Maximum Likelihood Estimation of Optimal Receiver Operating Characteristic Curves from Likelihood Ratio Observations

Bruce Hajek and Xiaohan Kang
University of Illinois at Urbana–Champaign
Electrical and Computer Engineering and Coordinated Science Laboratory
Urbana, Illinois
Email: b-hajek@illinois.edu, xiaohan.kang1@gmail.com

Abstract—The optimal receiver operating characteristic (ROC) curve, giving the maximum probability of detection as a function of the probability of false alarm, is a key information-theoretic indicator of the difficulty of a binary hypothesis testing problem (BHT). It is well known that the optimal ROC curve for a given BHT, corresponding to the likelihood ratio test, is theoretically determined by the probability distribution of the observed data under each of the two hypotheses. In some cases, these two distributions may be unknown or computationally intractable, but independent samples of the likelihood ratio can be observed. This raises the problem of estimating the optimal ROC for a BHT from independent samples. The maximum likelihood estimator is observed in simulation experiments to be considerably more accurate than the empirical estimator, especially when the number of samples obtained under one of the two hypotheses is small. The area under the maximum likelihood estimator is derived; it is a consistent estimator of the true area under the optimal ROC curve.

I. INTRODUCTION

Consider a binary hypothesis testing problem (BHT) with observation $X$. The observation $X$ could be high dimensional with continuous and/or discrete components. Suppose $g_0$ and $g_1$ are the probability densities of $X$ with respect to some reference measure, under hypothesis $H_0$ or $H_1$, respectively. Then the likelihood ratio is $R = g_1(x)/g_0(x)$. By the Neyman–Pearson lemma, the optimal decision rule for a specified probability of false alarm, is to declare $H_1$ to be true if either $R \geq \tau$, or if a biased coin comes up heads and $R = \tau$, for a suitable threshold $\tau$ and bias of the coin. The optimal receiver operating characteristic (ROC) curve, giving the maximum probability of detection as a function of the probability of false alarm, is a key information-theoretic indicator of the difficulty of the BHT. Because we focus on the optimal ROC, which is determined by the BHT rather than the specific decision rule, we use the terms “optimal ROC” and “ROC” interchangeably.

This paper addresses the problem of estimating the ROC curve for a BHT from independent samples $R_1, \ldots, R_n$ of the likelihood ratio. Specifically, we assume for some deterministic sequence, $(I_i : i \in [n])$, that $R_i$ is generated from an instance of the BHT such that hypothesis $H_{I_i}$ is true. This problem can arise if the densities $g_0$ and $g_1$ are unknown, but can be factored as $g_k(x) = u(x)h_k(x)$ for $k \in \{0, 1\}$, for some unknown (or very difficult-to-compute) function $u$ and known functions $h_0$ and $h_1$. Then the likelihood ratio can be computed for an observation $X$ using $R = h_1(X)/h_0(X)$ but the distribution of the likelihood ratio depends on the unknown function $u$. So if it is possible, through simulation or repeated physical trials, to generate independent instances of the BHT, it may be possible to generate the independent samples $R_1, \ldots, R_n$ as described.

To elaborate a bit more, we discuss a possible specific scenario related to Cox’s notion of partial likelihood [1]. Suppose $X = (Y_1, S_1, Y_2, S_2, \ldots, Y_T, S_T)$, where the components themselves may be vectors. The full likelihood under hypothesis $H_k$ for $k = 0, 1$ is the product of two factors given below, each of which is a product of $T$ factors:

$$\left(\prod_{i=1}^{T} f_{Y_i|Y_{t-1},S_{t-1}}(y_i|y_{t-1},s_{t-1};k)\right)$$

$$\cdot\left(\prod_{i=1}^{T} f_{S_i|Y_{t},S_{t-1}}(s_i|y',s_{t-1};k)\right),$$

where $y' \triangleq (y_{t'} : t' \in [t])$. Cox defined the first factor to be the partial likelihood based on $Y$ and the second factor to be the partial likelihood based on $S$. If the first factor is very complicated but does not depend on $k$, and the second factor is known and tractable, we arrive at the form of the total likelihood described above: $g_k(x) = u(x)h_k(x)$ for $k \in \{0, 1\}$.

To avoid possible confusion, we emphasize that the problem considered is an inference problem with independent observations, where the ROC is to be estimated. The space of ROCs...
is infinite-dimensional. The observations \( R_1, \ldots, R_n \) are not used for a binary hypothesis testing problem.

There is a large classical literature on ROC curves dating to the early 1940s. Much of the emphasis relating to estimating ROC curves is focused on estimating the area under the ROC curve (AUC). A popular approach is the binormal model such that the distribution of an observed score is assumed to be a monotonic transformation of Gaussian under either hypothesis, and maximum likelihood estimates of the parameters of the Gaussians are found. See [2], [3] and references therein. The first estimator we consider for the ROC curve, which we call the “empirical ROC curve,” is described by that name in [4]. The empirical ROC curve is the same up to a rotation as the “sample ordinal dominance graph” defined in [5], p. 400.

The paper is organized as follows. Some preliminaries about ROC curves are given in Section III. The empirical estimator of the optimal ROC curve based on using the empirical estimators for the two types of error probabilities is considered in Section III. A performance guarantee is derived based on a well-known bound for empirical estimators of CDFs. The ML estimator of the ROC curve is derived in Section IV. Consistency of the ML estimator with respect to the Lévy metric is demonstrated in Section V. The area under the ML estimator of the ROC curve is derived in Section VI and is shown to be a consistent estimator of AUC. Simulations comparing the accuracy of the empirical and ML estimators are given in Section VII and discussion is in Section VIII. Proofs are found in the appendix of the full version of this paper posted to arXiv.

II. PRELIMINARIES ABOUT OPTIMAL ROC CURVES

A. AN EXTENSION OF A CUMULATIVE DISTRIBUTION FUNCTION (CDF)

The CDF \( F \) for a random variable \( R \) is defined by \( F(\tau) = \mathbb{P}\{R \leq \tau\} \) for \( \tau \in \mathbb{R} \). In this paper, \( \mathbb{P} \) always means \( +\infty \). Given a CDF \( F \) with \( F(0^-) = 0 \) and possibly a point mass at \( +\infty \), we define an extended version of \( F \), and abuse notation by using \( F \) to denote both \( F \) and its extension. The extension is defined for \( \tau \in \mathbb{R} \cup \{+\infty\} \) and \( \eta \in [0, 1] \), by \( F(\tau, \eta) = (1 - \eta)F(\tau -) + \eta F(\tau) \), where \( F(-\infty) = \lim_{\tau \to -\infty} F(\tau) \) and \( F(+\infty) = 1 \). Let \( F(\tau) = F(\tau) - F(-\tau) \) denote the mass at \( \tau \). Thus, if \( R \) is an extended random variable with CDF \( F \), then \( F(\tau, \eta) = \mathbb{P}\{R < \tau\} + \eta \mathbb{P}\{R = \tau\} \). Note the extended version of \( F \) is continuous and nondecreasing in \((\tau, \eta)\) in the lexicographically order with \( F(0, 0) = 0 \) and \( F(+\infty, 1) = 1 \), and hence surjective onto \([0, 1]\). Also, let the extended complementary CDF for \( F \) be defined by \( F^c(\tau, \eta) = 1 - F(\tau, \eta) \), so that \( F^c(0, 1) = 1 \) and \( F^c(+\infty, \eta) = 1 \).

B. THE OPTIMAL ROC CURVE FOR A BHT

Consider a BHT and let \( F_1 \) denote the CDF of the likelihood ratio \( R \) under hypothesis \( H_0 \) and let \( F_1 \) denote the CDF of the observation \( R \) under hypothesis \( H_1 \). Then \( dF_1(\tau) = \tau dF_0(\tau) \) for \( \tau \in (0, +\infty) \) (see Appendix A for details), and \( F_1(0) = F_0(0) = 0 \), while it is possible that \( F_0(0) > 0 \) and/or \( F_1(0) > 0 \).

