimator to the arctangent of likelihood ratio and has a non-parametric convergence rate $O_p(n^{-\beta/4})$, where $\beta \in (0, 1]$ depends on the smoothness of the true arctangent likelihood ratio. Moreover, by parameterizing the ROC curves of *mixtures* of positive and negative distributions, the surface area sandwiched between the optimal ROC curve and the diagonal can be expressed via a similar variational objective. With the help of our arctangent ratio estimator, we can approximately maximize a lower bound to this surface area. We point out the similarity between this lower bound maximization and the classic AUC maximization [2] and show our approximated optimal score achieves comparable performance to an offline AUC maximizer in an imbalanced classification problem while being less computationally demanding.

2 Background

2.1 An Illustrative Example

ROC curves are frequently used to visualize binary classification performance, and we often rely on the Area Under the Curve (AUC) as a numerical metric for selecting a good classifier. In this section, we highlight an often overlooked ROC geometry: arc length. We illustrate its potential as a good discrepancy measure between positive and negative data distributions.

Let us consider the one-dimensional distributions and the ROC curves in Figure 1. In this example, ROC curves are generated using the identity score function $t(x) = x$.

In both (a) and (b), p_+ and p_- are quite different, and thus the discrepancies between positive and negative distributions should be large in both cases. In (c), the densities p_+ and p_- are the same and consequently their discrepancy should be smaller than that of both (a) and (b). Further, notice that both (a) and (b) have long ROC curves (≈ 2) while (c) has a shorter one ($\approx \sqrt{2}$). This example, therefore, suggests that the more similar p_+ and p_- are, the shorter the ROC curve is.

However, we can see that $t(x) = x$ is a special choice: If $t(x) = 0$, the arc length will not reflect any discrepancy between data distributions at all. This simple observation inspires the following questions: Why is the arc length in this example good at telling the differences between two data distributions? Are there other score functions whose ROC arc lengths are good discrepancy measures between positive and negative data distributions? What are the practical applications of studying the arc length of ROC? In the following sections, we study the arc length of a ROC curve under a probabilistic framework and provide answers to these questions.

2.2 ROC Curve in a Probabilistic Setting

Suppose we have positive and negative datasets $X_+ := \{\boldsymbol{x}_i^+\}_{i=1}^{n_+}$ and $X_- := \{\boldsymbol{x}_i^-\}_{i=1}^{n_-}$ drawn from two distributions $\mathbb{P}_+$ and $\mathbb{P}_-$ respectively. These distributions have respective probability density functions $p_+(\boldsymbol{x})$ and $p_-(\boldsymbol{x})$ that are both defined on the domain $\mathcal{X} \subseteq \mathbb{R}^d$. A classification score function (score function for short) takes a sample $\boldsymbol{x}$ as input and outputs a real-valued score. Suppose we have a score function $t(\boldsymbol{x}) \in \mathbb{R}$. We classify $\boldsymbol{x}$ as positive if $t(\boldsymbol{x}) > \tau$, where τ is a threshold.

Let F_+ and F_- denote the Cumulative Distribution Functions (CDFs) of $t(\boldsymbol{x}^+)$ and $t(\boldsymbol{x}^-)$ respectively. Then we can define the following quantities:

- False Positive Rate (FPR) at threshold τ, $\tilde{F}_-(\tau) := 1 - F_-(\tau)$
- True Positive Rate (TPR) at threshold τ, $\tilde{F}_+(\tau) := 1 - F_+(\tau)$
- The ROC curve of a score function t: the graph of function $\tilde{F}_+[\tilde{F}_-^{-1}(s)]$, where $s \in [0, 1]$.

The above definition of ROC curve requires F_- to be strictly increasing. In this paper, we assume F_+, F_- to be both strictly increasing. Both F_+ and F_- depend on the choice of score function t.

2.3 Arc Length of ROC Curve

Due to the strict monotonicity of CDFs, $\tilde{F}_+$ and $\tilde{F}_-$ form a bijective parameterization of the ROC curve in the sense that each point on this ROC curve can be written as $(\tilde{F}_-(\tau_0), \tilde{F}_+(\tau_0))$ for a unique $\tau_0 \in \mathbb{R}$. Using the line integral formula, the arc length of an ROC curve for a fixed score function t can be expressed using the derivatives of $\tilde{F}_-$ and $\tilde{F}_+$:

$$\widehat{\mathrm{ROC}}(t) := \int_{-\infty}^{\infty} \sqrt{\left[\partial_\tau \tilde{F}_+(\tau)\right]^2 + \left[\partial_\tau \tilde{F}_-(\tau)\right]^2} \mathrm{d}\tau = \int_{-\infty}^{\infty} \sqrt{f_+(\tau)^2 + f_-(\tau)^2} \mathrm{d}\tau, \qquad (1)$$

where $f_+(\tau)$ and $f_-(\tau)$ are the density functions of $t(\boldsymbol{x}^+)$ and $t(\boldsymbol{x}^-)$ respectively. Although (1) is an elementary result, it has been seldom discussed in the ROC literature. Authors in [3, 4] have proposed a performance metric computed over the "ROC hypersurface" and (1) is used to justify such a metric in a binary classification setting.

Using (1), we can confirm a simple geometric fact whose proof can be found in Appendix A:

Proposition 1. $\widehat{\mathrm{ROC}}(t) \in [\sqrt{2}, 2]$, *for all* t.

This result reflects the geometric observation that any monotone curve (such as ROC curve) starts and ends at two opposite corners of the ROC space $[0, 1]^2$ has an arc length between $\sqrt{2}$ and 2.

3 f-divergences Arising from ROC Arc Length

3.1 f-divergence of Score Distributions

Among many discrepancies measures, f-divergence has been widely used in many applications.

Definition 1. *Let p and q be densities of two continuous distributions. An f-divergence is defined as:* $\mathrm{D}_g(p|q) := \mathbb{E}_q\left[g\left(\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}\right)\right]$, *where g is convex and lower-semicontinuous satisfying $g(1) = 0$.*

Now let us slightly rewrite (1). Assuming f_- is strictly positive (in which case F_- is strictly increasing), we can write

$$\widehat{\mathrm{ROC}}(t) - \sqrt{2} = \mathbb{E}_{f_-} \sqrt{\left[\frac{f_+(\tau)}{f_-(\tau)}\right]^2 + 1} - \sqrt{2}. = \mathbb{E}_{f_-}\left[g\left(\frac{f_+(\tau)}{f_-(\tau)}\right)\right], \qquad (2)$$

where $g(s) = \sqrt{s^2+1} - \sqrt{2}$. Equation (2) yields the first important result of this paper: $\widehat{\mathrm{ROC}}(t) - \sqrt{2}$ is an f-divergence between score densities f_+ and f_- since $g(s)$ is a convex function and $g(1) = 0$. This result confirms, for any given t, $\widehat{\mathrm{ROC}}(t)$ is a good discrepancy for measuring positive and negative scores (i.e., score distributions). It also explains why the ROC arc length in Figure 1 is a good discrepancy measure: Given $t(x) = x$, the score distributions are same as the data distributions in each plot. The arc length of ROCs in Figure 1 plots are f-divergences of the score distributions hence are also f-divergences of data distributions.

Although $\widehat{\mathrm{ROC}}(t) - \sqrt{2}$ is an f-divergence of score distributions, it is not an f-divergence of the positive and negative *data* distributions for general t. The choice $t(x) = x$ in the toy example does not have simple analogues for higher dimensional datasets. Are there choices of t such that the arc lengths of their ROC curves are good discrepancy measures on data distribution? In what follows, we show when t is a bijective transform of the likelihood ratio $\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}$, $\widehat{\mathrm{ROC}}(t)$ encodes the differences between $p_+(\boldsymbol{x})$ and $p_-(\boldsymbol{x})$ in the form of an f-divergence between $p_+(\boldsymbol{x})$ and $p_-(\boldsymbol{x})$.

3.2 f-divergences of Data Distributions

Using the law of the unconscious statistician, we can express $\widehat{\mathrm{ROC}}(t)$ in terms of an expectation with respect to the negative data density $p_-(\boldsymbol{x})$: $\mathbb{E}_{p_-}\sqrt{\left[\frac{f_+(t(\boldsymbol{x}))}{f_-(t(\boldsymbol{x}))}\right]^2+1}$. Consider a special family of score functions: $t^*(\boldsymbol{x})=\gamma\left(\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right)$ where γ is any strictly increasing function. Due to the Neyman-Pearson lemma [25], $\mathrm{ROC}(t^*)$ has the highest TPR at any FPR level. Geometrically speaking, they dominate all other ROC cuves in an ROC plot. Hence, we refer to t^* as the optimal score and $\mathrm{ROC}(t^*)$ as the optimal ROC curve. For convenience, we denote the $\mathrm{ROC}(t^*)$ as ROC^* which reads "rock star". It can be shown that

$$\frac{f_+(t^*(\boldsymbol{x}_0))}{f_-(t^*(\boldsymbol{x}_0))}=\frac{\int_{\boldsymbol{x}:\gamma\left(\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right)=t^*(\boldsymbol{x}_0)}p_+(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\int_{\boldsymbol{x}:\gamma\left(\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right)=t^*(\boldsymbol{x}_0)}p_-(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}=\frac{p_+(\boldsymbol{x}_0)}{p_-(\boldsymbol{x}_0)},\quad\forall\gamma \tag{3}$$

where the second equality holds due to $\frac{\int_D a(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\int_D b(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}=\gamma^{-1}(C)$, when $\gamma\left(\frac{a(\boldsymbol{x})}{b(\boldsymbol{x})}\right)\equiv C,\ \forall\boldsymbol{x}\in D$. When $\gamma(\boldsymbol{x})=\boldsymbol{x}$, (3) expresses a known result [5], and is often given in plain English as "the density ratio of the likelihood ratio score is the likelihood ratio itself".

Finally, $\widehat{\mathrm{ROC}^*}$ takes an elegant form free from t^* or γ: $\mathbb{E}_{p_-(\boldsymbol{x})}\sqrt{\left[\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right]^2+1}$. Equivalently,

$$\widehat{\mathrm{ROC}^*}-\sqrt{2}=\mathbb{E}_{p_-(\boldsymbol{x})}\left[g\left(\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right)\right] \tag{4}$$

We can see that the same f-divergence arises from computing the arc length of ROC^*. However, unlike the f-divergence given in (2), (4) is an f-divergence between *data* distributions, not score distributions. It shows that as long as we use the the optimal scores, the arc length of the optimal ROC can indeed reflect the differences between *data* distributions. From now on, we will refer to $\widehat{\mathrm{ROC}^*}-\sqrt{2}$ as the ROC divergence. To the best of our knowledge, (4) has not been presented in literature before.

By definition, the ROC divergence is symmetric. Moreover, as a result of Proposition 1, the ROC divergence is upper bounded by $2-\sqrt{2}$ and lower bounded by 0. As we mentioned earlier, ROC^* has the maximal AUC. Some geometric properties of ROC^* can be found in Section B.

