On Estimating ROC Arc Length and Lower Bounding Maximal AUC for Imbalanced Classification

Song Liu [1]

Abstract

Some astrophysical datasets have extremely imbalanced classes. ROC curves are often used to measure the performance of classifiers on imbalanced datasets due to their insensitivity to class distributions. This paper studies the arc length of ROC curves and provides a novel way of lower bounding the maximal AUC. We show that when the data likelihood ratio is used as the score function, the arc length of the corresponding ROC curve gives rise to a novel f-divergence. This f-divergence can be expressed using a variational objective and estimated only using samples from the positive and negative data distributions. Moreover, we show the space below the optimal ROC curve can be expressed as a similar variational objective depending on the arctangent likelihood ratio. These new insights lead to a novel two-step procedure for finding a good score function by lower bounding the maximal AUC. Experiments on RR-Lyrae datasets show that the proposed two-step procedure achieves good AUC performance in imbalanced binary classification tasks while being less computationally demanding.

1. Introduction

In astrophysical datasets, classes are often imbalanced. For example, in RR-Lyrae classification dataset (Sesar et al., 2010), non-background observations only account for 0.52% samples in the dataset. In machine learning, ROC curves have been used to compare the performance of different classification algorithms on imbalanced datasets (Fawcett, 2006; Flach, 2016). Indeed, the Area Under the Curve (AUC) encodes a classifier's ranking accuracy, making it a preferable performance metric for imbalanced class classification (Fawcett, 2006; Cortes & Mohri, 2003). A classifier is considered superior if its AUC is larger than its competition. However, the classic AUC maximization has a computational complexity $O(n^2)$, which makes it impractical on large-scale datasets. Moreover, ROC curves have been applied to compare distributions recently. Examples include analyzing the mode collapsing issue of Generative Adversarial nets (GAN) (Lin et al., 2018), and diagnosing the performance of an amortized Markov Chain Monte Carlo (Hermans et al., 2020). Research along this line suggests that we can measure the differences between distributions by looking at ROC curves.

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 express the arc length using a variational objective and approximate it using only samples from two distributions. Moreover, by parameterizing the ROC curve of mixtures distributions, we can express the space below the optimal ROC curve via a similar variational objective and approximately maximize a lower bound to the AUC. We highlight the similarity between this lower bound maximization and the classic AUC maximization (Cortes & Mohri, 2003). While less computationally demanding, our approximated optimal score achieves comparable performance to an AUC maximizer in the RR-Lyrae classification dataset.

2. Expressing ROC Arc Length

2.1. 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 score function $t(\boldsymbol{x})$ transforms a sample $\boldsymbol{x}$ into a real-valued score.

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

- False Positive Rate (FPR), $\tilde{F}_- := 1 - F_-$
- True Positive Rate (TPR), $\tilde{F}_+ := 1 - F_+$
- The ROC curve of a score function t: the graph of

[1]University of Bristol, UK. Correspondence to: Song Liu <song.liu@bristol.ac.uk>.

ICML 2022 Workshop on Machine Learning for Astrophysics, Baltimore, Maryland, USA, 2022.

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.

2.2. 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 $\tilde{F}_+$:

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

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 (Edwards & Metz, 2007; 2008) have proposed a performance metric computed over the "ROC hypersurface" and (1) is used to justify such a metric in a binary classification setting.

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

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

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 shows, for any given t, $\widehat{\mathrm{ROC}}(t)$ is a good discrepancy for measuring positive and negative scores (i.e., score 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 $p_-(\boldsymbol{x})$:

$$\widehat{\mathrm{ROC}}(t) = \mathbb{E}_{p_-}\sqrt{\left[\frac{f_+(t(\boldsymbol{x}))}{f_-(t(\boldsymbol{x}))}\right]^2 + 1}.$$

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 (Neyman & Pearson, 1933), $\mathrm{ROC}(t^*)$ 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^*.

It can be shown that $\frac{f_+(t^*(\boldsymbol{x}_0))}{f_-(t^*(\boldsymbol{x}_0))} = \frac{p_+(\boldsymbol{x}_0)}{p_-(\boldsymbol{x}_0)}$. When $\gamma(\boldsymbol{x}) = \boldsymbol{x}$, this equality expresses a known result (Eguchi & Copas, 2002).

