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 $\partial_s^2F_+(F_-^{-1}(s;t^*)) \leq 0$. Moreover, at any FPR level $s \in [0,1]$, the Neyman-Pearson lemma [27] 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 $\mathrm{TV}(\mathbb{P}_+,\mathbb{P}_-)$ [34], 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 (8):

$$\tilde{v} := \operatorname*{argmin}_{v \in \mathcal{H}^*} \ell(v) + \frac{\lambda}{2}\|v\|^2 \tag{15}$$

Define $\tilde{u} := \tilde{v} - v^*$ and we have the following equality due to the KKT conditions of (15)

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

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 $g(v)$:

$$g(\tilde{v}) - g(v^*) = \nabla_v g(\bar{v})\tilde{u}, \tag{16}$$

where $\bar{v} = av^* + (1-a)\hat{v}$ 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 (16) into

$$\langle \tilde{s}, -\nabla_v \ell(v^*) - \lambda v^*\rangle = \langle \tilde{s}, [\nabla_f^2 \ell(\bar{v}) + \lambda \boldsymbol{I} + 2\tilde{\nu}\boldsymbol{I}]\tilde{u}\rangle, \tag{17}$$

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

$$\begin{aligned}\langle \tilde{s}, [\nabla_f^2 \ell(\bar{v}) + \lambda \boldsymbol{I} + 2\nu\boldsymbol{I}]\tilde{u}\rangle &\geq \langle \tilde{s}, [\nabla_f^2 \ell(\bar{v}) + \lambda \boldsymbol{I}]\tilde{u}\rangle \\ &\geq \underbrace{\langle (\boldsymbol{\Sigma}_{v^*} + \lambda \boldsymbol{I})^{-1}\tilde{u}, [\boldsymbol{\Sigma}_{v^*} + \lambda \boldsymbol{I}]\tilde{u}\rangle}_{\|\tilde{u}\|^2} - \underbrace{\langle \tilde{s}, [\boldsymbol{\Sigma}_{v^*} - \boldsymbol{\Sigma}_{\bar{v}}]\tilde{u}\rangle}_{a} - \underbrace{\langle \tilde{s}, [\boldsymbol{\Sigma}_{\bar{v}} - \nabla_f^2 \ell(\bar{v})]\tilde{u}\rangle}_{b} \\ &\geq \|\tilde{u}\|^2 - a - b. \end{aligned} \tag{18}$$

The first line is due to the fact that $\langle 2\nu\tilde{s}, \tilde{u}\rangle \geq 0$. Use the inequality (18) on (17), we get the inequality

$$\langle \tilde{s}, -\nabla_v \ell(v^*) - \lambda v^*\rangle \geq \|\tilde{u}\|^2 - a - b. \tag{19}$$

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

$$\sin\langle v^*, \varphi(\boldsymbol{x})\rangle - \sin\langle v^* + \boldsymbol{\delta}', \varphi(\boldsymbol{x})\rangle \leq \|\boldsymbol{\delta}'\| \cdot \|\varphi(\boldsymbol{x})\| \leq \delta \cdot \|\varphi(\boldsymbol{x})\|. \tag{20}$$

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 (20). 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

$$a \leq \frac{\|\tilde{u}\|^2}{2}. \tag{21}$$

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, [38]),

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

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 [38]), $\mathcal{R}_{n_+}(\mathcal{F}_Z)$ 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 [38]. 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,

$$|b| \leq \frac{C_0 \cdot \|\tilde{u}\|^2}{\lambda\sqrt{n_{\min}}} + \frac{\|\tilde{u}\|^2 \cdot \|\varphi(\boldsymbol{x})\|^2}{\lambda\sqrt{n_{\min}}}, \tag{22}$$

with high probability. Substituting (21)and (22) into (19), 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

$$-\langle \tilde{s}, \nabla_v \ell(v^*)\rangle + \|\tilde{u}\|\|(\boldsymbol{\Sigma}_{v^*} + \lambda \boldsymbol{I})^{-1}\lambda v^*\| + \frac{\max(C_0, 1)\|\tilde{u}\|^2}{\lambda\sqrt{n_{\min}}} \geq \frac{1}{2}\|\tilde{u}\|^2. \tag{23}$$