4 Estimating the Arc Length of ROC^*

4.1 A Variational Objective

To numerically approximate the arc length using samples alone, we leverage that $\widehat{\mathrm{ROC}^*}-\sqrt{2}$ is an f-divergence. Utilizing Fenchel's duality [18], Authors in [27] show that an f-divergence $\mathrm{D}_g(p_+|p_-)$ has a variational representation:

$$\begin{aligned}\mathrm{D}_g(p_+|p_-)=\int_{\mathcal{X}}p_-(\boldsymbol{x})g\left[\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right]\mathrm{d}\boldsymbol{x}&=\int_{\mathcal{X}}p_-(\boldsymbol{x})\sup_u\left\{u(\boldsymbol{x})\cdot\left[\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right]-g'[u(\boldsymbol{x})]\right\}\mathrm{d}\boldsymbol{x}\\&=\sup_u\int_{\mathcal{X}}p_+(\boldsymbol{x})u(\boldsymbol{x})-\int_{\mathcal{X}}p_-(\boldsymbol{x})g'[u(\boldsymbol{x})]\mathrm{d}\boldsymbol{x},\end{aligned}$$

where g' is the convex conjugate of g and the supremum is taken over all measurable functions. In the case of the ROC divergence, $g(z)=\sqrt{z^2+1}-\sqrt{2}$ with $z\in[0,\infty]$ thus g has a convex conjugate $g'(z')=-\sqrt{1-z'^2}-\sqrt{2}$, $z'\in[0,1]$. Rewriting $\widehat{\mathrm{ROC}^*}-\sqrt{2}$ using the above variational representation and dropping the $-\sqrt{2}$, we obtain:

$$\widehat{\mathrm{ROC}^*}=\sup_{u\in[0,1]}\mathbb{E}_{p_+}[u(\boldsymbol{x})]+\mathbb{E}_{p_-}[\sqrt{1-u^2(\boldsymbol{x})}].$$

We reparameterize $u(\boldsymbol{x}) = \sin[v(\boldsymbol{x})]$, where $v \in [0, \pi/2]$:

$$\widehat{\mathrm{ROC}^*} = \sup_{v \in [0,\pi/2]} \mathbb{E}_{p_+} \sin[v(\boldsymbol{x})] + \mathbb{E}_{p_-} \cos[v(\boldsymbol{x})]. \tag{5}$$

Differentiating the objective in (5) for v and set to zero, we can see the supremum is attained at $v^* = \operatorname{atan} \frac{p_+}{p_-}$. In other words, the optimal v^* is the arctangent likelihood ratio function.

It is also interesting to see how v^* is visualized in the ROC plot. We can see the tangent of ROC^* at an FPR level $s_0 \in [0, 1]$ is

$$\partial_s \tilde{F}_+(\tilde{F}_-^{-1}(s_0)) = \frac{f_+(\tilde{F}_-^{-1}(s_0))}{f_-(\tilde{F}_-^{-1}(s_0))} = \frac{p_+(\boldsymbol{x}_0)}{p_-(\boldsymbol{x}_0)},$$

where $\boldsymbol{x}_0$ is any point in $\mathcal{X}$ that satisfies the equality $\gamma\left(\frac{p_+(\boldsymbol{x}_0)}{p_-(\boldsymbol{x}_0)}\right) = \tilde{F}_-^{-1}(s_0)$. It can be seen that the likelihood ratio characterizes the slope of ROC^* and thus $v^* = \operatorname{atan} \frac{p_+}{p_-}$ is the *slope angle* of ROC^*.

Moreover, using (5), we can obtain a relationship between $\widehat{\mathrm{ROC}^*}$ and the total variation distance between $\mathbb{P}_+$ and $\mathbb{P}_-$ (denoted as $\mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-)$).

Proposition 2. $\mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-)$ *has a lower and upper bound expressed via* $\widehat{\mathrm{ROC}^*}$*:*

$$\max_{a \in [0,1]} \frac{2}{\pi}\left[\frac{\widehat{\mathrm{ROC}^*} - 2\sqrt{1-a^2}}{a} + \arccos(a) - \arcsin(a)\right] \leq \mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-) \leq \widehat{\mathrm{ROC}^*} - 1.$$

Proof of this proposition can be found in Section C. This proposition justifies that $\widehat{\mathrm{ROC}^*}$ is a valid measure of the discrepancy between $\mathbb{P}_+$ and $\mathbb{P}_-$. We compare this bound with other known TV bounds in Section 6.1.

4.2 A Tractable Objective for Estimating $\operatorname{atan} \frac{p_+}{p_-}$

To use (5) in practice, we need to find an appropriate function class $\mathcal{F}$. We can simply restrict v to a bounded (parametric/non-parametric) function class $\mathcal{F}$ and solve the sample version of (5):

$$\max_{v \in [0,\pi/2], v \in \mathcal{F}} \frac{1}{n_+} \sum_{i=1}^{n_+} \sin(v(\boldsymbol{x}_i^+)) + \frac{1}{n_-} \sum_{i=1}^{n_-} \cos(v(\boldsymbol{x}_i^-)). \tag{6}$$

In practice, enforcing the boundedness $v \in [0, \pi/2]$ over $\mathcal{X}$ is difficult. We can relax (6) by only enforcing the boundedness constraint of v on our finite sample dataset $X_+ \cup X_-$.

For example, by letting $\mathcal{F}$ be a Reproducing Kernel Hilbert Space (RKHS) [31], we can translate (6) into the following optimization problem:

$$\begin{aligned} &\hat{v} := \operatorname*{argmin}_{v \in \mathcal{H}} \ell(v) + \frac{\lambda}{2}\|v\|_{\mathcal{H}}^2, \quad \ell(v) := -\frac{1}{n_+}\sum_{i=1}^{n_+} \sin\langle v, \varphi(\boldsymbol{x}_i^+)\rangle - \frac{1}{n_-}\sum_{i=1}^{n_-} \cos\langle v, \varphi(\boldsymbol{x}_i^-)\rangle \\ &\text{s.t: } \langle v, \varphi(\boldsymbol{x})\rangle \in \left[0, \frac{\pi}{2}\right], \quad \forall \boldsymbol{x} \in X_+ \cup X_-, \end{aligned} \tag{7}$$

where $\mathcal{H}$ is a RKHS with a positive definite kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \varphi(\boldsymbol{x}), \varphi(\boldsymbol{x}')\rangle$, $\|\cdot\|_{\mathcal{H}}$ is the RKHS norm and $\frac{\lambda}{2}\|v\|_{\mathcal{H}}^2$ is the regularization term. The optimizer $\langle \hat{v}, \varphi(\boldsymbol{x})\rangle$ is an estimation of $v^*(\boldsymbol{x})$, the arctangent of the likelihood ratio. (7) is a strictly convex optimization and thus, if a solution $\hat{v}$ exists, it must be unique.

Instead of modelling $\operatorname{atan} \frac{p_+}{p_-}$, we can opt for modelling the log likelihood ratio $\log \frac{p_+}{p_-}$. However, this modelling choice results in an non-convex optimization thus presents extra challenges in the theoretical analysis. Details can be found in Section J.

4.3 Finite Sample Guarantee

We show that the solution of (7), $\langle \hat{v}, \varphi(\boldsymbol{x})\rangle$ converges to the true arctangent likelihood ratio (or its projection onto $\mathcal{H}$) as the number of samples goes to infinity. Below are a few regularity conditions.

Assumption 1. *There exists a unique $v^* \in \mathcal{H}$, such that $\mathbb{E}[\nabla_v \ell(v^*)] = 0$ and $\langle v^*, \varphi(\boldsymbol{x})\rangle \in [0, \pi/2]$ holds for all $\boldsymbol{x} \in \mathcal{X}$.*

A sufficient condition of the above condition is specified in the following proposition.

Proposition 3. *If there exists a unique $v^* \in \mathcal{H}$, such that $\langle v^*, \varphi(\boldsymbol{x})\rangle = \operatorname{atan}\left[\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right]$ then Assumption 1 holds.*

Proof of this Proposition can be found in Appendix F. Proposition 3 states if model is correctly specified and identifiable then Assumption 1 holds. It is possible that there exists a $v^* \in \mathcal{H}$ satisfying $\mathbb{E}[\nabla_v \ell(v^*)] = 0$ which does not meet the boundedness constraint $[0, \pi/2]$. In this paper we only consider situations where Assumption 1 holds, which includes all situations where the model is correctly specified and some situations where the model is misspecified.

Assumption 2. *Let $n_{\min} = \min(n_+, n_-)$. There exists a subspace $\mathcal{H}^* := \{v \in \mathcal{H} | \|v - v^*\|_{\mathcal{H}}^2 \leq \delta_{n_{\min}}^2\}$, such that $\forall v \in \mathcal{H}^*, \forall \boldsymbol{x} \in X_+ \cup X_-, \langle v, \varphi(\boldsymbol{x})\rangle \in (0, \frac{\pi}{2})$ holds with high probability. The sequence $\delta_{n_{\min}}$ is monotonically decreasing as $n_{\min}$ grows to infinity.*

Assumption 2 states all v within a vicinity of v^* are in the interior of (7)'s feasible region with high probability. The following proposition gives a sufficient condition under which Assumption 2 holds.

Proposition 4. *Suppose $\|\varphi(\boldsymbol{x})\|_{\mathcal{H}} \leq 1$. If our model is correctly specified as described in Proposition 3 and $\forall \boldsymbol{x} \in \mathcal{X}$, $\operatorname{atan}\left[\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right] \in [R_1, R_2]$, for some R_1 and R_2 such that $\frac{\pi}{2} > R_2 > R_1 > 0$, then there exists $N > 0$ such that Assumption 2 holds when $n_{\min} > N$.*

The proof can be found in Appendix D. Since we use RKHS as the estimator function class, our final assumption is that v^* should be reasonably smooth. In previous works such an assumption depends on the decay of the integral operator's eigenvalues [34, 8]. In this paper, we measure the smoothness using the *range space* technique which has recently been adopted in [9, 33]. We define

$\boldsymbol{\Sigma}_v := \mathbb{E}[\nabla_f^2 \ell(v)] = \mathbb{E}_{p_+}[\sin\langle v, \varphi(\boldsymbol{x})\rangle \cdot \varphi(\boldsymbol{x}, \cdot) \otimes \varphi(\boldsymbol{x}, \cdot)] + \mathbb{E}_{p_-}[\cos\langle v, \varphi(\boldsymbol{x})\rangle \cdot \varphi(\boldsymbol{x}, \cdot) \otimes \varphi(\boldsymbol{x}, \cdot)]$,

where $\otimes$ denotes the outer product. Given $v_0 \in \mathcal{H}$, $\boldsymbol{\Sigma}_{v_0}$ is an integral operator on $u \in \mathcal{H}$ and

$$\boldsymbol{\Sigma}_{v_0} u = \mathbb{E}_{p_+}[\sin\langle v_0, \varphi(\boldsymbol{x})\rangle \cdot \varphi(\boldsymbol{x}, \cdot) \cdot u(\boldsymbol{x})] + \mathbb{E}_{p_-}[\cos\langle v_0, \varphi(\boldsymbol{x})\rangle \cdot \varphi(\boldsymbol{x}, \cdot) \cdot u(\boldsymbol{x})].$$

By definition $\boldsymbol{\Sigma}_{v_0}$ is a positive, self-adjoint operator, in the sense that $\langle u, \boldsymbol{\Sigma}_{v_0} u\rangle \geq 0$, $\langle u, \boldsymbol{\Sigma}_{v_0} v\rangle = \langle \boldsymbol{\Sigma}_{v_0} u, v\rangle, \forall v, u \in \mathcal{H}$. Moreover, some algebra shows that $\boldsymbol{\Sigma}_{v_0}$ is also a bounded and compact operator. See Section G for more details.