Therefore, $\widehat{\mathrm{ROC}^*}$ takes an elegant form free from t^* or γ:

$$\widehat{\mathrm{ROC}^*} = \mathbb{E}_{p_-(\boldsymbol{x})}\sqrt{\left[\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right]^2 + 1}.$$

Equivalently, we can write

$$\widehat{\mathrm{ROC}^*} - \sqrt{2} = \mathbb{E}_{p_-(\boldsymbol{x})}\left[g\left(\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right)\right]. \tag{3}$$

We can see that the same f-divergence arises from computing the arc length of ROC^*. However, unlike the f-divergence given in (2), (3) is an f-divergence between *data* distributions, not score distributions. It shows that as long as we use the optimal scores, the arc length of 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, (3) has not been presented in literature before.

3. Estimating the Arc Length of ROC^*

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 (Hiriart-Urruty & Lemaréchal, 2004), authors in (Nguyen et al., 2010) 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 $\widehat{\mathrm{ROC}^*}$ using the above variational representation, we obtain:

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

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

$$\widehat{\mathrm{ROC}^*} = \sup_{v \in [0,\pi/2]} \mathbb{E}_{p_+} \sin[v(\boldsymbol{x})] + \mathbb{E}_{p_-} \cos[v(\boldsymbol{x})]. \tag{4}$$

Differentiating the objective in (4) for v and setting the derivative 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

$$\partial_s \tilde{F}_+(\tilde{F}_-^{-1}(s_0)) = \frac{f_+(\tilde{F}_-^{-1}(s_0))}{f_-(\tilde{F}_-^{-1}(s_0))} = \frac{p_+(\boldsymbol{x}_0)}{p_-(\boldsymbol{x}_0)}, \quad (5)$$

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^*. The boundary constraint for $v(\boldsymbol{x})$ in (4) coincides with the fact that the slope angle of any ROC curve is in between 0 and $\pi/2$.

We can simply restrict v to a bounded (parametric/non-parametric) function class $\mathcal{F}$ and solve the sample version of (4). Then we have the following sample objective for estimating the arctangent likelihood ratio.

$$\hat{v} := \underset{\substack{v \in [0, \pi/2], \\ v \in \mathcal{F}}}{\operatorname{argmax}} \frac{1}{n_+} \sum_{i=1}^{n_+} \sin(v(\boldsymbol{x}_i^+)) + \frac{1}{n_-} \sum_{i=1}^{n_-} \cos(v(\boldsymbol{x}_i^-)), \quad (6)$$

where $\hat{v}$ is an estimator of the arctangent likelihood ratio.

4. Lower Bounding the Maximal AUC

Finding a score function t that approximately maximizes AUC is an important task in binary classification (Cortes & Mohri, 2003; Gao et al., 2013; Ying et al., 2016). 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:

$$\begin{aligned} \mathrm{AUC}^* &= \underbrace{\sup_t \mathbb{E}_{p_-} \mathbb{E}_{p_+} \left[\mathbb{1}\left(t(\boldsymbol{x}^+) \geq t(\boldsymbol{x}^-) \right) \right]}_{(i)} \\ &\geq \sup_{t \in \mathcal{F}'} \underbrace{\mathbb{E}_{p_-} \mathbb{E}_{p_+} \left[L\left(t(\boldsymbol{x}^+), t(\boldsymbol{x}^-) \right) \right]}_{(ii)}, \end{aligned} \quad (7)$$

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 $t = \gamma\left(\frac{p_+}{p_-}\right)$ where γ is a strictly increasing function. Replacing the expectations in (ii) with sample averages, we obtain the optimization problem of AUC maximization (Cortes & Mohri, 2003; Gao et al., 2013). 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 $\operatorname{atan} \frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}$. We have seen that how $(\tilde{F}_-(\tau), \tilde{F}_+(\tau))$ parameterizes an ROC curve in Section 2.2. In fact, the space between ROC^* and the diagonal line can be similarly parameterized by considering ROC curves of *mixtures* between 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: When fixing α and varying τ, the coordinates give rise to a smooth curve 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 1 visualizes this parameterization. Now the surface area sandwiched between ROC^* and the diagonal line can be computed using a surface integration:

$$\begin{aligned} &\mathrm{AUC}^* - .5 = \\ &\int_{\mathrm{dom}(\tau)} \int_{[0,.5]} \left\| \partial_\tau \boldsymbol{r}(\tau, \alpha) \times \partial_\alpha \boldsymbol{r}(\tau, \alpha) \right\| \mathrm{d}\alpha \mathrm{d}\tau, \end{aligned} \quad (8)$$

where $\times$ denotes the cross product. After some algebra and applying the Fenchel duality technique in Section 3, we prove that AUC^* can be expressed as the supremum of a variational objective similar to (4):

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

$$\begin{aligned} A := \sup_{v \in [0, \pi/2]} &\mathbb{E}_{p_+} \left[w\left(\operatorname{atan} \frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})} \right) \sin[v(\boldsymbol{x})] \right] \\ &+ \mathbb{E}_{p_-} \left[w\left(\operatorname{atan} \frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})} \right) \cos[v(\boldsymbol{x})] \right], \end{aligned} \quad (9)$$

where $w(\tau) := \sin(\tau + \frac{\pi}{4}) \cdot |F_+^*(\tau) - F_-^*(\tau)|$. *The supremum of* (9) *is attained at* $v^* = \operatorname{atan} \frac{p_+}{p_-}$.