The likelihood ratio test with threshold \( \tau \) and randomization parameter \( \eta \) declares \( H_0 \) to be true if \( R < \tau \), declares \( H_1 \) to be true if \( R > \tau \), and declares \( H_1 \) to be true with probability \( \eta \) if \( R = \tau \). The optimal ROC curve is the graph of the function \( \text{ROC}(\tau) : 0 \leq \tau \leq 1 \) defined by \( \text{ROC}(\tau) = F_0^c(\tau, \eta) \) for \( \tau \) and \( \eta \) are selected such that \( F_0^c(\tau, \eta) = \max \{F_0(0), 1\} \). Note this is well-defined because \( F_0 \) is surjective and for any \( \tau, \tau', \eta, \eta' \) we have \( F_0^c(\tau, \eta) = F_0^c(\tau', \eta') \) if and only if \( F_0(\tau, \eta) = F_0(\tau, \eta') \). Equivalently, the optimal ROC curve is the set of points traced out by \( \text{ROC} = (F_0^c(\tau, \eta), F_1(\tau, \eta)) \) as \( \tau \) and \( \eta \) vary.

Proposition 1: Any one of the functions \( F_0, F_1 \), or ROC determines the other two. ROC is a continuous, concave function over \([0, 1]\).

C. THE LÉVY METRIC ON THE SPACE OF ROC CURVES

Given nondecreasing functions \( A, B \) mapping the interval \([0, 1]\) into itself, the Lévy distance between them, \( L(A, B) \), is the infimum of \( \epsilon > 0 \) such that

\[
A(p - \epsilon) - \epsilon \leq B(p) \leq A(p + \epsilon) + \epsilon \quad \text{for all } p \in \mathbb{R},
\]

with the convention that \( A(p) = B(p) = 0 \) for \( p < 0 \) and \( A(p) = B(p) = 1 \) for \( p > 1 \). A geometric interpretation of \( L(A, B) \) is as follows. It is the smallest value of \( \epsilon \) such that the graph of \( B \) is contained in the region bounded by the following two curves: An upper curve obtained by shifting the graph of \( A \) to the left by \( \epsilon \) and up by \( \epsilon \), and a lower curve obtained by shifting the graph of \( A \) to the right by \( \epsilon \) and down by \( \epsilon \).

Remark 1: It is easy to see the Lévy metric is dominated by the uniform norm \( \|A - B\|_{\infty} = \sup_{p \in [0, 1]} |A(p) - B(p)| \). The Lévy metric is typically more suitable than the uniform norm for functions with jumps. To see this, consider a perfect ROC curve ROC \( \equiv 1 \) and an estimate \( \tilde{\text{ROC}}(\tau) = \min\{1, \epsilon\} \). Then for large \( \epsilon \) the uniform norm of the difference is 1, while the Lévy distance is small.

Lemma 1: Let \( F_{a,0}, F_{a,1}, F_{b,0}, F_{b,1} \) denote CDFs for probability distributions on \([0, \infty]\). Let \( A \) be the function defined on \([0, 1]\) determined by \( F_{a,0}, F_{a,1}, F_{b,0}, F_{b,1} \) as follows. For any \( p \in [0, 1], A(p) = F_{a,1}(\tau, \eta) \), where \((\tau, \eta)\) is the lexicographically smallest point in \([0, \infty] \times [0, 1]\) such that \( F_{a,0}(\tau, \eta) = p \). If \( F_{a,0} \) and \( F_{a,1} \) are the CDFs of the likelihood ratio of a BHT, then \( A \) is the corresponding optimal ROC.) Let \( B \) be defined similarly in terms of \( F_{b,0} \) and \( F_{b,1} \). Then

\[
L(A, B) \leq \sup_{\tau \in [0, \infty]} \max\{|F_{a,0}(\tau) - F_{b,0}(\tau)|, |F_{a,1}(\tau) - F_{b,1}(\tau)|\}.
\]

(1)

We remark that [6] introduces a topology on binary input channels that is related to the Lévy metric used in this paper.

III. THE EMPIRICAL ESTIMATOR OF THE ROC

Consider a BHT and let \( F \) denote the CDF of the likelihood ratio \( R \) under hypothesis \( H_k \) for \( k = 0, 1 \). Suppose for some positive integers \( n_0 \) and \( n_1 \), independent random variables \( R_{0,1}, \ldots, R_{0,n_0}, R_{1,1}, \ldots, R_{1,n_1} \) are observed such that \( R_{k,i} \) has CDF \( F_k \) for \( k = 0, 1 \) and \( 1 \leq i \leq n_k \). A straightforward
approach to estimate ROC is to estimate $F_k$ using only the $n_k$ observations having CDF $F_k$ for $k = 0, 1$. In other words, let

$$
\hat{F}_k(\tau) = \frac{1}{n_k} \sum_{i=1}^{n_k} I(R_{ki} \leq \tau)
$$

for $k = 0, 1$ and let $\hat{\text{ROC}}_E$, the empirical estimator of ROC, have the graph swept out by the point $(\hat{F}_0(\tau), \hat{F}_1(\tau))$ as $\tau$ varies over $[0, \infty]$ and $\eta$ varies over $[0, 1]$. In general, $\text{ROC}_E$ is a step function with all jump locations at multiples of $\frac{1}{n}$ and the jump sizes being multiples of $\frac{1}{n_1}$. Moreover, $\text{ROC}_E$ depends on the numerical values of the observations only through the ranks (i.e., the order, with ties accounted for) of the observations.

The estimator $\text{ROC}_E$ as we have defined it is typically not concave, and hence is typically not the optimal ROC curve for a BHT. This suggests the concavified empirical estimator $\hat{\text{ROC}}_{CE}$ defined to be the least concave majorant of $\text{ROC}_E$. Equivalently, the region under the graph of $\text{ROC}_{CE}$ is the convex hull of the region under $\text{ROC}_E$.

The following provides some performance guarantees for the empirical and concavified empirical estimators.

**Proposition 2:** Let $n = n_0 + n_1$ and $\alpha = \frac{n_0}{n_1 + n_0}$. Then the empirical estimator satisfies

$$
P\{L(\text{ROC}, \hat{\text{ROC}}_E) \geq \delta\} \leq 2e^{-2n_0\delta^2} + 2e^{-2n(1-\alpha)\delta^2}.
$$

Moreover, if $\alpha \in (0, 1)$ is fixed and $n_k \to \infty$ for $k = 0, 1$ with $\frac{n_0}{n} = \frac{1-\alpha}{\alpha}$, then $L(\text{ROC}, \hat{\text{ROC}}_E) \to 0$ a.s. as $n \to \infty$.

In other words, $\text{ROC}_E$ is consistent in the $L_1$ metric. In general, $L(\text{ROC}, \hat{\text{ROC}}_{CE}) \leq L(\text{ROC}, \hat{\text{ROC}}_E)$, so the above statements are also true with $\text{ROC}_E$ replaced by $\hat{\text{ROC}}_{CE}$.

**Proof:** The Dvoretzky–Kiefer–Wolfowitz (DKW) inequality with the optimal constant proved by Massart implies:

$$
P\left\{ \sup_{\tau \in [0, \infty]} |F_k(\tau) - \hat{F}_k(\tau)| \geq \delta \right\} \leq 2e^{-2n_k\delta^2}.
$$

Combining with Lemma 1 implies 2. The consistency of $\hat{\text{ROC}}_E$ follows from the Borel–Cantelli lemma and the fact the sum of the right-hand side of (2) over $n$ is finite for any $\delta > 0$.

The final inequality follows from the following observations: $\hat{\text{ROC}}_{CE}(p) \geq \text{ROC}_E(p)$ for $p \in [0, 1]$, and if $\text{ROC}_E$ is less than or equal to the concave function $p \to \text{ROC}(p+\epsilon) + \epsilon$, then so is $\hat{\text{ROC}}_{CE}$, by the definition of least concave majorant.

**Remark 2:** A strong consistency result for the empirical estimator in terms of the uniform norm with some restrictions on the distributions $F_0$ and $F_1$ has been developed in [3].

While the bound (2) seems reasonably tight for $\alpha$ near 1/2, the bound is degenerate if $\alpha$ is very close to zero or one. The maximum likelihood estimator derived in the next section is consistent even if all the observations are generated under a single hypothesis.

**IV. The ML Estimator of the ROC**

Consider a BHT and let $F_k$ denote the CDF of the likelihood ratio $R$ under hypothesis $H_k$ for $k = 0, 1$, and suppose for some $n \geq 1$ and deterministic binary sequence $I_i : i \in \{n\}$, independent random variables $R_1, \ldots, R_n$ are observed such that for each $i \in \{n\}$, the distribution of $R_i$ is $F_i$. The likelihood of the set of observations is the probability the observations take their particular values, and that is determined by $F_0$ and $F_1$, and hence, by Proposition 1 also by ROC or by $F_0$ alone or by $F_1$ alone. Hence, it makes sense to ask what is the maximum likelihood (ML) estimator of ROC, or equivalently, what is the ML estimator of the triplet $(F_0, F_1, \text{ROC})$, given $I_i : i \in \{n\}$ and $R_i, i \in \{n\}$. The answer is given by the proposition in this section.