Next, we assume the true arctangent ratio function (or its projection) is in the range space of $\boldsymbol{\Sigma}_{v^*}$.

Assumption 3. *Let $\mathcal{R}(\boldsymbol{\Sigma}_{v^*})$ denote the range space of $\boldsymbol{\Sigma}_{v^*}$. There exists $0 < \beta \leq 1$, $v^* \in \mathcal{R}(\boldsymbol{\Sigma}_{v^*}^{\beta})$, where C^{β} is the fraction power of a compact, positive and self-adjoint operator C.*

Note that the larger β is, the smoother the functions in the range space are. More discussions on the range space assumption can be found in Section 4.2, [33]. Now we are ready to state our theorem:

Theorem 1 (Convergence Rate of $\hat{v}$)**.** *Suppose Assumptions 1, 2 and 3 hold and $\hat{v}$ exists. If $\|\varphi(\boldsymbol{x})\|_{\mathcal{H}} \leq 1$ and*

$$\lambda = \frac{T}{n_{\min}^{1/4}}, \quad \frac{K}{n_{\min}^{\beta/4}} \leq \delta_{n_{\min}} \leq \frac{4}{\max(B_+, B_-)},$$

where

$$B_+ = \|(\boldsymbol{\Sigma}_{v^*} + \lambda \boldsymbol{I})^{-1} \mathbb{E}_{p_+}[\varphi(\boldsymbol{x})]\|_{\mathcal{H}}, \;\; B_- = \|(\boldsymbol{\Sigma}_{v^*} + \lambda \boldsymbol{I})^{-1} \mathbb{E}_{p_-}[\varphi(\boldsymbol{x})]\|_{\mathcal{H}},$$

and $T \geq 1, K > 0$ are constants that do not depend on $n_{\min}$, then there exists a constant $N > 0$ such that when $n_{\min} > N$, $\|\hat{v} - v^*\|_{\mathcal{H}} = O_p(n_{\min}^{-\beta/4})$.

The proof can be found at Appendix E. Theorem 1 shows that, under mild assumptions, $\hat{v}$ is indeed a good estimator for $\operatorname{atan}\left[\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right]$. Some discussions comparing our results with convergence results proved by Nguyen et al. [26] can be found in Section L. Since $\operatorname{atan}\left[\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right]$ is an optimal score that gives rise to ROC^*, our estimator may have some interesting applications such as outlier detection [17] or Neyman-Pearson classification [35]. We will defer discussions on those applications in future works. In the next section, we employ our arctangent likelihood ratio estimator in an application of lower bounding the maximal AUC.

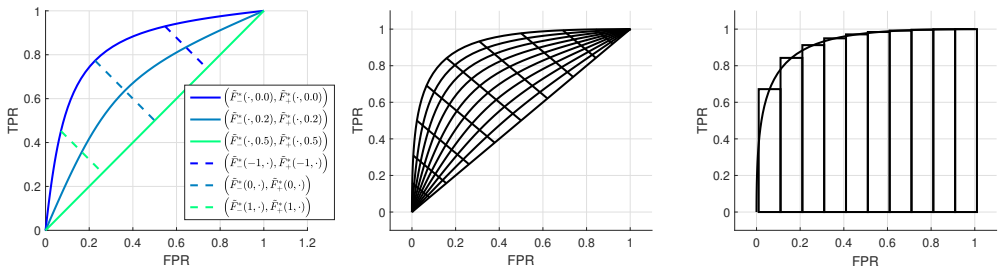


Figure 2: Left: $(\tilde{F}_-^*(\tau,\alpha), \tilde{F}_+^*(\tau,\alpha))$ parameterizes the surface between ROC^* and the diagonal line in $[0,1]^2$. This plot is created by setting $p_+ = \mathcal{N}(1,1)$, $p_- = \mathcal{N}(-1,1)$ and $t^*(x) = \frac{1}{2}\log\frac{p_+(x)}{p_-(x)} = x$. Middle: Our parameterization "mesh" divides AUC into surface elements (small patches on the plot). Right: Wilcoxon-Mann-Whitney statistic divides AUC into histogram bars.

5 Lower Bounding the Maximal AUC

Finding a score function t that approximately maximizes AUC is an important task in binary classification. Due to the Neyman-Pearson lemma, ROC^* has the maximum AUC among all ROC curves. Let us denote the AUC of ROC^* as AUC^*. Consider the following inequalities:

$$\mathrm{AUC}^* = \underbrace{\sup_t \mathbb{E}_{p_-}\mathbb{E}_{p_+}\left[\mathbb{1}\left(t(\boldsymbol{x}^+) \geq t(\boldsymbol{x}^-)\right)\right]}_{(i)} \geq \sup_{t\in\mathcal{F}'} \underbrace{\mathbb{E}_{p_-}\mathbb{E}_{p_+}\left[L\left(t(\boldsymbol{x}^+), t(\boldsymbol{x}^-)\right)\right]}_{(ii)}, \tag{8}$$

where $L(a,b)$ is a continuous and concave lower bound of the indicator function $\mathbb{1}(a > b)$. It can be seen that $\mathbb{E}_{p_-}\mathbb{E}_{p_+}\left[\mathbb{1}\left(t(\boldsymbol{x}_+) \geq t(\boldsymbol{x}_-)\right)\right] = \int_{[0,1]} \tilde{F}_+(\tilde{F}_-^{-1}(s))\mathrm{d}s$ is the AUC of $\mathrm{ROC}(t)$. Due to the Neyman-Pearson lemma, the supremum of (i) is only attained when $t(\boldsymbol{x}) = \gamma\left(\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right)$ where γ is a strictly increasing function. Replacing the expectations in (ii) with sample averages yields the optimization problem of AUC maximization [2, 10]:

$$\max_{t\in\mathcal{F}'} \frac{1}{n_- n_+} \sum_{i=1}^{n_-}\sum_{j=1}^{n_+} L(t(\boldsymbol{x}_j^+), t(\boldsymbol{x}_i^-)). \tag{9}$$

Therefore, we can see that AUC maximization is a procedure that approximates an optimal score function by maximizing a lower bound of AUC^*.

Now, we show a different way of lower bounding AUC^* with the help of $\mathrm{atan}\,\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}$. We have seen that how $(\tilde{F}_-(\tau), \tilde{F}_+(\tau))$ parameterizes an ROC curve in Section 2.3. In fact, the surface between ROC^* and the diagonal line between $(0,0)$ and $(1,1)$ can be similarly parameterized by considering ROC curves of *mixtures* of positive and negative score distributions. Let F_+^* and F_-^* denote CDFs of any optimal score. For $\alpha \in [0,.5]$, we define FPR and TPR for α-mixtures of F_+^* and F_-^*:

$$\begin{aligned}\tilde{F}_-^*(\tau,\alpha) :=& 1 - [(1-\alpha)F_-^*(\tau) + \alpha F_+^*(\tau)],\\ \tilde{F}_+^*(\tau,\alpha) :=& 1 - [\alpha F_-^*(\tau) + (1-\alpha)F_+^*(\tau)].\end{aligned}$$

We can see the 2-dimensional coordinate $\boldsymbol{r}(\tau,\alpha) := (\tilde{F}_-^*(\tau,\alpha), \tilde{F}_+^*(\tau,\alpha))$ parameterizes the area between ROC^* and the diagonal line in $[0,1]^2$: When fixing α and varying τ, the coordinates give rise to a smooth curve in ROC space from $[0,0]$ to $[1,1]$. When $\alpha = 0$, such a curve is ROC^*. When $\alpha = .5$, such a curve is the diagonal line. When fixing $\tau = \tau_0$ and varying α, the coordinates produce a straight line segment connecting $(\tilde{F}_-^*(\tau_0,0), \tilde{F}_+^*(\tau_0,0))$ and $(\tilde{F}_-^*(\tau_0,.5), \tilde{F}_+^*(\tau_0,.5))$. The left plot in Figure 2 visualizes this parameterization. Now the surface area sandwiched between ROC^* and the diagonal line can be computed using a surface integration:

$$\mathrm{AUC}^* - .5 = \int_{\mathrm{dom}(\tau)}\int_{[0,.5]} \|\partial_\tau \boldsymbol{r}(\tau,\alpha) \times \partial_\alpha \boldsymbol{r}(\tau,\alpha)\|\,\mathrm{d}\alpha\mathrm{d}\tau, \tag{10}$$

where $\times$ denotes the cross product. After some algebra and applying the Fenchel duality technique in Section 4.1, we prove that AUC^* can be expressed as the supremum of a variational objective similar to (5):

Proposition 5. $\mathrm{AUC}^* = \frac{\sqrt{2}A}{2} + \frac{1}{2}$,

$$A := \sup_{v \in [0, \pi/2]} \mathbb{E}_{p_+}\left[w\left(\operatorname{atan}\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right)\sin[v(\boldsymbol{x})]\right] + \mathbb{E}_{p_-}\left[w\left(\operatorname{atan}\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right)\cos[v(\boldsymbol{x})]\right], \quad (11)$$

where $w(\tau) := \sin(\tau + \frac{\pi}{4}) \cdot |F_+^*(\tau) - F_-^*(\tau)|$. *The supremum of* (11) *is attained at* $v^* = \operatorname{atan}\frac{p_+}{p_-}$.

The proof can be found in Appendix H in the supplementary material. A lower bound of A can be obtained by restricting v to a function class.

Evaluating w requires us to evaluate $\operatorname{atan}\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}$, F_+^* and F_-^* which are not readily available. However, Section 4.3 shows that the empirical estimator (7) is a consistent estimator of $\operatorname{atan}\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}$ under mild conditions. Therefore, we propose the following two-step procedure to approximately lowerbound A:

Algorithm 1 Two-step Procedure for lower bounding A

1. Obtain $\hat{t}(\boldsymbol{x}) := \langle \hat{v}, \phi(\boldsymbol{x}) \rangle$ using (7). Approximate F_+^* and F_-^* using $\hat{F}_+$ and $\hat{F}_-$ which are empirical CDFs of $\hat{t}(\boldsymbol{x}^+)$ and $\hat{t}(\boldsymbol{x}^-)$.
2. Optimize the empirical version of (11) by restricting v to a feasible function class and plugging in estimates obtained in the earlier step, i.e.,

$$\hat{v} := \operatorname*{argmax}_{v \in [0, \pi/2], v \in \mathcal{F}} \frac{1}{n_+}\sum_{i=1}^{n_+} \hat{w}\left[\hat{t}(\boldsymbol{x}_i)\right] \cdot \sin[v(\boldsymbol{x}_i)] + \frac{1}{n_-}\sum_{i=1}^{n_-} \hat{w}\left[\hat{t}(\boldsymbol{x}_i)\right] \cdot \cos[v(\boldsymbol{x}_i)], \quad (12)$$

where $\hat{w}(\tau) := \sin(\tau + \frac{\pi}{4})\left|\hat{F}_+(\tau) - \hat{F}_-(\tau)\right|$.

Note that (12) is nothing but a weighted sample objective (6). Thus, it can be easily optimized by the algorithm that solves a weighted version of (6) given the approximated weights in the first step. In practice, we simply run solver for (7) twice: The first time we run it without weights then run it again with weights $\hat{w}\left[\hat{t}(\boldsymbol{x}_i)\right]$ calculated from the first run.