Readers can find the proof in Appendix A 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. Therefore, we propose the following two-step procedure in Algorithm 1 to approximately lowerbound A.

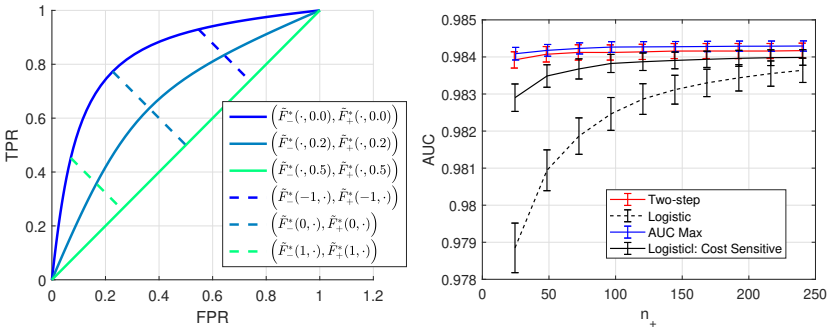

Figure 1. Left: $(\tilde{F}_-^*(\tau,\alpha), \tilde{F}_+^*(\tau,\alpha))$ parameterizes the space between ROC* and the diagonal line in $[0,1]^2$. This plot is created by setting $p_+ = \mathcal{N}(1,1)$, $p_- = \mathcal{N}(-1,1)$ and $t^*(x) = \frac{1}{2}\log\frac{p_+(x)}{p_-(x)} = x$. Right: Testing AUC on RR-Lyrae dataset. Error bars represent standard error over 96 runs.

Algorithm 1 Two-step Procedure for lower bounding A

(1) Obtain $\hat{t}(\boldsymbol{x}) := \langle \hat{v}, \phi(\boldsymbol{x}) \rangle$ using (6). 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 (9) by restricting v to a feasible function class and plugging in estimates obtained in the earlier step, i.e.,

$$\hat{v} := \underset{\substack{v\in[0,\pi/2],\\ v\in\mathcal{F}}}{\operatorname{argmax}} \frac{1}{n_+}\sum_{i=1}^{n_+} \hat{w}\left[\hat{t}(\boldsymbol{x}_i^+)\right]\cdot\sin[v(\boldsymbol{x}_i^+)] + \frac{1}{n_-}\sum_{i=1}^{n_-} \hat{w}\left[\hat{t}(\boldsymbol{x}_i^-)\right]\cdot\cos[v(\boldsymbol{x}_i^-)], \quad (10)$$

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

Note that (10) in Algorithm 1 is nothing but a weighted sample objective (6). Thus, it can be easily optimized by the same algorithm that solves a weighted version of (6) given the approximated weights in the first step. In practice, we simply run (6) 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*, we test the maximizer $\hat{v}$ of (10) in AUC maximization tasks. In the next section, we show that our two-step procedure Algorithm 1 achieves a promising AUC performance compared to a dedicated AUC maximizer.

Comparing to the traditional AUC maximizer, the proposed algorithm has a lower computational cost. Without loss of generality, assuming $n_+ = n_- = n$, the AUC maximization problem derived from the Wilcoxon-Mann-Whitney statistic (Hanley & McNeil, 1982) has a computational complexity $O(n^2)$. In comparison, (10) has a computational complexity $O(n)$. Computing $\hat{F}_+$ and $\hat{F}_-$ requires sorting our datasets which has an average computational 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 $\{\hat{t}(\boldsymbol{x}_i)\}_{i=1}^{n_+}$.