Let $\varphi_n$ be defined by

$$
\varphi_n(\lambda) = \begin{cases} 
\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda+(1-\lambda)R_i} & \text{if } 0 \leq \lambda < 1, \\
1 & \text{if } \lambda = 1.
\end{cases}
$$

Note that $\varphi_n$ is finite over $(0, 1)$, continuous over $[0, 1)$, and convex over $[0, 1]$. Moreover, $\varphi_n(0) = \infty$ if and only if $R_i = 0$ for some $i$, and $\varphi_n$ has a jump discontinuity at 1 if and only if $R_i = \infty$ for some $i$.

**Proposition 3:** The ML estimator $(\hat{F}_0, \hat{F}_1, \hat{\text{ROC}}_{ML})$ is unique and is determined as follows. $\hat{\text{ROC}}_{ML}$ is the optimal ROC curve corresponding to $\hat{F}_0$ and/or $\hat{F}_1$, where:

1. If $\frac{1}{n} \sum_{i=1}^{n} R_i \leq 1$, then for $\tau \in [0, \infty)$

   $$
   \hat{F}_0(\tau) = \frac{1}{n} \sum_{i=1}^{n} I(R_i \leq \tau); \quad \hat{F}_1(\tau) = \frac{1}{n} \sum_{i=1}^{n} I(R_i \leq \tau) R_i.
   $$

2. If $\frac{1}{n} \sum_{i=1}^{n} R_i > 1$, then for $\tau \in [0, \infty)$

   $$
   \hat{F}_0(\tau) = \frac{1}{n} \sum_{i=1}^{n} I(R_i > \tau); \quad \hat{F}_1(\tau) = \frac{1}{n} \sum_{i=1}^{n} I(R_i \leq \tau).
   $$

3. If neither of the previous two cases holds, then for $\tau \in (0, \infty)$

   $$
   \hat{F}_0(\tau) = \frac{1}{n} \sum_{i=1}^{n} I(R_i \leq \tau) \frac{1}{\lambda_n + (1-\lambda_n)R_i}
   $$

and

   $$
   \hat{F}_1(\tau) = \frac{1}{n} \sum_{i=1}^{n} I(R_i \leq \tau) \frac{R_i}{\lambda_n + (1-\lambda_n)R_i},
   $$

where $\lambda_n$ is the unique value in $(0, 1)$ so that $\varphi_n(\lambda_n) = 1$.

**Remark 3:**

1. The estimator does not depend on the indicator variables $I_i : i \in \{n\}$. That is, the estimator does not take into account which observations are generated using which hypothesis.

2. Cases 1) and 2) can both hold only if $R_i = 1$ for all $i$, because $r + \frac{1}{r} \geq 2$ for $r \in [0, \infty]$ with equality if and only if $r = 1$.

3. If case 1) holds with strict inequality, then $\hat{F}_1(\{\infty\}) > 0$, even though $R_i < \infty$ for all $i$. 


4) Similarly, if case 2) holds with strict inequality, then \( F_0(0) > 0 \) even though \( R_i > 0 \) for all \( i \).

5) Suppose case 3) holds. The existence and uniqueness of \( \lambda_n \) can be seen as follows. Since case 2) does not hold, \( \varphi_n(0) > 1 \). If \( R_i = \infty \) for some \( i \) then \( \varphi_n(1) < 1 \); and if \( R_i < \infty \) for all \( i \), then \( \varphi_n'(1) = \frac{1}{n} \sum_{i=1}^{n} (R_i - 1) > 0 \), where we have used the fact case 1) does not hold. Thus, \( \varphi_n(\lambda) < 1 \) if \( \lambda < 1 \) and \( \lambda \) is sufficiently close to 1. Therefore the existence and uniqueness of \( \lambda_n \) in case 3) follow from the properties of \( \varphi_n' \).

6) The proof of Proposition 5 is in Appendix C. Maximizing the likelihood is reduced to a convex optimization problem and the KKT optimality conditions are used.

The following corollary presents an alternative version of Proposition 5 that consolidates the three cases of Proposition 3. It is used in the proof of consistency of the ML estimator.

**Corollary 1:** The ML estimator is unique and is determined as follows. For \( \tau \in [0, \infty) \),
\[
\hat{F}_0^c(\tau) = \frac{1}{n} \sum_{i=1}^{n} I_{\{R_i > \tau\}} \frac{1}{\lambda_n + (1 - \lambda_n)R_i}
\]
and
\[
\hat{F}_1(\tau) = \frac{1}{n} \sum_{i=1}^{n} I_{\{R_i \leq \tau\}} \frac{R_i}{\lambda_n + (1 - \lambda_n)R_i},
\]
where \( \lambda_n = \min \{ \lambda \in [0, 1] : \varphi_n(\lambda) \leq 1 \} \).

**Remark 4:** It is shown in the proof of Proposition 5 that \( \lambda_n \to \alpha \) a.s. as \( n \to \infty \) if \( F_0 \) is not identical to \( F_1 \). Thus, for \( n \) large, \( \lambda_n \) is approximately the prior probability that a given observation is generated under hypothesis \( H_0 \) and \( n\lambda_n \) is approximately the number of observations generated under \( H_0 \). The ML estimator \( \hat{F}_0 \) can be written as
\[
\hat{F}_0^c(\tau) = \frac{1}{n\lambda_n} \sum_{i=1}^{n} I_{\{R_i > \tau\}} \frac{\lambda_n}{\lambda_n + (1 - \lambda_n)R_i},
\]
where \( \frac{\lambda_n}{\lambda_n + (1 - \lambda_n)R_i} \) can be interpreted as an estimate of the posterior probability that observation \( R_i \) was generated under \( H_0 \).

**V. CONSISTENCY OF THE ML ESTIMATE OF THE ROC**

Suppose \( R \) has CDF \( F_0 \) under \( H_0 \) and CDF \( F_1 \) under \( H_1 \), and \( \lambda \) is also the likelihood ratio. Let \( \alpha \) be fixed with \( \alpha \in [0, 1] \) and suppose the observations \( R_1, R_2, \ldots \) are independent, identically distributed random variables with the mixture distribution \( \alpha F_0 + (1 - \alpha) F_1 \). We are considering the problem of estimating the ROC curve for the BHT for distributions \( F_0 \) and \( F_1 \) using the ML estimator \( \hat{F}_0(\tau), \hat{F}_1(\tau) \) based on the observations \( R_1, \ldots, R_n \) as \( n \to \infty \). For brevity we suppress \( n \) in the notation for \( \hat{F}_0(\tau), \hat{F}_1(\tau) \), and \( \hat{ROC} \) and we write \( "X \to c \) a.s. as \( n \to \infty \)" where a.s. is the abbreviation for "almost surely," to mean \( \Pr[\lim_{n \to \infty} X = c] = 1 \).

**Proposition 4 (Consistency of the ML estimator of the ROC curve):** The ML estimator of the ROC curve for \( H_0 \) vs. \( H_1 \) is consistent. That is, \( L(\hat{ROC}, ROC) \to 0 \) a.s. as \( n \to \infty \).

The proof of Proposition 4 is given in Appendix C. The first part of the proof is to establish that if \( F_0 \) is not identical to \( F_1 \), then \( \lambda_n \to \alpha \) a.s. as \( n \to \infty \). Thus, the estimators \( \hat{F}_0 \) and \( \hat{F}_1 \) are close to functions obtained by replacing \( \lambda_n \) by \( \alpha \), and those resulting functions converge to \( F_0 \) and \( F_1 \), respectively, by the law of large numbers.

**VI. AREA UNDER THE ML ROC CURVE**

The area under ROC ML, which we denote by \( \hat{AUC}_{ML} \), is a natural candidate for an estimator of AUC, the area under ROC for the BHT. An expression for it is given in the following proposition. Let \( \lambda_n \) be defined as in Corollary 1 and for \( i, i' \in [n] \), let
\[
T_{i,i'} = \frac{\max \{ R_i, R_{i'} \}}{2(\lambda_n + (1 - \lambda_n)R_i)(\lambda_n + (1 - \lambda_n)R_{i'})},
\]
with the following understanding. Recall that if \( R_i = 0 \) for some \( i \in [n] \) then \( \lambda_n > 0 \), so the denominator in \( T_{i,i'} \) is always strictly positive. Also recall that if \( R_i = \infty \) for some \( i \in [n] \) then \( \lambda_n < 1 \), and the following is based on continuity: If \( R_i = R_{i'} = \infty \) set \( T_{i,i'} = 0 \). If \( R_i < R_{i'} = \infty \), set \( T_{i,i'} = \frac{1}{2(\lambda_n + (1 - \lambda_n)R_i)(1 - \lambda_n)} \).