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

$$|\langle \tilde{s}, \nabla_v \ell(v^*) - \mathbb{E}[\nabla_v \ell(v^*)]\rangle| \leq \frac{\|\tilde{u}\|}{\lambda}\|\nabla_v \ell(v^*) - \mathbb{E}[\nabla_v \ell(v^*)]\| \leq \frac{C_1\|\tilde{u}\|}{\lambda\sqrt{n_{\min}}} \tag{24}$$

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,

$$\begin{aligned} \|(\boldsymbol{\Sigma}_{v^*} + \lambda \boldsymbol{I})^{-1} v^* \lambda\| = \|(\boldsymbol{\Sigma}_{v^*} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_{v^*}^{\beta} g\lambda\| \leq& \left\|\sum_i \langle \boldsymbol{\psi}_i, g\rangle \boldsymbol{\psi}_i \cdot \frac{\alpha_i^{\beta}\lambda}{\alpha_i + \lambda}\right\| \\ \leq& \lambda^{\beta}\left\|\sum_i \langle \boldsymbol{\psi}_i, g\rangle\right\| \leq \|\boldsymbol{\Sigma}_{v^*}{}^{-\beta} v^*\| \cdot \lambda^{\beta}. \end{aligned} \tag{25}$$

Combine (23), (24) and (25) 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 (15). Moreover, $\tilde{v} \in \mathcal{H}^*$, so it is in the feasible region of (8) thanks to Assumption 2. This further indicates that $\tilde{v}$ is also a solution to (8). As (8) 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 [32], 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 $\mathbb{E}[\nabla_v \ell(v^*)]$:

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

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:

$$\begin{aligned}
A_0 &= \int_{\mathrm{dom}(\tau)} \int_{[0,.5]} \left| \left[F^*_-(\tau) - F^*_+(\tau)\right] \partial_\tau \tilde{F}^*_+(\tau, \alpha) - \left[F^*_+(\tau) - F^*_-(\tau)\right] \partial_\tau \tilde{F}^*_-(\tau, \alpha) \right| \mathrm{d}\alpha \mathrm{d}\tau, \\
&= \int_{\mathrm{dom}(\tau)} \int_{[0,.5]} \left| \left[F^*_-(\tau) - F^*_+(\tau)\right] \left(\partial_\tau F^*_+(\tau) + \partial_\tau F^*_-(\tau)\right) \right| \mathrm{d}\alpha \mathrm{d}\tau, \\
&= \int_{\mathrm{dom}(\tau)} \int_{[0,.5]} \|\boldsymbol{a}(\tau) \times \boldsymbol{b}(\tau)\| \, \mathrm{d}\alpha \mathrm{d}\tau, \qquad (26)
\end{aligned}$$

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 (26) in a different form, we obtain

$$\begin{aligned}
A_0 &= \sqrt{2} \int_{\mathrm{dom}(\tau)} \int_{[0,.5]} \sin(\theta(\tau)) \left|F^*_-(\tau) - F^*_+(\tau)\right| \sqrt{\partial_\tau F^*_+(\tau)^2 + \partial_\tau F^*_-(\tau)^2} \, \mathrm{d}\alpha \mathrm{d}\tau, \\
&= \frac{\sqrt{2}}{2} \int_{\mathrm{dom}(\tau)} \sin(\theta(\tau)) \left|F^*_-(\tau) - F^*_+(\tau)\right| \sqrt{\partial_\tau F^*_+(\tau)^2 + \partial_\tau F^*_-(\tau)^2} \, \mathrm{d}\tau, \\
&= \frac{\sqrt{2}}{2} \int_{\mathrm{dom}(\tau)} \sin(\theta(\tau)) f^*_-(\tau) \left|F^*_-(\tau) - F^*_+(\tau)\right| \sqrt{\left(\frac{f^*_+(\tau)}{f^*_-(\tau)}\right)^2 + 1} \, \mathrm{d}\tau, \qquad (27)
\end{aligned}$$

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 $\sqrt{\left(\frac{p_+(\boldsymbol{x})}{p_-(\boldsymbol{x})}\right)^2 + 1}$ 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 (12) with respect to v and setting the derivative to zero, we can see that superemum is attained at $v^* = \mathrm{atan} \, \frac{p_+}{p_-}$.