5. Imbalanced Classification on RR-Lyrae Dataset (Sesar et al., 2010)

In this section, we test if the $\hat{v}$ obtained in our two-step procedure (10) is indeed a good score function in terms of maximizing AUC in imbalanced classification. The performance is compared with a logistic regression classifier, a logistic regression classifier whose "class costs" are adjusted according to the class proportions, and an AUC maximizer which maximizes the empirical lower bound in (7). We set $L(a,b) := -(1-(a-b))^2$ in the AUC maximizer, as suggested in (Gao et al., 2013). Since our models are simple and we have sufficient samples, all methods use linear models with no regularization terms. Particularly, the $\hat{t}$ and $\hat{v}$ in our two-step algorithm is obtained using (6).

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. We repeated the experiment 96 times for each class using different random samples. We use 80% samples for training and 20% samples for testing. The training and testing subset is split randomly.

The average AUCs computed on the testing dataset, and their standard errors over 96 runs over different n_+ sample sizes are shown in the right plot of Figure 1. 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 $\hat{v}$ can be a good score function in imbalanced classification.

References

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

Edwards, D. C. and Metz, C. E. 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, pp. 21 – 30. SPIE, 2007.

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

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

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

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

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

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

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

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

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

Neyman, J. and Pearson, E. S. 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.

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

Sesar, B., Ivezić, Z., Grammer, S. H., Morgan, D. P., Becker, A. C., Jurić, M., De Lee, N., Annis, J., Beers, T. C., Fan, X., Lupton, R. H., Gunn, J. E., Knapp, G. R., Jiang, L., Jester, S., Johnston, D. E., and Lampeitl, H. Light curve templates and galactic distribution of rr lyrae stars from sloan digital sky survey stripe 82. *The Astrophysical Journal*, 708(1):717–741, January 2010.

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

A. Proof of Proposition 1

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 surface between ROC^* and the diagonal from $(0, 0)$ to $(1, 1)$. We can compute the surface area using the surface integral formula:

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

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:

$$\begin{aligned} A_0 &= \int_{\mathrm{dom}(\tau)} \int_{[0,.5]} \left| \left[F^*_-(\tau) - F^*_+(\tau)\right] \partial_\tau \tilde{F}^*_+(\tau, \alpha) - \left[F^*_+(\tau) - F^*_-(\tau)\right] \partial_\tau \tilde{F}^*_-(\tau, \alpha) \right| \mathrm{d}\alpha \mathrm{d}\tau, \\ &= \int_{\mathrm{dom}(\tau)} \int_{[0,.5]} \left| \left[F^*_-(\tau) - F^*_+(\tau)\right] \left(\partial_\tau F^*_+(\tau) + \partial_\tau F^*_-(\tau)\right) \right| \mathrm{d}\alpha \mathrm{d}\tau, \\ &= \int_{\mathrm{dom}(\tau)} \int_{[0,.5]} \|\boldsymbol{a}(\tau) \otimes \boldsymbol{b}(\tau)\| \, \mathrm{d}\alpha \mathrm{d}\tau, \end{aligned} \tag{11}$$

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 (11) in a different form, we obtain

$$\begin{aligned} A_0 &= \sqrt{2} \int_{\mathrm{dom}(\tau)} \int_{[0,.5]} \sin(\theta(\tau)) \left|F^*_-(\tau) - F^*_+(\tau)\right| \sqrt{\partial_\tau F^*_+(\tau)^2 + \partial_\tau F^*_-(\tau)^2} \, \mathrm{d}\alpha \mathrm{d}\tau, \\ &= \frac{\sqrt{2}}{2} \int_{\mathrm{dom}(\tau)} \sin(\theta(\tau)) \left|F^*_-(\tau) - F^*_+(\tau)\right| \sqrt{\partial_\tau F^*_+(\tau)^2 + \partial_\tau F^*_-(\tau)^2} \, \mathrm{d}\tau, \\ &= \frac{\sqrt{2}}{2} \int_{\mathrm{dom}(\tau)} \sin(\theta(\tau)) f^*_-(\tau) \left|F^*_-(\tau) - F^*_+(\tau)\right| \sqrt{\left(\frac{f^*_+(\tau)}{f^*_-(\tau)}\right)^2 + 1} \, \mathrm{d}\tau, \end{aligned} \tag{12}$$

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 (5)) and $\boldsymbol{a}(\tau)$ points at the 45 degree downward regardless of τ, we can see $\theta(\tau) = \left[\mathrm{atan}\, \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right] + \frac{\pi}{4}$.

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

Replacing $\sqrt{\left(\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right)^2 + 1}$ with its Fenchel dual as introduced in Section 3 and pulling the $\sup$ out of the expectation yields the desired result.

Differentiating the objective (9) with respect to v and setting the derivative to zero, we can see that superemum is attained at $v^* = \mathrm{atan}\, \frac{p_+}{p_-}$. □