**Proposition 5:**

1) The area under \( \hat{ROC}_{ML} \) is given by
\[
\hat{AUC}_{ML} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{i'=1}^{n} T_{i,i'}. \quad (5)
\]

2) The estimator \( \hat{AUC}_{ML} \) is consistent: \( \hat{AUC}_{ML} \to AUC \) a.s. as \( n \to \infty \).

3) Let \( R, R' \) be independent random variables and use \( \mathbb{E}_0 \) to denote expectation when they both have CDF \( F_0 \). Then
\[
AUC = \frac{1}{2} \mathbb{E}_0[\max \{ R, R' \}] + F_1(\{ \infty \}) \quad (6)
\]
\[
= 1 - \frac{1}{2} \mathbb{E}_0[\min \{ R, R' \}]. \quad (7)
\]

4) For \( i \neq i' \), \( E[T_{i,i'}^{(\alpha)}] = AUC \), where \( T_{i,i'}^{(\alpha)} \) is the same as \( T_{i,i'} \) with \( \lambda_n \) replaced by \( \alpha \).

**Remark 5:**

1) The expression \( (5) \) can be verified by checking that it reduces to \( (6) \) in case \( \mathbb{E}_0 \) also is \( \hat{AUC}_{ML} \) replaced by expectation using \( \hat{F}_0 \) and \( \hat{F}_1 \) is replaced by \( \hat{F}_1 \). A more direct proof of \( (5) \) is given.

2) The true AUC for the BHT is invariant under swapping the two hypotheses. Similarly, \( \hat{AUC}_{ML} \) is invariant under replacing \( \lambda_n \) by \( 1 - \lambda_n \) and \( R_i \) by \( \frac{1}{R_i} \) for all \( i \). If \( R_i = 1 \) for all \( i \), \( \hat{AUC}_{ML} = 1/2 \).

3) Part 4) of the proposition is to be expected due to the consistency of \( \hat{AUC}_{ML} \) and the law of large numbers, because if \( n \) is large, most of the \( n^2 \) terms in \( (5) \) are indexed by \( i, i' \) with \( i \neq i' \), and we know, if \( F_0 \) is not identical to \( F_1 \), that \( \lambda_n \to \alpha \) a.s. as \( n \to \infty \).
 VII. SIMULATIONS

In this section we test the estimators in a simple binormal setting. Let $X$ be distributed by $\mathcal{N}(0, 1)$ under $H_0$ and by $\mathcal{N}(1, 1)$ under $H_1$. Then the likelihood ratio is $R = \exp(X - \frac{1}{2})$ and the ROC curve is given by $\text{ROC}(p) = 1 - \Phi(\Phi^{-1}(1 - p) - 1)$, where $\Phi$ is the CDF of the standard Gaussian distribution. Simulation results for the three ROC estimators are shown in Fig. 1 with various numbers of observations under the two hypotheses $(n_0, n_1)$. For each pair of $(n_0, n_1)$ two figures are shown. The left figure shows the estimated ROC curves together with the true ROC curve for a single sample instance of $n_0 + n_1$ likelihood observations. The right figure shows the average Lévy distances of the estimators over $N = 500$ such sample instances together with error bars (i.e., plus or minus sample standard deviations divided by $\sqrt{N}$). The simulation code can be found at [7].

The two empirical estimators have similar performance, while CE outperforms E slightly in terms of the Lévy distance. Note $\text{ROC}_{CE}$, as the least concave majorant of $\text{ROC}_E$, could be biased toward higher probability of detection as evidenced by the sample instances.

It can be seen that the ML estimator (MLE) achieves much smaller Lévy distance than E or CE. The difference is more pronounced when the number of observations under one hypothesis is significantly smaller than that under the other, as seen in Figs. 1d, 1f. This is because E and CE calculate the empirical distributions based on the likelihood ratio observations under the two hypotheses separately before combining the empirical distributions into an estimated ROC curve. As a result, having very few samples under either hypothesis results in errors in estimating the ROC curve regardless of how accurate the estimated distribution under the other hypothesis is. In contrast, every observation contributes to the joint estimation of the pair of distributions in MLE, so the ROC curve can be accurately estimated even when there are very few samples from one hypothesis. In fact, as Section VIII suggested, MLE works even if all samples are generated from the same hypothesis (see Fig. 1g), while E and CE do not work because one of the distributions cannot be estimated at all.

VIII. DISCUSSION

The qualitative differences between the empirical estimator $\text{ROC}_E$ and the ML estimator $\text{ROC}_{ML}$ are striking. Only the rank ordering of the samples is used by the empirical estimator—not the numerical values. So it is important to track which samples are generated with which distribution. The ML estimator does not depend on which samples were generated with which distribution and exact numerical values are used.

We proved a consistency result for $\text{ROC}_{ML}$, but perhaps it also satisfies a bound similar to $\text{CE}$. It may be interesting to explore the accuracy of the ML estimator for large, fixed $n$ as a function of the fraction, $\alpha$, of observations that are taken under hypothesis $H_0$.

A BHT is the same as a binary input channel (BIC). Work of Blackwell and others working on the comparison

Fig. 1: Sample instances and average Lévy distances.
of experiments has led to canonical channel descriptions that are equivalent to the ROC curve, such as the Blackwell measure. The Blackwell measure is the distribution of the posterior probability that hypothesis $H_0$ is true for equal prior probabilities 1/2 for the hypotheses. See [8] and references therein. It may be of interest to explore estimation of various canonical channel descriptions besides the ROC under various metrics.

REFERENCES

[1] D. R. Cox, “Partial likelihood,” Biometrika, vol. 62, no. 2, pp. 269–276, 1975. [Online]. Available: http://www.jstor.org/stable/2335362

[2] C. Metz and X. Pan, “Proper binormal ROC curves: Theory and maximum-likelihood estimation,” Biometrics, vol. 43, pp. 1–33, 1999.

[3] F. Hsieh and B. W. Turnbull, “Nonparametric and semiparametric estimation of the receiver operating characteristic curve,” The Annals of Statistics, vol. 24, no. 1, Feb. 1996.

[4] E. DeLong, D. DeLong, and D. Clark-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach,” Biometrics, vol. 44, pp. 837–845, 1988.

[5] D. Barber, “The area above the ordinal dominance graph and the area below the receiver operating characteristic graph,” J. Math. Psychology, vol. 12, pp. 387–415, 1975.

[6] R. Nasser, “Topological structures on DMC spaces,” Entropy, vol. 20, no. 5, 2018. [Online]. Available: https://www.mdpi.com/1099-4300/20/5/143

[7] X. Kang, “ML estimator of optimal roc curve simulations,” Feb. 2022. [Online]. Available: https://github.com/Veggente/mleroc

[8] N. Goela and M. Raginsky, “Channel polarization through the lens of Blackwell measures,” IEEE Transactions on Information Theory, vol. 66, no. 10, pp. 6222–6241, 2020.

[9] A. Ben-Tal and A. Nemirovski, “Optimization III: Convex analysis, nonlinear programming theory, nonlinear programming algorithms,” 2013. [Online]. Available: https://www2.isye.gatech.edu/~nemirovski/OPTIII

[10] K. B. Athreya and S. N. Lahiri, Measure Theory and Probability Theory. New York, NY: Springer, 2006.

APPENDIX A

RELATION OF $F_0$ AND $F_1$

Let $P_k$ and $g_k$ denote the probability distribution and the probability density function with respect to some reference measure $\mu$ of the observation $X$ in a measurable space $(\mathcal{X}, \Sigma)$ under hypothesis $H_k$ for $k = 0, 1$. In other words, $P_k(A) = \int_A g_k(x) \mu(dx)$ for any $A \in \Sigma$. Let $\rho: \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ be defined by

$$
\rho(x) = \begin{cases} 
\frac{g_1(x)}{g_0(x)} & \text{if } g_0(x) > 0, \\
\infty & \text{if } g_0(x) = 0.
\end{cases}
$$

Then $\rho$ is a Borel measurable function denoting the likelihood ratio given an observation. The probability distribution of the extended random variable $R = \rho(X)$ under $H_k$ is the push-forward of the measure $P_k$ induced by the function $\rho$ for $k = 0, 1$, denoted by $\nu_k$. The probability distribution $\nu_k$ restricted to $\mathbb{R}$ is also the unique Borel measure (known as the Lebesgue–Stieljes (L–S) measure corresponding to $F_k$, the CDF of $R$) on $[0, \infty)$ such that $\nu_k([0, \tau]) = F_k(\tau)$ for all $\tau \in [0, \infty)$.