Since the above algorithm also approximates an optimal score ($\operatorname{atan}\frac{p_+}{p_-}$) by maximizing a lower bound of AUC^*, it is natural to wonder how the maximizer $\hat{v}$ of (12) would perform in AUC maximization tasks. In the next section, we show that our two-step algorithm achieves a promising AUC performance compared to an AUC maximizer.

Computing AUC^* using (10) is different from using the conventional Wilcoxon-Mann-Whitney statistic: Our approach divides the space between ROC^* and the diagonal into small surface elements and then adds them up. Wilcoxon-Mann-Whitney statistic adds up all histogram bars, which are TPRs at different FPR levels. Our approach requires $\tilde{F}_+$ and $\tilde{F}_-$ to be differentiable with respect to τ, which means the score distributions cannot be discrete. However, Wilcoxon-Mann-Whitney can compute the AUC of discrete score distributions without a problem. This difference is visualized in the middle and right plots of Figure 2.

6 Experiments

6.1 Numerical Comparison of Divergences and TV Bounds

In this experiment, we numerically compare the ROC divergence, the upper and lower bound in Proposition 2 with several other divergences and some known bounds of TV in Figure 3. In this numerical simulation, $p_+ = \mathcal{N}(0, 1)$ and $p_- = \mathcal{N}(\delta, 1)$. We plot ROC divergence, JS divergence, Wasserstein distance and TV between p_+ and p_- as δ grows from 0 to 5. We can see the (rescaled)

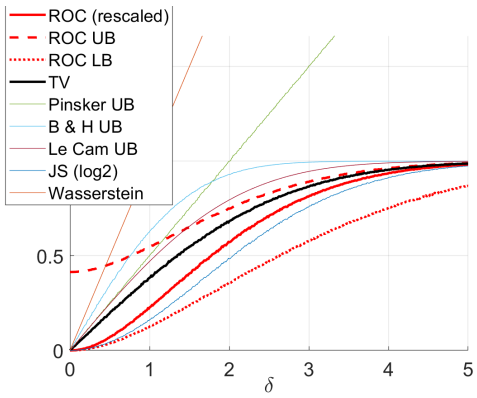


Figure 3: Comparison of various divergences and bounds of TV.

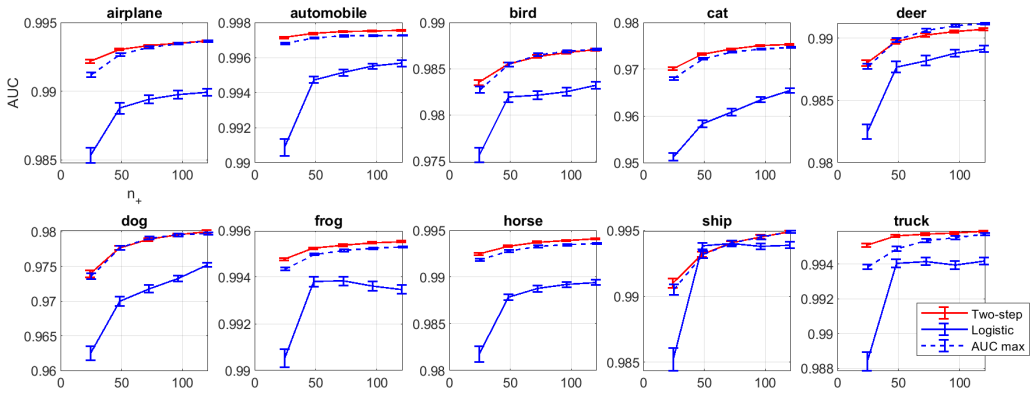


Figure 4: Testing AUC of one-versus-the-rest classification on CIFAR-10 dataset.

ROC divergence closely resembles TV. When $\delta > 1.45$, the upper bound given in Proposition 2 is the tightest among known TV upper bounds (Pinsker's upperbound, Bretagnolle & Huber's upper bound, Le Cam's upper bound). This suggests that combining our upperbound with existing bounds may produce an even tighter bound for TV.

6.2 Imbalanced Classification on CIFAR-10

In this section, we test if the $\hat{v}$ obtained in our two-step procedure (12) is indeed a good score function in terms of AUC in imbalanced classification tasks. We use a widely known image classification dataset CIFAR-10 [20]. The performance is compared with an AUC maximizer which maximizes the empirical lower bound in (9) and a vanilla logistic regression classifier. We set the surrogate loss $L(a, b) := -(1 - (a - b))^2$ in the AUC maximizer, as suggested in [10]. All methods use linear models with no regularization terms since our models are simple and we have sufficient samples. Particularly, the $\hat{t}$ and $\hat{v}$ in our two-step algorithm is obtained using (7) by setting $\phi(\boldsymbol{x}) = \boldsymbol{x}$ and $\lambda = 0$. The AUC maximization (AUC max) is implemented using SPAUC method [22].

Instead of using the raw features, we extract 50 dimensional bounded features by training a residual network [15] on the training dataset using the 10-class cross entropy loss. The structure of the network is included in the supplementary material. After obtaining features, we construct datasets for 10 different one-versus-the-rest classification tasks. For a single task, we pick a class and attain X_+ by randomly sampling from this class n_+ times in the training set. Similarly, X_- is attained by randomly sampling from the rest of the classes n_- times. In our experiments, we set $n_+ = 24, 48, 72, 96, 120$ and fix $n_- = 1000$ to create imbalanced positive and negative datasets. We run all three methods and obtain the corresponding score functions. For each class, we repeat the experiment 96 times using different random samples. We use the testing and training split provided by the dataset itself.

Our experiments can be seen as a transfer learning task which reuses predictive features trained for a multi-class classifier for one-versus-the-rest binary classification tasks.

The average AUCs computed on the testing dataset and their standard errors over 96 runs over different n_+ sample sizes are shown in Figure 4. Our method has approximately equal performance with the AUC maximizer despite not directly maximizing the AUC and having lower computational complexity. This observation indicates that $\hat{\hat{v}}$ can be a good score function in AUC maximization tasks. Both of the methods significantly outperform vanilla logistic regression.

7 Discussions on Computational Complexity

Without loss of generality, assume $n_+ = n_- = n$. The naive caclulation of objective (9) has a computational complexity $O(n^2)$ since we evaluate the loss function L at each pair of samples. In comparison, the objective (12) is a summation of $\sin / \cos(v(\boldsymbol{x}))$ evaluated at each datum, so computing the objective/gradient has a computational complexity $O(n)$. However, authors in [19] have shown that the objective function in (9) can also be computed with $O(n \log(n))$ complexity. A recent work [37] has simplified the computation of (9) for the squared surrogate loss function L and linear-in-parameter models, so computing the objective has a computational complexity $O(n)$. When implemented in an online fashion, it has a computational complexity of one datum.

Computing $\hat{F}_+$ and $\hat{F}_-$ requires sorting our dataset, which has an average complexity $O(n \log n)$. However, once our datasets are sorted, $\hat{F}_+\left(\hat{t}(\boldsymbol{x}_0)\right) = \frac{i}{n_+}$, where i is the index of $\hat{t}(\boldsymbol{x}_0)$ in the sorted set $\{\hat{t}(\boldsymbol{x}_i)\}_{i=1}^{n_+}$.

Acknowledgments and Disclosure of Funding

The author would like to thank Prof. Peter Flach and Dr. Hao Song for their helpful discussions. The author would like to thank four anonymous reviewers for their insightful comments.

References

[1] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[2] C. Cortes and M. Mohri. Auc optimization vs. error rate minimization. In *Advances in Neural Information Processing Systems 16 (NeurIPS 2003)*, volume 16, 2003.

[3] D. C. Edwards and C. E. Metz. A utility-based performance metric for roc analysis of n-class classification tasks. In *Medical Imaging 2007: Image Perception, Observer Performance, and Technology Assessment*, volume 6515, pages 21 – 30. SPIE, 2007.

[4] D. C. Edwards and C. E. Metz. Optimality of a utility-based performance metric for roc analysis. In *Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment*, volume 6917, pages 122 – 127. SPIE, 2008.

[5] S. Eguchi and J. Copas. A class of logistic-type discriminant functions. *Biometrika*, 89(1):1–22, 2002.

[6] T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[7] P. A. Flach. Roc analysis. In *Encyclopedia of machine learning and data mining*, pages 1–8. Springer, 2016.

[8] K. Fukumizu. *Exponential manifold by reproducing kernel Hilbert spaces*, page 291–306. Cambridge University Press, 2009.

[9] K. Fukumizu, L. Song, and A. Gretton. Kernel bayes' rule: Bayesian inference with positive definite kernels. *The Journal of Machine Learning Research*, 14(1):3753–3783, 2013.

[10] W. Gao, R. Jin, S. Zhu, and Z-H Zhou. One-pass auc optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, volume 28, pages 906–914, 2013.

[11] D. J. Goodenough, K. Rossmann, and L. B Lusted. Radiographic applications of receiver operating characteristic (roc) curves. *Radiology*, 110(1):89–95, 1974.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27 (NeurIPS 2014)*, pages 2672–2680, 2014.

[13] D. M. Green and J. A. Swets. *Signal detection theory and psychophysics*, volume 1. Wiley New York, 1966.

[14] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] J. Hermans, V. Begy, and G. Louppe. Likelihood-free MCMC with amortized approximate ratio estimators. In *Proceedings of the 37th International Conference on Machine Learning (ICML2020)*, volume 119, pages 4239–4248, 2020.

[17] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori. Statistical outlier detection using direct density ratio estimation. *Knowledge and information systems*, 26(2):309–336, 2011.

[18] J.B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

[19] T. Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd international conference on Machine learning (ICML)*, pages 377–384, 2005.

[20] A. Krizhevsky. Learning multiple layers of features from tiny images. 2009. techincal report.

[21] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[22] Y. Lei and Y. Ying. Stochastic proximal auc maximization. *Journal of Machine Learning Research*, 22(61):1–45, 2021.

[23] Z. Lin, A. Khetan, G. Fanti, and S. Oh. Pacgan: The power of two samples in generative adversarial networks. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, volume 31, 2018.

[24] L. B. Lusted. Signal detectability and medical decision-making. *Science*, 171(3977):1217–1219, 1971.

[25] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

[26] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems 20*, pages 1089–1096. 2008.

[27] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56 (11):5847–5861, 2010.

[28] S. Nowozin, B. Cseke, and R. Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems 29 (NeurIPS 2016)*, pages 271–279, 2016.

[29] M. D. Reid and R. C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12(22):731–817, 2011.

[30] L. Rosasco, M. Belkin, and E. De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11(2), 2010.

[31] B. Scholkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[32] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. RG Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6: 1550–1599, 2012.

[33] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, A. Hyvärinen, and R. Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57): 1–59, 2017.

[34] I. Steinwart, D. R. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Proceedings of Conference on Learning Theory (COLT) 2009*, pages 79–93, 2009.

[35] X. Tong. A plug-in approach to neyman-pearson classification. *Journal of Machine Learning Research*, 14(56):3011–3040, 2013.