□

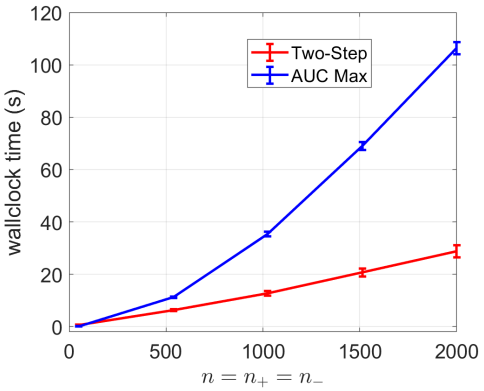

Figure 7: The time comparison over different sample sizes $n = n_+ = n_-$.

I Wall Clock Comparison

In this experiment, we evaluate the computation time of the our two-step algorithm and the naive implementation of the offline AUC maximization (10) by plotting the wall clock time in Figure 7. Both the AUC maximization and two-step procedure are implemented using MATLAB's optimization toolbox. See Section J for details. The two-step procedure's computation time grows at a much slower rate than the offline AUC maximization via a pairwise loss function. Note that as we explained in Section 6.3, if the surrogate loss is decomposable, the objective can be computed with a computational complexity $O(n \log(n))$ [21]. If it the loss is squared loss, the offline algorithm can be performed with a $O(n)$ computational complexity [39].

In this experiment, both methods are written in fully vectorized code. The first order derivatives are provided to the fmincon and fminunc to accelerate the computation. Code can be found in the supplementary material.

J Experiment Setup

In Section 6.2, we reduce the dimension of CIFAR-10 dataset to 50. We first train a residual neural network [17] 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 6 can be found in the supplementary materials. We invite reviewers to reproduce our results.

K 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 (7) becomes

$$(\hat{\boldsymbol{v}}, \hat{v}_0) := \underset{\substack{\boldsymbol{v}\in\mathbb{R}^m,\\ v_0\in\mathbb{R}}}{\operatorname{argmax}} \frac{1}{n_+}\sum_{i=1}^{n_+} \sin[\operatorname{atan}\exp(\langle \boldsymbol{v}, \boldsymbol{h}(\boldsymbol{x}_i)\rangle + v_0)] + \frac{1}{n_-}\sum_{i=1}^{n_-} \cos[\operatorname{atan}\exp(\langle \boldsymbol{v}, \boldsymbol{h}(\boldsymbol{x}_i)\rangle + v_0)]. \tag{28}$$

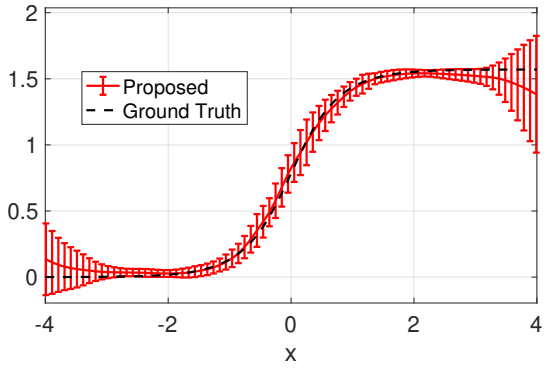

Figure 8: Estimation of the arctangent density ratio function.

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 (8) since the objective in (28) is non-convex with respect to $\boldsymbol{v}$ thus presents extra challenges in the theoretical analysis, although (28) is easier to implement in practice due to its unconstrained nature.

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

We draw 100 samples from $X_+ \sim \mathcal{N}(1,1)$ and $X_- \sim \mathcal{N}(-1,1)$ and solve (8) to estimate the arctangent density ratio. The estimated arctangent density ratio with standard deviation (over 72 runs) are plotted in Figure 8. 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.

M Comparison with Convergence Result in Nguyen et al. [28]

The convergence of (log) density ratio estimation have been developed for two KL divergence based estimators [28]. 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 [28]. Specifically, Nguyen et al. [28] 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.