Throughout this paper, integrals of the form $\int h(r) \, dF(r)$ are understood to be Lebesgue–Stieljes integrals (for the extended real numbers). That is,

$$
\int_{\mathbb{R}} h(r) \, dF(r) \triangleq \int_{\mathbb{R}} h(r) \nu_F(dr),
$$

for any Borel measurable function $h$.

Proposition 6: For any Borel set $A$ in $\mathbb{R}$,

$$
\nu_1(A) = \int_A r\nu_0(dr).
$$

In other words, when restricted to the Borel sets in $\mathbb{R}$, $\nu_1$ is absolutely continuous with respect to $\nu_0$, and the Radon–Nikodym derivative is the identity function almost everywhere with respect to $\nu_0$.

Proof: By the change-of-variables formula for push-forward measures, for any Borel set $A$ in $\mathbb{R}$,

$$
\nu_1(A) = \int_{\mathbb{R}} I_A(r)\nu_1(dr)
$$

$$
= \int_{\mathbb{R}} I_A(\rho(x))P_1(dx)
$$

$$
= \int_{\mathbb{R}} I_A(\rho(x))g_1(x)\mu(dx)
$$

$$
= \int_{\mathbb{R}} I_A(\rho(x))\rho(x)g_0(x)\mu(dx)
$$

$$
= \int_{\mathbb{R}} I_A(\rho(x))r\nu_0(dr)
$$

$$
= \int_A \nu_0(dr),
$$

implying the proposition.

APPENDIX B

PROOFS FOR SECTION III

Proof of Proposition 7: The function $F_0$ determines $F_1$ by $F_1(\tau) = \int_{[0,\tau]} r \, dF_0(r)$ for $\tau \in [0, \infty)$. Conversely, $F_1$ determines $F_0$ by $F_0^\tau(\tau) = \int_{[\tau,\infty)} \frac{1}{r} \, dF_1(r)$ for $\tau \in [0, \infty)$. So either one of $F_0$ or $F_1$ determines the other, and hence also determines ROC as described above. To complete the proof it suffices to show that ROC determines $F_0$. The function ROC is concave so it has a right-hand derivative which we denote by ROC'. Then $F_0^\tau(\tau) = \sup\{p : \text{ROC}'(p) > \tau\}$ for $\tau \in [0, \infty)$.

Proof of Lemma 7: Let the right-hand side of (I) be denoted by $\epsilon$. Note that

$$
\epsilon = \sup_{\tau \in (0, \infty)} \max_{\eta \in [0,1]} |F_{a,0}^\tau(\tau, \eta) - F_{b,0}^\tau(\tau, \eta)|,
$$

$$
|F_{a,1}^\tau(\tau, \eta) - F_{b,1}^\tau(\tau, \eta)|,
$$

because for $\tau$ fixed, the right-hand side of (8) is the maximum of a convex function of $\eta$ and the value at $\eta = 0$ and $\eta = 1$ is obtained by the right-hand side of (I) at $\tau$ and $\tau$, respectively. We appeal to the geometric interpretation of $L(A, B)$. Consider any point $(p, B(p))$ on the graph of $B$. It is equal to $(F_{b,0}^\tau(\tau, \eta), F_{b,1}^\tau(\tau, \eta))$ for some choice of $(\tau, \eta)$. Let $(p', A(p'))$ denote the point on the graph of $A$ for the same choice of $(\tau, \eta)$. In other words, it is the point $(F_{a,0}^\tau(\tau, \eta), F_{a,1}^\tau(\tau, \eta))$. Then $(p, B(p))$ can be reached from $(p', A(p'))$ by moving horizontally at most $\epsilon$ and moving.
vertically at most $\epsilon$. So $(p, B(p))$ is contained in the region bounded between the upper and lower shifts of the graph of $A$ as claimed.

**APPENDIX C**

**DERIVATION OF $\text{ROC}_{\text{ML}}$**

Proposition [3] and its corollary are proved in this section.

**Proof of Proposition [3]**

Given the binary sequence $(I_i; i \in \mathbb{N})$ and the likelihood ratio samples $R_1, \ldots, R_m$, let $0 = v_0 < v_1 < v_2 < \cdots < v_m < v_{m+1} = \infty$ be the set of unique values of the samples, augmented by $v_0 = 0$ and $v_{m+1} = \infty$ even if $0$ and/or $\infty$ is not among the observed samples. Let $(c^0_0, c^0_1, c^0_2, \ldots, c^0_m)$ denote the multiplicities of the values from among $(R_i; I_i = 0)$ and let $(c^1_1, c^1_2, \ldots, c^1_m, c^1_{m+1})$ denote the multiplicities of the values from among $(R_i; I_i = 1)$.

Let $a_j = F_0(\{v_j\})$ for $0 \leq j \leq m$ and let $b = F_1(\{\infty\})$. Thus $a_j$ is the probability mass at $v_j$ under hypothesis $H_0$ for $0 \leq j \leq m$. The corresponding probability mass at $v_j$ under hypothesis $H_1$ is $a_jv_j$ for $0 \leq j \leq m$ and the probability mass $v_j$ under hypothesis $H_1$ is $b$.

The log-likelihood to be maximized is given by

$$\sum_{j=0}^{m} c_j \log a_j + \sum_{j=1}^{m} c_j \log (a_j v_j) + c_{m+1} \log b,$$

where $0 \leq 0 \leq m$ and $0 \leq 0 \leq m$ is understood as $0$ and $0 \leq 0 \leq m$ is understood as negative infinity. Equivalently, dropping the term $\sum_{j=1}^{m} c_j \log (v_j)$ which does not depend on $F_0$ (or $F_1$ or ROC), the ML estimator is to maximize

$$\sum_{j=0}^{m} c_j \log a_j + c_{m+1} \log b,$$

where $c_0 \triangleq c^0_0, c_{m+1} \triangleq c^1_{m+1}$ and $c_j \triangleq c^0_j + c^1_j$ for $1 \leq j \leq m$. In other words, $c_j$ is the total multiplicity of $v_j$ in all samples regardless of the hypothesis.

For any choice of $F_0$ (or $F_1$ or ROC), the probabilities satisfy the constraint:

$$\sum_{j=0}^{m} a_j \leq 1 \text{ and } \sum_{j=1}^{m} a_j v_j + b \leq 1. \tag{9}$$

The inequalities in (9) both hold with equality if the distribution $F_0$ (or equivalently $F_1$) assigns probability one to the set $\{v_0, \ldots, v_{m+1}\}$. Otherwise, both inequalities are strict. We claim and now prove that any ML estimator is such that both inequalities in (9) hold with equality. It is true in the degenerate special case that $R_i \in \{0, \infty\}$ for all $i$, in which case an ML estimator is given by $\text{ROC}(p) \equiv 1$, $F_0(0) = 1$ and $F_1(\{\infty\}) = 1$. So we can assume $m \geq 1$ and there is a value $j_0$ (for example, $j_0 = 1$) such that $1 \leq j_0 \leq m$. If $F_0$ does not assign probability one to $\{v_0, \ldots, v_{m+1}\}$ then the same is true for $F_1$, so that strict inequality must hold in both constraints in (9). Then the probability mass from $F_0$ (and $F_1$) that is not on the set $\{v_0, \ldots, v_{m+1}\}$ can be removed and mass can be added to $F_0$ at $0$ and $v_{j_0}$ and to $F_1$ at $v_{j_0}$ and $\infty$ such that both constraints in (9) hold with equality and the likelihood is strictly increased. This completes the proof of the claim.

Therefore, any ML estimator is such that the distributions are supported on the set $\{v_0, \ldots, v_{m+1}\}$ and the probabilities assigned to the points give an ML estimator if and only if they are solutions to the following convex optimization problem:

$$\max_{a \geq 0, b \geq 0} \sum_{j=0}^{m} c_j \log a_j + c_{m+1} \log b \tag{10}$$

s.t. $\sum_{j=0}^{m} a_j = 1$ and $\sum_{j=1}^{m} a_j v_j + b = 1$.