[36] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[37] Y. Ying, L. Wen, and S. Lyu. Stochastic online auc maximization. In *Advances in Neural Information Processing Systems 29 (NeurIPS 2016)*, volume 29, 2016.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   [Yes] In introduction, we discussed the background/usages ROC curves, f-divergence estimation and provided a full paragraph summary of the contribution of this paper. We give the same contribution rundown in the abstract as well.
   (b) Did you describe the limitations of your work?
   [Yes] In Section 5, we included a brief discussion comparing the proposed maximal AUC lower bounding approach vs. the classic Wilcoxon-Mann-Whitney statistic.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them?
   [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results?
   [Yes] See Section 4.3
   (b) Did you include complete proofs of all theoretical results?
   [Yes] See Appendix
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)?
   [Yes] Full instructions on how to reproduce our experiments is provided in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
   [Yes] See Section 6.2
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)?
   [Yes] See Section 6.2

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)?
[N/A] The Computation time is not compared in this paper.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators?
   [Yes] See Section 6.2
   (b) Did you mention the license of the assets?
   [No] The authors of CIFAR-10 do not specify a license for the dataset.
   (c) Did you include any new assets either in the supplemental material or as a URL?
   [No] We did not create any new asset in this research.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating?
   [No] We did not collect any data ourselves.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?
   [N/A] CIFAR-10 dataset we use in our experiment does not contain any identifiable human information.

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable?
   [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable?
   [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation?
   [N/A]

A Proof of Proposition 1

Proof. Jensen's inequality:

$$\widehat{\mathrm{ROC}}(t) = \sqrt{2}\int \sqrt{\frac{f_+^2(\tau)}{2} + \frac{f_-^2(\tau)}{2}}\mathrm{d}\tau \geq \sqrt{2}\int \frac{|f_+(\tau)|}{2} + \frac{|f_-(\tau)|}{2}\mathrm{d}\tau = \sqrt{2}.$$

Triangle inequality: $\widehat{\mathrm{ROC}} \leq \int |f_+(\tau)| + |f_-(\tau)|\mathrm{d}\tau = 2.$ □

B Geometric Properties of ROC^*

Here we prove a result regarding some other geometric properties of ROC^*.

Proposition 6. ROC^* *is a convex curve and* $\widehat{\mathrm{ROC}^*}$ *is the longest among all convex ROC curves.*

Proof. First, we show ROC^* is a convex curve. To show ROC^* is convex, we only need to show $\tilde{F}_+(\tilde{F}_-^{-1}(s))$ is a concave function. This can be verified by checking the sign of $\partial_s^2 \tilde{F}_+(\tilde{F}_-^{-1}(s))$:

$$\partial_s \tilde{F}_+(\tilde{F}_-^{-1}(s)) = \frac{f_+[\tilde{F}_-^{-1}(s)]}{f_-[\tilde{F}_-^{-1}(s)]} = \frac{p_+(\boldsymbol{x}_0)}{p_-(\boldsymbol{x}_0)} = \gamma^{-1}\left(\tilde{F}_-^{-1}(s)\right),$$

where the second equality is due to (3) and $\boldsymbol{x}_0$ is any point in $\mathcal{X}$ that satisfies the equality $\gamma\left(\frac{p_+(\boldsymbol{x}_0)}{p_-(\boldsymbol{x}_0)}\right) = \tilde{F}_-^{-1}(s)$. Further, we can show that,

$$\partial_s^2 \tilde{F}_+(\tilde{F}_-^{-1}(s;t^*)) = -\frac{1}{\partial_s \gamma[\gamma^{-1}(\tilde{F}_-^{-1}(s))]} \cdot \frac{1}{f_-[(\tilde{F}_-^{-1}(s)]}.$$

Since γ is a strictly monotone increasing function, the first factor is non-negative and the second factor is also strictly positive due to the our assumption on the positivity of f_-. We have $\partial_s^2 F_+(F_-^{-1}(s;t^*)) \leq 0$. Moreover, at any FPR level $s \in [0,1]$, the Neyman-Pearson lemma [25] implies

$$\tilde{F}_+(\tilde{F}_-^{-1}(s)) \geq \tilde{F}'_+(\tilde{F}'^{-1}_-(s)),$$

where $\tilde{F}'_+$ and $\tilde{F}'_-$ are TPR and FPR of any other score function. In words, ROC^* dominates all other ROC curves. Since ROC^* is convex and encloses all other ROC curves, our claim follows Archimedes's Second Axiom: among all convex curves with the same endpoints, the one encloses all other curves has the longest arc length. □

C Proof of Proposition 2

Proof. Using the integral probability metric representation of $\mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-)$ [32], we can write:

$$\begin{aligned}\frac{\pi}{2}\mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-) &= \sup_{\|v\|_\infty \leq 1} \mathbb{E}_{p_+}\left[\frac{\pi}{2}\cdot\frac{(v(\boldsymbol{x})+1)}{2}\right] - \mathbb{E}_{p_-}\left[\frac{\pi}{2}\cdot\frac{(v(\boldsymbol{x})+1)}{2}\right] \\ &= \sup_{v' \in [0,\pi/2]} \mathbb{E}_{p_+}[v'(\boldsymbol{x})] + \mathbb{E}_{p_-}\left[-v'(\boldsymbol{x})\right]\end{aligned}$$

Some algebra can show that $z \geq \frac{\sin(z)}{a} + \arccos(a) - \frac{\sqrt{1-a^2}}{a}$ and $-z \geq \frac{\cos(z)}{a} - \arcsin(a) - \frac{\sqrt{1-a^2}}{a}$ for all $a \in [0,1]$ and $z \in [0, \pi/2]$. Therefore

$$\begin{aligned}\frac{\pi}{2}\mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-) &\geq \sup_{v' \in [0,\pi/2]} \mathbb{E}_{p_+}\left[\frac{\sin(\boldsymbol{x})}{a}\right] + \mathbb{E}_{p_-}\left[\frac{\cos(\boldsymbol{x})}{a}\right] + \arccos(a) - \arcsin(a) - \frac{2\sqrt{1-a^2}}{a} \\ &\geq \frac{\widehat{\mathrm{ROC}^*}}{a} + \arccos a - \arcsin a - \frac{2\sqrt{1-a^2}}{a}.\end{aligned}$$

Similarly, multiplying both sides of the second equality above by $\frac{2}{\pi}$, we obtain

$$\begin{aligned}\mathrm{TV}(\mathbb{P}_+, \mathbb{P}_-) &= \sup_{v' \in [0,\pi/2]} \mathbb{E}_{p_+} \frac{2}{\pi} v'(\boldsymbol{x}) + \mathbb{E}_{p_-}\left[-\frac{2}{\pi} v'(\boldsymbol{x})\right] \\ &\leq \sup_{v' \in [0,\pi/2]} \mathbb{E}_{p_+} \sin(v'(\boldsymbol{x})) + \mathbb{E}_{p_-}\left[\cos v'(\boldsymbol{x}) - 1\right] \\ &= \widehat{\mathrm{ROC}}^* - 1.\end{aligned}$$

□

D Proof of Proposition 4

Proof. $\forall v \in \mathcal{H}^*, |\langle v, \varphi(\boldsymbol{x})\rangle - \langle v^*, \varphi(\boldsymbol{x})\rangle| \leq \|v - v^*\|_{\mathcal{H}} \|\varphi(\boldsymbol{x})\|_{\mathcal{H}} \leq \delta_{n_{\min}}$. If $\delta_{n_{\min}} < \min(R_1, \frac{\pi}{2} - R_2)$ then $\langle v, \varphi(\boldsymbol{x})\rangle \in (0, \frac{\pi}{2})$ holds uniformly for every $\boldsymbol{x} \in \mathcal{X}$. As $\delta_{n_{\min}}$ is a decaying sequence, there always exists an N such that $\delta_{n_{\min}} \leq \min(R_1, \frac{\pi}{2} - R_2)$ holds for $n_{\min} \geq N$. □

E Proof of Theorem 1

To reduce the visual clutter, in this section, $\|v\|$ represents the Hilbert space norm of v, defined as $\sqrt{\langle v, v\rangle}$. We simplify $\mathbb{E}_{p_+}[v(\boldsymbol{x})]$ as $\mathbb{E}_+[v(\boldsymbol{x})]$ whenever it does not lead to confusion. For ease, we write $\sum_{i=1}^{n_+} f(\boldsymbol{x}_i^+)$ as $\sum_{i=1}^{n_+} f(\boldsymbol{x}_i)$, a convention which will adopted henceforth.

Proof. Define $\mathcal{H}^* := \{v \in \mathcal{H} | \|v - v^*\|^2 \leq \delta^2\}$. Consider an optimization that is similar to (7):

$$\tilde{v} := \operatorname*{argmin}_{v \in \mathcal{H}^*} \ell(v) + \frac{\lambda}{2}\|v\|^2 \tag{13}$$

Define $\tilde{u} := \tilde{v} - v^*$ and we have the following equality due to the KKT conditions of (13)

$$\nabla_v \ell(\tilde{v}) + \lambda \tilde{v} + 2\nu \tilde{u} = 0,$$

where ν is a Lagrangian multiplier and $\nu \geq 0$. Multiplying both sides by $\tilde{s} = (\boldsymbol{\Sigma}_{v^*} + \lambda \boldsymbol{I})^{-1} \tilde{u}$, we have

$$\langle \tilde{s}, \nabla_v \ell(\tilde{v}) + \lambda \tilde{v} + 2\nu \tilde{u}\rangle = 0.$$

Let $g(v) := \langle \tilde{s}, \nabla_v \ell(v) + \lambda v + 2\nu(v - v^*)\rangle$, we can applying Mean Value Theorem (MVT) on the scalar valued function $g(v)$:

$$g(\tilde{v}) - g(v^*) = \nabla_v g(\bar{v}) \tilde{u}, \tag{14}$$

where $\bar{v} = a v^* + (1 - a)\hat{v}$ for some $a \in [0, 1]$. Knowing $g(\tilde{v}) = 0$ and $g(v^*) = \langle \tilde{s}, \nabla_v \ell(v^*) + \lambda v^*\rangle$, we can translate (14) into

$$\langle \tilde{s}, -\nabla_v \ell(v^*) - \lambda v^*\rangle = \langle \tilde{s}, [\nabla_f^2 \ell(\bar{v}) + \lambda \boldsymbol{I} + 2\tilde{\nu} \boldsymbol{I}]\tilde{u}\rangle, \tag{15}$$

where $\boldsymbol{I}$ is the identify matrix. Focusing on the RHS, we have

$$\begin{aligned}\langle \tilde{s}, [\nabla_f^2 \ell(\bar{v}) + \lambda \boldsymbol{I} + 2\nu \boldsymbol{I}]\tilde{u}\rangle \geq& \langle \tilde{s}, [\nabla_f^2 \ell(\bar{v}) + \lambda \boldsymbol{I}]\tilde{u}\rangle \\ \geq& \underbrace{\langle (\boldsymbol{\Sigma}_{v^*} + \lambda \boldsymbol{I})^{-1} \tilde{u}, [\boldsymbol{\Sigma}_{v^*} + \lambda \boldsymbol{I}]\tilde{u}\rangle}_{\|\tilde{u}\|^2} - \underbrace{\langle \tilde{s}, [\boldsymbol{\Sigma}_{v^*} - \boldsymbol{\Sigma}_{\bar{v}}]\tilde{u}\rangle}_{a} - \underbrace{\langle \tilde{s}, [\boldsymbol{\Sigma}_{\bar{v}} - \nabla_f^2 \ell(\bar{v})]\tilde{u}\rangle}_{b} \\ \geq& \|\tilde{u}\|^2 - a - b.\end{aligned} \tag{16}$$

The first line is due to the fact that $\langle 2\nu \tilde{s}, \tilde{u}\rangle \geq 0$. Use the inequality (16) on (15), we get the inequality

$$\langle \tilde{s}, -\nabla_v \ell(v^*) - \lambda v^*\rangle \geq \|\tilde{u}\|^2 - a - b. \tag{17}$$

First, let us inspect a. Using MVT on $\sin\langle v, \varphi(\boldsymbol{x})\rangle, v \in \mathcal{H}^*$ and applying Hölder's inequality, we get

$$\sin\langle v^*, \varphi(\boldsymbol{x})\rangle - \sin\langle v^* + \boldsymbol{\delta}', \varphi(\boldsymbol{x})\rangle \leq \|\boldsymbol{\delta}'\| \cdot \|\varphi(\boldsymbol{x})\| \leq \delta \cdot \|\varphi(\boldsymbol{x})\|. \tag{18}$$

Define $\mathbf{\Sigma}_v^+ := \mathbb{E}_+ \left[\sin\langle v, \varphi(\boldsymbol{x})\rangle \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})\right]$ and $\hat{\mathbf{\Sigma}}_v^+$ as its empirical counterpart approximated using X_+. We can see that $a = \langle \tilde{s}, [\mathbf{\Sigma}_{v^*}^+ - \mathbf{\Sigma}_{\bar{v}}^+]\tilde{u}\rangle + \langle \tilde{s}, [\mathbf{\Sigma}_{v^*}^- - \mathbf{\Sigma}_{\bar{v}}^-]\tilde{u}\rangle$. Moreover,

$$\begin{aligned}\langle \tilde{s}, [\mathbf{\Sigma}_{v^*}^+ - \mathbf{\Sigma}_{\bar{v}}^+]\tilde{u}\rangle &\overset{\text{i}}{\leq} \mathbb{E}_+ \left\{\delta \cdot \|\varphi(\boldsymbol{x})\| \cdot \langle \tilde{s}, \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})\tilde{u}\rangle\right\} \\ &\leq \delta \langle \tilde{u}, \mathbb{E}_+ \left\{(\mathbf{\Sigma}_{v^*} + \lambda \boldsymbol{I})^{-1}\varphi(\boldsymbol{x})\rangle \cdot \|\varphi(\boldsymbol{x})\|\right\} \cdot \|\tilde{u}\| \\ &\leq \delta \|\tilde{u}\| \cdot \|(\mathbf{\Sigma}_{v^*} + \lambda \boldsymbol{I})^{-1}\mathbb{E}_+\varphi(\boldsymbol{x})\| \cdot \|\tilde{u}\|\end{aligned}$$

(i) is due to (18). Following a similar line of reasoning, we can see

$$\langle \tilde{s}, [\mathbf{\Sigma}_{v^*}^- - \mathbf{\Sigma}_{\bar{v}}^-]\tilde{u}\rangle \leq \delta \| \left(\mathbf{\Sigma}_{v^*} + \lambda \boldsymbol{I}\right)^{-1} \mathbb{E}_-[\varphi(\boldsymbol{x})]\| \cdot \|\tilde{u}\|^2.$$

By setting $\delta \leq 4 \max\left(\| \left(\mathbf{\Sigma}_{v^*} + \lambda \boldsymbol{I}\right)^{-1} \mathbb{E}_+[\varphi(\boldsymbol{x})]\|, \| \left(\mathbf{\Sigma}_{v^*} + \lambda \boldsymbol{I}\right)^{-1} \mathbb{E}_-[\varphi(\boldsymbol{x})]\|\right)^{-1}$, we have

$$a \leq \frac{\|\tilde{u}\|^2}{2}. \tag{19}$$

Now we inspect b. We can see $|b| \leq \left|\tilde{s}\hat{\mathbf{\Sigma}}_{\bar{v}}^+\tilde{u} - \mathbb{E}_+\left[\tilde{s}\hat{\mathbf{\Sigma}}_{\bar{v}}^+\tilde{u}\right]\right| + \left|\tilde{s}\hat{\mathbf{\Sigma}}_{\bar{v}}^-\tilde{u} - \mathbb{E}_-\left[\tilde{s}\hat{\mathbf{\Sigma}}_{\bar{v}}^-\tilde{u}\right]\right|$. Define a scalar random variable

$$Z_f^{(i)} := \sin\langle v, \varphi(\boldsymbol{x}_i)\rangle \cdot \langle \tilde{s}, \varphi(\boldsymbol{x}_i) \otimes \varphi(\boldsymbol{x}_i)\tilde{u}\rangle.$$

By definition $\frac{1}{n_+}\sum_{i=1}^{n_+} Z_f^{(i)} = \tilde{s}^\top \hat{\mathbf{\Sigma}}_f^+ \tilde{u}$. Therefore

$$\left|\frac{1}{n_+}\sum_{i=1}^{n_+} Z_{\bar{v}}^{(i)} - \mathbb{E}Z_{\bar{v}}\right| \leq \sup_v \left|\frac{1}{n_+}\sum_{i=1}^{n_+} Z_v^{(i)} - \mathbb{E}Z_v\right|.$$

Since $0 \leq Z_v^{(i)} \leq \|\tilde{s}\| \cdot \|\tilde{u}\| \cdot \|\varphi(\boldsymbol{x})\|^2 \leq \| \left(\mathbf{\Sigma}_{v^*} + \lambda \boldsymbol{I}\right)^{-1} \tilde{u}\|\|\tilde{u}\| \leq \frac{\|\tilde{u}\|^2}{\lambda}$, using Uniform Law of Large Number for bounded random variable (Theorem 4.10, [36]),

$$\sup_v \left|\frac{1}{n_+}\sum_{i=1}^{n_+} Z_v^{(i)} - \mathbb{E}Z_v\right| \leq 2\mathcal{R}_{n_+}(\mathcal{F}_Z) + \frac{\|\tilde{u}\|^2 \cdot \|\varphi(\boldsymbol{x})\|^2}{\lambda\sqrt{n_+}},$$

with high probability, where $\mathcal{R}_{n_+}(\mathcal{F}_Z)$ is the Rademacher complexity of the function class of Z_v. It remains to bound $\mathcal{R}_{n_+}(\mathcal{F}_Z)$. It can be seen that $Z_f = h[\langle v, \varphi(\boldsymbol{x})\rangle]$ where h is a Lipschitz continuos function with Lipschitz constant $\frac{\|\tilde{u}\|^2}{\lambda}$. Hence, due to Ledoux–Talagrand contraction inequality (see, e.g., (5.61) in [36]), $\mathcal{R}_{n_+}(\mathcal{F}_Z)$ is upperbounded by,

$$\mathcal{R}_{n_+}(\mathcal{F}_Z) \leq \frac{2\|\tilde{u}\|^2}{\lambda} \cdot \mathcal{R}_{n_+}(\mathcal{H}^*) \leq \frac{C_0 \cdot \|\tilde{u}\|^2}{\lambda\sqrt{n_+}},$$

where C_0 is a universal constant. The last inequality is due to Corollary 14.5 in [36]. Therefore

$$\left|\tilde{s}\hat{\mathbf{\Sigma}}_{\bar{v}}^+\tilde{u} - \mathbb{E}_+\left[\tilde{s}\hat{\mathbf{\Sigma}}_{\bar{v}}^+\tilde{u}\right]\right| \leq \frac{C_0 \cdot \|\tilde{u}\|^2}{\lambda\sqrt{n_+}} + \frac{\|\tilde{u}\|^2 \cdot \|\varphi(\boldsymbol{x})\|^2}{\lambda\sqrt{n_+}}$$

and similarly,

$$\left|\tilde{s}\hat{\mathbf{\Sigma}}_{\bar{v}}^-\tilde{u} - \mathbb{E}_-\left[\tilde{s}\hat{\mathbf{\Sigma}}_{\bar{v}}^-\tilde{u}\right]\right| \leq \frac{C_0 \cdot \|\tilde{u}\|^2}{\lambda\sqrt{n_-}} + \frac{\|\tilde{u}\|^2 \cdot \|\varphi(\boldsymbol{x})\|^2}{\lambda\sqrt{n_-}}.$$

Therefore,

$$|b| \leq \frac{C_0 \cdot \|\tilde{u}\|^2}{\lambda\sqrt{n_{\min}}} + \frac{\|\tilde{u}\|^2 \cdot \|\varphi(\boldsymbol{x})\|^2}{\lambda\sqrt{n_{\min}}}, \tag{20}$$

with high probability. Substituting (19)and (20) into (17), we get

$$\langle \tilde{s}^\top, -\nabla_v \ell(v^*) - \lambda v^*\rangle + \frac{\max\left(C_0, \|\varphi(\boldsymbol{x})\|^2\right) \cdot \|\tilde{u}\|^2}{\lambda\sqrt{n_{\min}}} \geq \|\tilde{u}\|^2 - \frac{1}{2}\|\tilde{u}\|^2.$$

Using triangle inequality and Hölder's inequality, we have

$$-\langle \tilde{s}, \nabla_v \ell(v^*)\rangle + \|\tilde{u}\| \|(\boldsymbol{\Sigma}_{v^*} + \lambda \boldsymbol{I})^{-1} \lambda v^*\| + \frac{\max(C_0, 1)\|\tilde{u}\|^2}{\lambda\sqrt{n_{\min}}} \geq \frac{1}{2}\|\tilde{u}\|^2. \tag{21}$$

Due to Assumption 1, $\mathbb{E}[\nabla_v \ell(v^*)] = 0$. Hence,

$$\begin{aligned} -\langle \tilde{s}, \nabla_v \ell(v^*)\rangle =& -\langle \tilde{s}, \mathbb{E}[\nabla_v \ell(v^*)]\rangle - \langle \tilde{s}, \nabla_v \ell(v^*) - \mathbb{E}[\nabla_v \ell(v^*)]\rangle \\ =& 0 - \langle \tilde{s}, \nabla_v \ell(v^*) - \mathbb{E}[\nabla_v \ell(v^*)]\rangle. \end{aligned}$$

We can see

$$|\langle \tilde{s}, \nabla_v \ell(v^*) - \mathbb{E}[\nabla_v \ell(v^*)]\rangle| \leq \frac{\|\tilde{u}\|}{\lambda} \|\nabla_v \ell(v^*) - \mathbb{E}[\nabla_v \ell(v^*)]\| \leq \frac{C_1 \|\tilde{u}\|}{\lambda\sqrt{n_{\min}}} \tag{22}$$

holds with high probability and C_1 is a universal constant (due to Lemma 2).