The relaxed Slater constraint qualification condition is satisfied for (10), so there exists a solution and dual variables satisfying the KKT conditions (see Theorem 3.2.4 in [9]). The Lagrangian is

$$L(a, b, \lambda, \mu) = \sum_{j=0}^{m} c_j \log a_j + c_{m+1} \log b$$

$$- \left( \sum_{j=0}^{m} a_j - 1 \right) \lambda - \left( \sum_{j=1}^{m} a_j v_j + b - 1 \right) \mu.$$

The KKT conditions on $(a, b, \lambda, \mu)$ are

$$a \geq 0, b \geq 0; \sum_{j=0}^{m} a_j = 1; \sum_{j=1}^{m} a_j v_j + b = 1; \frac{\partial L}{\partial a_0} \leq 0; a_0 \cdot \frac{\partial L}{\partial a_0} = 0; \frac{\partial L}{\partial a_j} = 0 \text{ for } j \in [m]; \frac{\partial L}{\partial b} \leq 0; b \cdot \frac{\partial L}{\partial b} = 0,$$

where

$$\frac{\partial L}{\partial a_0}(a, b, \lambda, \mu) = \begin{cases} \frac{c_0}{a_0} - \lambda & \text{if } c_0 > 0, \\ -\lambda & \text{if } c_0 = 0; \end{cases}$$

$$\frac{\partial L}{\partial a_j}(a, b, \lambda, \mu) = \begin{cases} \frac{c_j}{a_j} - \lambda - v_j \mu & \text{for } j \in [m]; \\ -\mu & \text{if } c_{m+1} > 0, \end{cases}$$

$$\frac{\partial L}{\partial b}(a, b, \lambda, \mu) = \begin{cases} \frac{c_{m+1}}{b} - \mu & \text{if } c_{m+1} > 0, \\ -\mu & \text{if } c_{m+1} = 0. \end{cases}$$

Solving the KKT conditions yields:

1) If $c_{m+1} = 0$ and $\sum_{j=1}^{m} v_j c_j \leq \sum_{j=0}^{m} c_j$, then

$$\hat{a}_j = \frac{c_j}{\sum_{k=0}^{m} c_k} \quad \text{for } 0 \leq j \leq m;$$

$$\hat{b} = 1 - \sum_{j=1}^{m} \frac{v_j c_j}{\sum_{j=0}^{m} c_j}; \quad \hat{\lambda} = \sum_{j=0}^{m} c_j; \quad \hat{\mu} = 0.$$

2) Otherwise, if $c_0 = 0$ and $\sum_{j=1}^{m} c_j / v_j \leq \sum_{j=1}^{m+1} c_j$, then

$$\hat{a}_j = \frac{c_j}{v_j \mu} \quad \text{for } 1 \leq j \leq m;$$

$$\hat{a}_0 = 1 - \frac{\sum_{j=1}^{m} c_j / v_j}{\sum_{j=1}^{m+1} c_j};$$

$$\hat{b} = \frac{1}{\sum_{j=0}^{m} c_j} - \frac{\sum_{j=0}^{m} c_j / v_j}{\sum_{j=1}^{m+1} c_j}.$$
\[ \tilde{b} = 0; \quad \tilde{\lambda} = 0; \quad \tilde{\mu} = \sum_{k=1}^{m+1} c_k. \]

3) Otherwise, \( \tilde{\lambda} > 0, \tilde{\mu} > 0 \) are determined by solving
\begin{align}
\sum_{j=0}^{m} \frac{c_j}{\lambda + v_j \mu} &= 1, \\
\sum_{j=1}^{m} \frac{c_j}{\lambda / v_j + \mu} + \frac{c_{m+1}}{\mu} &= 1, \\
\text{and for } 0 \leq j \leq m, \quad \tilde{\alpha}_j &= \frac{c_j}{\lambda + v_j \mu}, \quad \tilde{b} = \frac{c_{m+1}}{\mu}.
\end{align}

Multiplying both sides of (11) by \( \lambda \) and both sides of (12) by \( \mu \) and adding the respective sides of the two equations obtained, yields \( \lambda + \mu = \sum_{j=0}^{m+1} c_j = n \). The above conditions can be expressed in terms of the variables \( R_i \), and then replacing \( \mu \) by \( n - n\lambda \), and \( \lambda \) by \( n\lambda \), yields the proposition.

**Proof of Corollary** Corollary [1] is deduced from Proposition [3] as follows. If \( R_i = 1 \) for \( 1 \leq i \leq n \) then the corollary gives that both \( \hat{F} \) and \( \tilde{F} \) have all their mass at \( r = 1 \), in agreement with Proposition [3]. So for the remainder of the proof suppose \( R_i \neq 1 \) for some \( i \).

Consider the three cases of Proposition [3]. If case 1) holds then \( \varphi_n(1) = 0 \) and \( \varphi_n'(1) = \frac{1}{n} \sum_{i=1}^{n} (R_i - 1) \leq 0 \). Also, \( R_i < \infty \) for \( 1 \leq i \leq n \). Since \( R_i \notin \{1, \infty\} \) for at least one value of \( i \), \( \varphi_n(\lambda) \) is strictly convex over \([0, 1]\). Therefore, \( \varphi_n(\lambda) > 1 \) for \( \lambda \in (0, 1) \). Thus, \( \lambda_n \) defined in the corollary is given by \( \lambda_n = 1 \), and the corollary agrees with Proposition [3].

If case 2) holds then \( \varphi_n(0) \leq 1 \). Thus, \( \lambda_n \) defined in the corollary is given by \( \lambda_n = 0 \), and the corollary agrees with Proposition [3].

If neither case 1) nor case 2) holds, then \( \lambda_n \) in the corollary is the same as \( \lambda_n \) in Proposition [3] and the corollary again agrees with Proposition [3].

**APPENDIX D**

**From Pointwise to Uniform Convergence of CDFs**

The following basic lemma shows that uniform convergence of a sequence \( (F_n: n \geq 1) \) of CDFs to a fixed limit is equivalent to pointwise convergence of both the sequence and the corresponding sequence of left limit functions, at each of a suitable countably infinite set of points. The CDFs in this section may correspond to probability distributions with positive mass at \( -\infty \) and/or \( \infty \).

**Lemma 2 (Finite net lemma for CDFs):** Given a CDF \( F \) and any integer \( L \geq 1 \), there exist \( c_1, \ldots, c_{L-1} \in \mathbb{R} \cup \{-\infty, \infty\} \) such that for any CDF \( G, \sup_{r \in \mathbb{R}} |F(c_r) - G(c_r)| \leq \delta + \frac{1}{L} \) where
\[ \delta = \max_{1 \leq \ell \leq L-1} \max \{|F(c_\ell) - G(c_\ell)|, |F(c_{\ell+1}) - G(c_{\ell+1})|\}. \]

**Proof:** Let \( c_\ell = \min \{c \in \mathbb{R} \cup \{-\infty, \infty\}: F(c) \geq \frac{\ell}{L} \} \) for \( 1 \leq \ell \leq L - 1 \). Also, let \( c_0 = -\infty \) and \( c_L = \infty \). The fact \( F(c_{\ell+1} - c_\ell) \leq \frac{1}{L} \) for \( 0 \leq \ell \leq L - 1 \) and the monotonicity of \( F \) and \( G \) implies the following. For \( 0 \leq \ell \leq L - 1 \) and \( c \in (c_\ell, c_{\ell+1}) \),
\[ G(c) \geq G(c_\ell) \geq F(c_\ell) - \delta \geq F(c) - \delta - \frac{1}{L}. \]

and similarly
\[ G(c) \leq G(c_{\ell+1}) \leq F(c_{\ell+1}) + \delta \leq F(c) + \delta + \frac{1}{L}. \]

Since \( \mathbb{R} \subset \{c_1, \ldots, c_L\} \cup [F(c_1), F(c_{L-1})] \), it follows that \( |F(c) - G(c)| \leq \delta + \frac{1}{L} \) for all \( c \in \mathbb{R} \), as was to be proved.

**Corollary 2:** If \( F \) is a CDF, there is a countable sequence \( (c_\ell: \ell \geq 1) \) such that, for any sequence of CDFs \( (F_n: n \geq 1) \), \( \sup_{r \in \mathbb{R}} |F_n(c) - F(c)| \to 0 \) if and only if \( F_n(c_\ell) \to F(c_\ell) \) and \( F_n(c_{\ell+1}) \to F(c_{\ell+1}) \) as \( n \to \infty \) for all \( \ell \geq 1 \).