Moreover, since $v^* \in \mathcal{R}(\boldsymbol{\Sigma}_{v^*}^{\beta})$, there exists $g \in \mathcal{H}$, $v^* = \boldsymbol{\Sigma}_{v^*}^{\beta} g$. Notice $\boldsymbol{\Sigma}_{v^*}$ is a bounded, compact, self-adjoint linear operator (see Section G). Therefore, Hilbert-Schmidt Theorem indicates, $\boldsymbol{\Sigma}_{v^*} = \sum_i \alpha_i \psi_i \langle \psi_i, \cdot\rangle$, where ψ_i, α_i are eigenfunctions and eigenvalues of $\boldsymbol{\Sigma}_{v^*}$ respectively. Hence,

$$\begin{aligned} \|(\boldsymbol{\Sigma}_{v^*} + \lambda \boldsymbol{I})^{-1} v^* \lambda\| = \|(\boldsymbol{\Sigma}_{v^*} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\Sigma}_{v^*}^{\beta} g \lambda\| \leq& \left\| \sum_i \langle \boldsymbol{\psi}_i, g\rangle \boldsymbol{\psi}_i \cdot \frac{\alpha_i^{\beta} \lambda}{\alpha_i + \lambda} \right\| \\ \leq& \lambda^{\beta} \left\| \sum_i \langle \boldsymbol{\psi}_i, g\rangle \right\| \leq \|\boldsymbol{\Sigma}_{v^*}{}^{-\beta} v^*\| \cdot \lambda^{\beta}. \end{aligned} \tag{23}$$

Combine (21), (22) and (23) and cancel $\|\tilde{u}\|$, we can conclude that

$$\frac{C_1}{\lambda\sqrt{n_{\min}}} + \|\boldsymbol{\Sigma}_{v^*}{}^{-\beta} v^*\| \cdot \lambda^{\beta} + \frac{\max\left(C_0, \|\varphi(\boldsymbol{x})\|^2\right)}{\lambda\sqrt{n_{\min}}} \geq \frac{1}{2}\|\tilde{u}\|,$$

with high probability. Set $\lambda = \frac{\max(C_1, C_0, 1)}{n_{\min}^{1/4}}$, we have

$$\frac{2}{n_{\min}^{1/4}} + \frac{\max(C_1, C_0, 1)^{\beta} \|\boldsymbol{\Sigma}_{v^*}^{-\beta} v^*\|}{n_{\min}^{\beta/4}} \geq \frac{1}{2}\|\tilde{u}\|,$$

holds with high probability. Therefore, $\exists N_2$, when $n_{\min} > N_2$, $\|\tilde{u}\| = O_p(n_{\min}^{-\beta/4})$.

Since $\|\tilde{u}\| = o_p(1)$, as long as $\delta \geq K \cdot n_{\min}^{-\beta/4}$ where $K > 0$ is a constant, there exists a constant N such that, when $n_{\min} > N$, $\tilde{v}$ is in the interior of $\mathcal{H}^*$ with high probability. When this happens, the constraint $v \in \mathcal{H}^*$ is no longer active. This means $\tilde{v}$ is a stationary point of the objective function in (13). Moreover, $\tilde{v} \in \mathcal{H}^*$, so it is in the feasible region of (7) thanks to Assumption 2. This further indicates that $\tilde{v}$ is also a solution to (7). As (7) is a strictly convex optimization problem, $\tilde{v}$ is also its only solution. Therefore $\tilde{v} = \hat{v}$ and $\|\hat{v} - v^*\| = \|\tilde{v} - v^*\| = O_p(n_{\min}^{-\beta/4})$.

□

Lemma 2. *Given any $v^* \in \mathcal{H}$ such that $\mathbb{E}[\nabla_v \ell(v^*)] = 0$, if $\|\varphi(\boldsymbol{x})\|_{\mathcal{H}} \leq B$ then*

$$P(\|\nabla_v \ell(v^*)\|_{\mathcal{H}} > \delta) \leq 4 \exp\left(-\frac{n_{\min}\delta^2}{B^2}\right).$$

Proof. Write down the definition of $\nabla_v \ell(v^*)$. Notice

$$\nabla_v \ell(v^*) = \underbrace{-\frac{1}{n_+} \sum_{i=1}^{n_+} \cos\langle v, \varphi(\boldsymbol{x}_i)\rangle \varphi(\boldsymbol{x}_i)}_{a} + \underbrace{\frac{1}{n_-} \sum_{i=1}^{n_-} \sin\langle v, \varphi(\boldsymbol{x}_i)\rangle \varphi(\boldsymbol{x}_i)}_{b}.$$

By using Hilbert-space Hoeffding's inequality [30], we know for all $\delta_a, \delta_b > 0$

$$P(\|a - \mathbb{E}[a]\|_{\mathcal{H}} > \delta_a) \leq 2\exp\left(-\frac{Cn_+\delta_a^2}{B^2}\right) \text{ and } P(\|b - \mathbb{E}[b]\|_{\mathcal{H}} > \delta_b) \leq 2\exp\left(-\frac{Cn_-\delta_b^2}{B^2}\right),$$

where C is a constant. Let $\delta = \delta_a = \delta_b$,

$$\begin{aligned}
P(\|a+b\|_{\mathcal{H}} > 2\delta) =& P(\|a+b-(\mathbb{E}[a]+\mathbb{E}[b])\|_{\mathcal{H}} > 2\delta) \\
\leq& P(\|a-\mathbb{E}[a]\|_{\mathcal{H}} + \|b-\mathbb{E}[b]\|_{\mathcal{H}} > \delta_a + \delta_b) \\
\leq& P(\|a-\mathbb{E}[a]\|_{\mathcal{H}} > \delta_a) + P(\|b-\mathbb{E}[b]\|_{\mathcal{H}} > \delta_b) \\
\leq& 4\exp\left(-\frac{Cn_{\min}\delta^2}{B^2}\right),
\end{aligned}$$

where the first equality used the condition that $\mathbb{E}[\nabla_v \ell(v^*)] = \mathbb{E}[a] + \mathbb{E}[b] = 0$. This completes the proof. □

F Proof of Proposition 3

Proof. We start from the definition of $\mathbb{E}[\nabla_v \ell(v^*)]$:

$$\begin{aligned}
-\mathbb{E}[\nabla_v \ell(v^*)] =& \mathbb{E}_+\left[\frac{1}{n_+}\sum_{i=1}^{n}\cos\langle v^*, \varphi(\boldsymbol{x}_i)\rangle\varphi(\boldsymbol{x}_i)\right] - \mathbb{E}_-\left[\frac{1}{n_-}\sum_{i=1}^{n}\sin\langle v^*, \varphi(\boldsymbol{x}_i)\rangle\varphi(\boldsymbol{x}_i)\right] \\
=& \mathbb{E}_+\left[\cos\langle v^*, \varphi(\boldsymbol{x})\rangle\varphi(\boldsymbol{x})\right] - \mathbb{E}_-\left[\frac{\sin\langle v^*, \varphi(\boldsymbol{x})\rangle}{\cos\langle v^*, \varphi(\boldsymbol{x})\rangle}\cos\langle v^*, \varphi(\boldsymbol{x})\rangle\varphi(\boldsymbol{x})\right] \\
=& \mathbb{E}_+\left[\cos\langle v^*, \varphi(\boldsymbol{x})\rangle\varphi(\boldsymbol{x})\right] - \mathbb{E}_-\left[\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\cos\langle v^*, \varphi(\boldsymbol{x})\rangle\varphi(\boldsymbol{x})\right] \\
=& \mathbb{E}_+\left[\cos\langle v^*, \varphi(\boldsymbol{x})\rangle\varphi(\boldsymbol{x})\right] - \mathbb{E}_+\left[\cos\langle v^*, \varphi(\boldsymbol{x})\rangle\varphi(\boldsymbol{x})\right] = 0,
\end{aligned}$$

where the third equality is due to the fact that $\langle v^*\varphi(\boldsymbol{x})\rangle = \operatorname{atan}\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}$. Since $p_+/p_- \in [0,\infty)$, $\langle v, \varphi(\boldsymbol{x})\rangle \in [0, \pi/2)$. As v^* is unique by assumption, Assumption 1 holds. □

G Properties of Operator $\boldsymbol{\Sigma}_{v_0}$

By construction, it is easy to verify that $\boldsymbol{\Sigma}_{v_0}$ is self-adjoint.

First, we prove that the integral operator

$$\boldsymbol{\Sigma}_{v_0} u = \mathbb{E}_+[\sin\langle v_0, \varphi(\boldsymbol{x})\rangle\varphi(\boldsymbol{x}) \cdot u(\boldsymbol{x})] + \mathbb{E}_-[\cos\langle v_0, \varphi(\boldsymbol{x})\rangle\varphi(\boldsymbol{x}) \cdot u(\boldsymbol{x})],$$

is a bounded operator. For all $u \in \mathrm{Ball}(0,1)$, where $\mathrm{Ball}(0,1)$ is the unit ball in $\|\cdot\|_{\mathcal{H}}$,

$$\begin{aligned}
\|\boldsymbol{\Sigma}_{v_0} u\|_{\mathcal{H}} \leq& \|\mathbb{E}_+[\sin\langle v_0, \varphi(\boldsymbol{x})\rangle\varphi(\boldsymbol{x}) \cdot u(\boldsymbol{x})]\|_{\mathcal{H}} + \|\mathbb{E}_-[\cos\langle v_0, \varphi(\boldsymbol{x})\rangle\varphi(\boldsymbol{x}) \cdot u(\boldsymbol{x})]\|_{\mathcal{H}} \\
\leq& \mathbb{E}_+[\|\sin\langle v_0, \varphi(\boldsymbol{x})\rangle\varphi(\boldsymbol{x})\|_{\mathcal{H}} \cdot \|u(\boldsymbol{x})\|_{\mathcal{H}}] + \mathbb{E}_-[\|\cos\langle v_0, \varphi(\boldsymbol{x})\rangle\varphi(\boldsymbol{x})\|_{\mathcal{H}} \cdot \|u(\boldsymbol{x})\|_{\mathcal{H}}] \\
\leq& \mathbb{E}_+[\|\varphi(\boldsymbol{x})\|_{\mathcal{H}} \cdot \|u(\boldsymbol{x})\|_{\mathcal{H}}] + \mathbb{E}_-[\|\varphi(\boldsymbol{x})\|_{\mathcal{H}} \cdot \|u(\boldsymbol{x})\|_{\mathcal{H}}] \\
\leq& \mathbb{E}_+[\|\varphi(\boldsymbol{x})\|_{\mathcal{H}}] + \mathbb{E}_-[\|\varphi(\boldsymbol{x})\|_{\mathcal{H}}].
\end{aligned}$$

Hence, $\boldsymbol{\Sigma}_{v_0}$ is a bounded operator as long as $\|\varphi(\boldsymbol{x})\|_{\mathcal{H}}$ is bounded.

Second, we show $\boldsymbol{\Sigma}_{v_0}$ is trace class hence compact. Let $\boldsymbol{\psi}_i, i \in \mathbb{N}$ be an orthonormal basis in $\mathcal{H}$, then

$$\begin{aligned}
&\sum_i \langle \boldsymbol{\psi}_i, \boldsymbol{\Sigma}_{v_0}\boldsymbol{\psi}_i\rangle \\
=& \mathbb{E}_+[\sin\langle v_0, \varphi(\boldsymbol{x})\rangle \sum_{i\in\mathbb{N}}\langle \boldsymbol{\psi}_i, \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})\boldsymbol{\psi}_i\rangle] + \mathbb{E}_-[\cos\langle v_0, \varphi(\boldsymbol{x})\rangle \sum_{i\in\mathbb{N}}\langle \boldsymbol{\psi}_i, \varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})\boldsymbol{\psi}_i\rangle] \\
=& \mathbb{E}_+[\sin\langle v_0, \varphi(\boldsymbol{x})\rangle \sum_{i\in\mathbb{N}}\langle \boldsymbol{\psi}_i, \varphi(\boldsymbol{x})\rangle^2] + \mathbb{E}_-[\cos\langle v_0, \varphi(\boldsymbol{x})\rangle \sum_{i\in\mathbb{N}}\langle \boldsymbol{\psi}_i, \varphi(\boldsymbol{x})\rangle^2] \\
=& \mathbb{E}_+[\sin\langle v_0, \varphi(\boldsymbol{x})\rangle \cdot \|\varphi(\boldsymbol{x})\|_{\mathcal{H}}^2] + \mathbb{E}_-[\cos\langle v_0, \varphi(\boldsymbol{x})\rangle \cdot \|\varphi(\boldsymbol{x})\|_{\mathcal{H}}^2] < \infty
\end{aligned}$$

holds as long as $\|\varphi(\boldsymbol{x})\|_{\mathcal{H}}$ is bounded. This shows $\boldsymbol{\Sigma}_{v_0}$ is trace-class and therefore, compact.

H Proof of Proposition 5

Proof. Let us define for $\alpha \in [0, .5]$,

$$\tilde{F}_-^*(\cdot, \alpha) := 1 - [(1-\alpha)F_-^*(\cdot) + \alpha F_+^*(\cdot)], \quad \tilde{F}_+^*(\cdot, \alpha) := 1 - [\alpha F_-^*(\cdot) + (1-\alpha)F_+^*(\cdot)].$$

We can see that $\boldsymbol{r}(\tau, \alpha) := (\tilde{F}_-^*(\tau, \alpha), \tilde{F}_+^*(\tau, \alpha))$ is a parameterization for the space between ROC^* and the diagonal from $(0, 0)$ to $(1, 1)$. We can compute the surface area using the surface integral formula:

$$A_0 := \int_{\mathrm{dom}(\tau)} \int_{[0,.5]} \|\partial_\tau \boldsymbol{r}(\tau, \alpha) \times \partial_\alpha \boldsymbol{r}(\tau, \alpha)\| \, \mathrm{d}\alpha \mathrm{d}\tau,$$

where $\partial_\tau \boldsymbol{r}(\tau, \alpha) = \begin{bmatrix} \partial_\tau \tilde{F}_-^*(\tau, \alpha) \\ \partial_\tau \tilde{F}_+^*(\tau, \alpha) \\ 0 \end{bmatrix}$ and $\partial_\alpha \boldsymbol{r}(\tau, \alpha) = \begin{bmatrix} \partial_\alpha \tilde{F}_-^*(\tau, \alpha) \\ \partial_\alpha \tilde{F}_+^*(\tau, \alpha) \\ 0 \end{bmatrix}$. It can be seen that $\partial_\alpha \boldsymbol{r}(\tau, \alpha) = \begin{bmatrix} F_-^*(\tau) - F_+^*(\tau) \\ F_+^*(\tau) - F_-^*(\tau) \\ 0 \end{bmatrix}$ for all α. Rewrite A_0:

$$\begin{aligned} A_0 &= \int_{\mathrm{dom}(\tau)} \int_{[0,.5]} \left| \left[F_-^*(\tau) - F_+^*(\tau)\right] \partial_\tau \tilde{F}_+^*(\tau, \alpha) - \left[F_+^*(\tau) - F_-^*(\tau)\right] \partial_\tau \tilde{F}_-^*(\tau, \alpha) \right| \, \mathrm{d}\alpha \mathrm{d}\tau, \\ &= \int_{\mathrm{dom}(\tau)} \int_{[0,.5]} \left| \left[F_-^*(\tau) - F_+^*(\tau)\right] \left(\partial_\tau F_+^*(\tau) + \partial_\tau F_-^*(\tau)\right) \right| \, \mathrm{d}\alpha \mathrm{d}\tau, \\ &= \int_{\mathrm{dom}(\tau)} \int_{[0,.5]} \|\boldsymbol{a}(\tau) \times \boldsymbol{b}(\tau)\| \, \mathrm{d}\alpha \mathrm{d}\tau, \end{aligned} \tag{24}$$

where $\boldsymbol{a}(\tau) = \begin{bmatrix} F_-^*(\tau) - F_+^*(\tau) \\ F_+^*(\tau) - F_-^*(\tau) \\ 0 \end{bmatrix}$ and $\boldsymbol{b}(\tau) = \begin{bmatrix} \partial_\tau F_-^*(\tau) \\ \partial_\tau F_+^*(\tau) \\ 0 \end{bmatrix}$. Both $\boldsymbol{a}$ and $\boldsymbol{b}$ are free from α. Rewriting the cross product in (24) in a different form, we obtain

$$\begin{aligned} A_0 &= \sqrt{2} \int_{\mathrm{dom}(\tau)} \int_{[0,.5]} \sin(\theta(\tau)) \left|F_-^*(\tau) - F_+^*(\tau)\right| \sqrt{\partial_\tau F_+^*(\tau)^2 + \partial_\tau F_-^*(\tau)^2} \, \mathrm{d}\alpha \mathrm{d}\tau, \\ &= \frac{\sqrt{2}}{2} \int_{\mathrm{dom}(\tau)} \sin(\theta(\tau)) \left|F_-^*(\tau) - F_+^*(\tau)\right| \sqrt{\partial_\tau F_+^*(\tau)^2 + \partial_\tau F_-^*(\tau)^2} \, \mathrm{d}\tau, \\ &= \frac{\sqrt{2}}{2} \int_{\mathrm{dom}(\tau)} \sin(\theta(\tau)) f_-^*(\tau) \left|F_-^*(\tau) - F_+^*(\tau)\right| \sqrt{\left(\frac{f_+^*(\tau)}{f_-^*(\tau)}\right)^2 + 1} \, \mathrm{d}\tau, \end{aligned} \tag{25}$$

where $\theta(\tau)$ is the angle between $\boldsymbol{a}(\tau)$ and $\boldsymbol{b}(\tau)$. $\boldsymbol{b}(\tau)$ is the tangent vector of the ROC^*. Knowing the slope of ROC^* is the likelihood ratio (see Section 4.1) and $\boldsymbol{a}(\tau)$ points at the 45 degree downward regardless of τ, we can see $\theta(\tau) = \left[\mathrm{atan}\, \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right] + \frac{\pi}{4}$. Using the fact that $\frac{f_+^*(t(\boldsymbol{x}))}{f_-^*(t(\boldsymbol{x}))} = \frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}$ and the law of unconscious statistician:

$$A_0 = \frac{\sqrt{2}}{2} \mathbb{E}_{p_-} \left\{ \sin\left[\left(\mathrm{atan}\, \frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right) + \frac{\pi}{4}\right] \left|F_-^*(t^*(\boldsymbol{x})) - F_+^*(t^*(\boldsymbol{x}))\right| \sqrt{\left(\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right)^2 + 1} \right\}$$

Replacing $\sqrt{\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}^2 + 1}$ with its Fenchel dual as introduced in Section 4.1 and pulling the $\sup$ out of the expectation yields the desired result.

Differentiating the objective (11) with respect to v and setting the derivative to zero, we can see that superemum is attained at $v^* = \mathrm{atan}\, \frac{p_+}{p_-}$.

□

I Experiment Setup

In the wall clock time comparison in Section 5, we implement the AUC maximization and two-step procedure using MATLAB's Optimization Toolbox. Both methods are written in fully vectorized code. The first order derivatives are provided to the $\mathrm{fmincon}$ and $\mathrm{fminunc}$ to accelerate the computation. Code can be found in the supplementary material.

In Section 6.2, we reduce the dimension of CIFAR-10 dataset to 50. We first train a residual neural network [15] using logistic regression on all 10 classes. This 103-layer network structure was borrowed from a MATLAB tutorial (`https://www.mathworks.com/help/deeplearning/ug/train-residual-network-for-image-classification.html`). MATLAB provides a pre-trained version of this network. To obtain bounded features, we append a fully connected linear layer (output dimension 50) and a bounded activation layer (clipped-relu) to the last average pooling layer in the network. We freeze the earlier layers and only train the last two layers for 5 epochs.

The dataset and the code that reproduces Figure 4 can be found in the supplementary materials. We invite reviewers to reproduce our results.

J Estimating $\log\left[\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right]$

We can also leverage that v^* is the arctangent of the likelihood ratio and introduce mild assumptions on p_+ and p_-. When $p_+(\boldsymbol{x})$ and $p_-(\boldsymbol{x})$ are both members of the exponential family and share the same sufficient statistic $\boldsymbol{h}(\boldsymbol{x}) \in \mathbb{R}^m$, then $\exists \boldsymbol{v}^* \in \mathbb{R}^m$ such that $\log\left[\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right] = \langle \boldsymbol{v}^*, \boldsymbol{h}(\boldsymbol{x})\rangle + C$, where C is a constant. If we choose to parameterize the log likelihood ratio using a linear model, $\langle \boldsymbol{v}, \boldsymbol{h}(\boldsymbol{x})\rangle + v_0$, then (6) becomes

$$(\hat{\boldsymbol{v}}, \hat{v}_0) := \operatorname*{argmax}_{\substack{\boldsymbol{v}\in\mathbb{R}^m,\\ v_0\in\mathbb{R}}} \frac{1}{n_+}\sum_{i=1}^{n_+} \sin[\operatorname{atan}\exp(\langle \boldsymbol{v}, \boldsymbol{h}(\boldsymbol{x}_i)\rangle + v_0)] + \frac{1}{n_-}\sum_{i=1}^{n_-} \cos[\operatorname{atan}\exp(\langle \boldsymbol{v}, \boldsymbol{h}(\boldsymbol{x}_i)\rangle + v_0)]. \tag{26}$$

Note we do not have to restrict the optimization to a bounded function family as $\log\left[\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right] \in \mathbb{R}$, and $\langle \hat{\boldsymbol{v}}, \boldsymbol{h}(\boldsymbol{x})\rangle + \hat{v_0}$ is an estimate of the likelihood ratio.

In this paper, we focus on (7) since the objective in (26) is non-convex with respect to $\boldsymbol{v}$ thus presents extra challenges in the theoretical analysis, although (26) is easier to implement in practice due to its unconstrained nature.

K Numerical Simulation of $\operatorname{atan}\frac{p_+}{p_-}$ Estimation

We draw 100 samples from X_+ and X_- and solve (7) to estimate the arctangent density ratio. The estimated arctangent density ratio with standard deviation (over 72 runs) are plotted in Figure 5. We use Gaussian kernel and hyperparameters (kernel bandwidth and regularization parameter) are tuned using cross validation.

We observe that the estimated arctangent ratio using the proposed method is very close to the ground truth and has a small standard deviation.

L Comparison with Convergence Result in Nguyen et al. [26]

The convergence of (log) density ratio estimation have been developed for two KL divergence based estimators [26]. However, these convergence theories are not general theories for arbitrary f-divergences. Thus their proofs cannot be easily applied to our ROC divergence.

Moreover, Theorem 1 is not a minor modification of convergence theories in [26]. Specifically, Nguyen et al. [26] prove the likelihood ratio converges in Hellinger distance, while we prove the arctangent likelihood ratio converges in Hilbert space norm. The proofs rely on completely different machinery and assumptions. These technical results depend on the variational objective functions the estimators maximize and are not interchangeable.

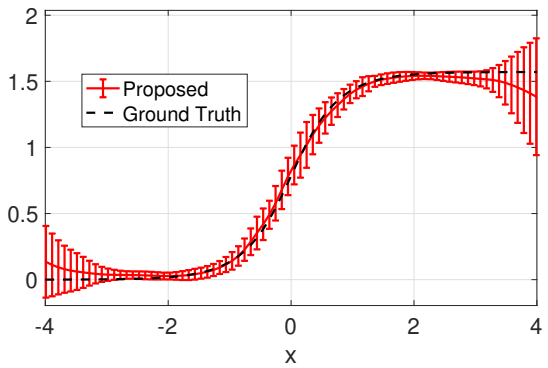


Figure 5: Estimation of the arctangent density ratio function.