**Proof:** Given \( F \), let \( (L_j: j \geq 1) \) be a sequence of integers converging to \( \infty \). For each \( j \), Lemma 2 implies the existence of \( L_j - 1 \) values \( c_\ell \) with a specified property. Let the infinite sequence \( (c_\ell: \ell \geq 1) \) be obtained by concatenating those finite sequences.

Alternatively, Corollary 2 is a consequence of Polya’s theorem, which states uniform convergence of CDFs is equivalent to pointwise convergence on the union of a dense subset of \( \mathbb{R} \) and \( \{-\infty, \infty\} \) (see, e.g., [10]).

**APPENDIX E**

**Proof of Consistency of ML Estimator**

The proof of the Proposition is given in this section after some preliminary results. Define \( \varphi(\lambda) \) for \( 0 \leq \lambda \leq 1 \) by
\[ \varphi(\lambda) \triangleq \begin{cases} \mathbb{E} \left[ \frac{1}{\lambda + (1-\lambda) R} \right] & \text{if } 0 \leq \lambda < 1, \\
1 & \text{if } \lambda = 1. \end{cases} \]
(13)

For any fixed \( \lambda \in [0, 1] \), \( \varphi_n(\lambda) \) is the average of \( n \) independent random variables with mean \( \varphi(\lambda) \), so by the law of large numbers, \( \varphi_n(\lambda) \to \varphi(\lambda) \) a.s. as \( n \to \infty \). Note that \( \varphi \) is finite over \([0, 1]\), continuous over \([0, 1]\), and convex over \([0, 1]\).

**Lemma 3:** If \( F_0 \) is not identical to \( F_1 \), exactly one of the following happens:
1) \( \varphi(1) < 1 \) and \( \varphi \) is convex;
2) \( \varphi(1) = 1 \) and \( \varphi \) is strictly convex.

**Proof:** Note that \( \varphi(\lambda) \leq \sup_{r \in [0, \infty]} \frac{1}{\lambda + (1-\lambda) r} = \frac{1}{\lambda} \) for \( \lambda \in (0, 1] \), so \( \varphi(1) \leq 1 \). The function \( \varphi \) is convex because it is the expectation of a convex function. If \( \varphi(1) = 1 \), then \( \mathbb{P} \{R_i = \infty\} = 0 \) and since it is also assumed that \( F_0 \) is not identical to \( F_1 \), \( \mathbb{P} \{R_i \notin \{1, \infty\}\} > 0 \). Hence, the function in the expectation defining \( \varphi \) is strictly convex with positive probability, so \( \varphi \) is strictly convex.

**Lemma 4:** Suppose \( F_0 \) is not identical to \( F_1 \) and let \( \lambda_n \) be defined as in Corollary 1. Then \( \lambda_n \to \alpha \) a.s. as \( n \to \infty \).

**Proof:** Suppose \( \alpha = 0 \). Then
\[ \varphi(0) = \mathbb{E}_1 \left[ \frac{1}{R} \right] = \int_0^\infty r dF_0(r) = 1 - F_0(0) \leq 1. \]

If \( \varphi(0) < 1 \), then, since \( \varphi_n(0) \to \varphi(0) \) a.s. as \( n \to \infty \), \( \varphi_n(0) < 1 \) for all sufficiently large \( n \). So \( \lambda_n = 0 = \alpha \) for all sufficiently large \( n \), with probability one.
If $\varphi(0) = 1$, then by Lemma 5 it follows that $\varphi(\lambda) < 1$ for $\lambda \in (0, 1)$. So for any such $\lambda$ fixed, $\varphi_n(\lambda) < 1$ for all sufficiently large $n$ with probability one. Thus, for any fixed $\lambda \in (0, 1)$, $\lambda_n \leq \lambda$ for all large $n$ with probability one, so $\lambda_n \to 0$ a.s. as $n \to \infty$. This implies the lemma for $\alpha = 0$.

Suppose $\alpha \in (0, 1)$. Note that

$$
\varphi(\alpha) = \int_0^\infty \frac{1}{\alpha + (1 - \alpha)r} (\alpha + (1 - \alpha) r) F_0(r) = 1.
$$

Therefore, Lemma 3 implies that, for any $\epsilon > 0$ such that $\alpha + \epsilon < 1$ and $\alpha - \epsilon > 0$, it holds that $\varphi(\alpha + \epsilon) < 0$ and $\varphi(\alpha - \epsilon) > 0$. Therefore, with probability one, $\varphi_n(\alpha + \epsilon) < 0$ and $\varphi_n(\alpha - \epsilon) > 0$ for all sufficiently large $n$, and therefore $|\lambda_n - \alpha| < \epsilon$ for all sufficiently large $n$ with probability one. This implies the lemma for $\alpha \in (0, 1)$.

Suppose $\alpha = 1$. Since $P \{ R_i < \infty \} = P \{ R_i = \infty \} = 1$ it holds that $\varphi(1) = 1$, so by Lemma 3, $\varphi$ is strictly convex. Furthermore, $\varphi(1) = E_0[R] - 1 \leq 0$. Therefore, $\varphi(\lambda) > 1$ for $\lambda \in [0, 1)$. Thus, for any fixed $\lambda \in [0, 1)$, $\varphi_n(\lambda) > 1$ for all sufficiently large $n$, with probability one. This implies the lemma for $\alpha = 1$, as needed. The proof of the lemma is complete.

Define cumulative distribution functions $\hat{G}_0$ and $\hat{G}_1$ by

$$
\hat{G}_0^c(\tau) = \min \left\{ \frac{1}{n} \sum_{i=1}^n I(\tau > r) \frac{1}{\alpha + (1 - \alpha) R_i} \right\},
$$

$$
\hat{G}_1(\tau) = \min \left\{ \frac{1}{n} \sum_{i=1}^n I(\tau \leq r) \frac{R_i}{\alpha + (1 - \alpha) R_i} \right\},
$$

for $\tau \in [0, \infty)$.

**Lemma 5:** As $n \to \infty$,

$$
\sup_{\tau \in [0, \infty]} |\hat{F}_k(\tau) - \hat{G}_k(\tau)| \to 0 \quad \text{a.s. for } k \in \{0, 1\}. \quad (14)
$$

**Proof:** The following conditions are equivalent: $F_0$ is identical to $F_1$: $F_0(\{1\}) = 1$; $F_1(\{1\}) = 1$; $P \{ R_i = 1 \} = 1$. If any of these conditions hold then $R_i = 1$ for all $i$ with probability one, so by Corollary $1$, $\hat{F}_0(\{1\}) = \hat{F}_1(\{1\}) = 1$. Also, $\hat{G}_0(\{1\}) = \hat{G}_1(\{1\}) = 1$. So the lemma is true if $F_0$ is identical to $F_1$. For the remainder of the proof suppose $F_0$ is not identical to $F_1$, which by Lemma 4 implies that $\lambda_n \to \alpha$ a.s. as $n \to \infty$.

If $0 < \alpha < 1$, the convergence $(14)$ follows immediately from the fact (based on $\lambda_n \to \alpha$) that, as $n \to \infty$, the function $r \mapsto \frac{1}{\lambda_n + (1 - \lambda_n) r}$ converges uniformly over all $r \in [0, \infty]$ to $\frac{1}{\alpha + (1 - \alpha) r}$, and the function $r \mapsto \frac{R_i}{\lambda_n + (1 - \lambda_n) r}$ converges uniformly over all $r \in [0, \infty]$ to $\frac{R_i}{\alpha + (1 - \alpha) r}$.

The proof of $(14)$ in case $\alpha = 0$ or $\alpha = 1$ is more subtle. Here we give the proof for $\alpha = 0$ and $k = 0$. The other three possibilities for $\alpha$ and $k$ follow in the same way. So consider the case $\alpha = 0$. The random variables $R_1, R_2, \ldots$ are independent and all have CDF $F_1$, and

$$
\hat{G}_0^c(\tau) = \min \left\{ \frac{1}{n} \sum_{i=1}^n I(\tau > r) \frac{1}{R_i} \right\} \quad \text{for } \alpha = 0.
$$

Fix an arbitrary $\delta > 0$ and let $\epsilon > 0$ be so small that $\epsilon < 1$ and $2(F_0(\epsilon) - F_0(0)) < \delta$. For any CDF $F$ and $\tau \in [0, \epsilon]$,

$F_c(\tau) = (F_c(\tau) - F_0(\epsilon)) + F_0(\epsilon)$ and $|F_c(\tau) - F_0(\epsilon)| \leq F_0(\epsilon) - F_0(\epsilon)$. Also note that (since $\epsilon < 1$) $\hat{F}_0(0) - \hat{F}_0(\epsilon) \leq \hat{G}_0^c(0) - \hat{G}_0^c(\epsilon)$. Therefore, for $\tau \in [0, \epsilon]$,

$$
|\hat{F}_0^c(\tau) - \hat{G}_0^c(\tau)| \leq |\hat{F}_0^c(\tau) - \hat{G}_0^c(\tau)| + 2|\hat{G}_0^c(0) - \hat{G}_0^c(\epsilon)|.
$$

Thus,

$$
\sup_{\tau \in [0, \epsilon]} |\hat{F}_0^c(\tau) - \hat{G}_0^c(\tau)| \leq \sup_{\tau \in [0, \epsilon]} |\hat{F}_0^c(\tau) - \hat{G}_0^c(\tau)| + 2|\hat{G}_0^c(0) - \hat{G}_0^c(\epsilon)|. \quad (15)
$$

Since $\lambda_n \to 0$ with probability one, the function $r \mapsto \frac{1}{\lambda_n + (1 - \lambda_n) r}$ converges uniformly over all $r \in [\epsilon, \infty]$ to $\frac{1}{\epsilon}$.

It follows that the supremum term on the right side of $(15)$ converges to zero a.s. as $n \to \infty$. Since

$$
|\hat{G}_0^c(0) - \hat{G}_0^c(\epsilon)| \leq \frac{1}{n} \sum_{i=1}^n I(0 < R_i \leq \epsilon) \frac{1}{R_i},
$$

and

$$
E \left[ I(0 < R_i \leq \epsilon) \frac{1}{R_i} \right] = \int_{\epsilon}^\infty \frac{1}{r} dF_0(r) = \int_{\epsilon}^\infty dF_0(r) = F_0(\epsilon) - F_0(0),
$$

the law of large numbers implies

$$
\lim_{n \to \infty} \sup_{\tau \in [0, \epsilon]} |\hat{F}_0^c(\tau) - \hat{G}_0^c(\tau)| \leq 2(F_0(\epsilon) - F_0(0)) < \delta
$$

with probability one. So $\sup_{\tau \in [0, \epsilon]} |\hat{F}_0^c(\tau) - \hat{G}_0^c(\tau)| < \delta$ for all sufficiently large $n$, with probability one. Since $\delta > 0$ was selected arbitrarily, this completes the proof of $(14)$ for $k = 0$ in case $\alpha = 0$, and hence the proof of Lemma 5 overall.

**Lemma 6:** As $n \to \infty$,

$$
\sup_{\tau \in [0, \epsilon]} |\hat{G}_k(\tau) - F_k(\tau)| \to 0 \quad \text{a.s. for } k \in \{0, 1\}. \quad (16)
$$

**Proof:** Note that

$$
E \left[ I(R_i > \tau) \frac{1}{\alpha + (1 - \alpha) R_i} \right] = \int_{\tau}^{\infty} \frac{1}{\alpha + (1 - \alpha) r} (\alpha + (1 - \alpha) r) dF_0(r) = F_0^2(\tau),
$$

$$
E \left[ I(R_i \leq \tau) \frac{R_i}{\alpha + (1 - \alpha) R_i} \right] = \int_{0}^{\tau} \frac{R_i}{\alpha + (1 - \alpha) r} (\alpha + (1 - \alpha)) dF_1(r) = F_1(\tau).
$$

Hence, by the law of large numbers, for any fixed $\tau \in [0, \epsilon]$,

$\hat{G}_k(\tau) \to F_k(\tau)$ with probability one as $n \to \infty$, for $k \in \{0, 1\}$. It can similarly be shown that $\hat{G}_k(\tau) \to F_k(\tau)$ with probability one as $n \to \infty$, for $k \in \{0, 1\}$ for each $\tau$. 

fixed. Pointwise convergence of CDFs and their corresponding left limits implies uniform convergence (see Appendix D), implying [16].

**Proof of Proposition 4** Lemmas 5 and 6 and the triangle inequality:

$$|\hat{F}_k(\tau) - F_k(\tau)| \leq |\hat{F}_k(\tau) - \hat{G}_k(\tau)| + |\hat{G}_k(\tau) - F_k(\tau)|,$$

imply that as $n \to \infty$,

$$\sup_{\tau \in [0, \infty)} |\hat{F}_k(\tau) - F_k(\tau)| \to 0 \quad \text{a.s. for } k \in \{0, 1\}.$$

Application of Lemma 1 completes the proof.

**APPENDIX F**

**DERIVATION OF EXPRESSIONS FOR AUC AND \(\overline{\text{AUC}}_{\text{ML}}\)**

**Proof of Proposition 3** (Proof of 1) Let $R_{[1]} \leq R_{[2]} \leq \ldots \leq R_{[n]}$ denote a reordering of the samples $R_{1}, \ldots, R_{n}$. Then the region under $\text{ROC}_{\text{ML}}$ can be partitioned into a union of trapezoidal regions, such that there is one trapezoid for each $R_{[i]}$ such that $R_{[i]} < \infty$. The trapezoids are numbered from right to left. If a value $v_j \in (0, \infty)$ is taken on by $c_j$ of the samples, then the union of the trapezoidal regions corresponding to those samples is also a trapezoidal region.

The area of the $i$th trapezoidal region is the width of the base times the average of the lengths of the two sides. The width of the base is $\frac{1}{n} \cdot \lambda_n + (1 - \lambda_n)R_{[i]}$, corresponding to a term in $\hat{F}_0$. The length of the left side is $\frac{1}{n} \sum_{i'=i+1}^{n} \frac{1}{\lambda_n + (1 - \lambda_n)R_{[i']}}$, and the length of the right side is greater than the length of the left side by $\frac{1}{n} \cdot \frac{1}{\lambda_n + (1 - \lambda_n)R_{[i]}}$. Summing the areas of the trapezoids yields:

$$\overline{\text{AUC}}_{\text{ML}} = \frac{1}{n^2} \sum_{i=1}^{n} \left\{ \frac{1}{(\lambda_n + (1 - \lambda_n)R_{[i]})} \cdot \left( \sum_{i'=i+1}^{n} \frac{R_{[i']}}{\lambda_n + (1 - \lambda_n)R_{[i']}} \right) + \frac{1}{2} \frac{R_{[i]}}{\lambda_n + (1 - \lambda_n)R_{[i]}} \right\},$$

which is equivalent to the expression given 1) of the proposition.

(Proof of 2) The consistency of $\overline{\text{AUC}}_{\text{ML}}$ follows from Proposition 4, the consistency of $\text{ROC}_{\text{ML}}$.

(Proof of 3) Let $\tau(p)$ and $\eta(p)$ denote values $\tau(p) \in [0, \infty)$ and $\eta(p) \in [0, 1]$ such that $F_0^\prime(\tau(p), \eta(p)) = p$. Then

$$\text{AUC} = \int_0^1 \text{ROC}(p) \, dp = \int_0^1 F_1^\prime(\tau(p), \eta(p)) \, dp$$

$$\begin{align*}
&= \int_0^1 (\eta(p)F_1^\prime(\tau(p)) + (1 - \eta(p))F_1^\prime(\tau(p) -)) \, dp \\
&\quad (a) \int_0^1 F_1^\prime(\tau(p)) + F_1^\prime(\tau(p) -) \, dp \\
&\quad (b) \sqrt{\mathbb{E}_0 F_1^\prime(R) + \mathbb{E}_0 F_1^\prime(R -)} \\
&\quad = \frac{\mathbb{E}_0 F_1^\prime(R) + \mathbb{E}_0 F_1^\prime(R -)}{2} \\
&\quad = \frac{\mathbb{E}_0 \left[ \int_{R^+} r' \, dF_0(r') + \int_{R^+} r' \, dF_0(r') + F_1(\{\infty\}) \right]}{2} \\
&\quad = \mathbb{E}_0 \left\{ \int_{R^+} r' \, dF_0(r') + \frac{\int_{R^+} r' \, dF_0(r')}{2} + F_1(\{\infty\}) \right\}
\end{align*}$$

where (a) follows from the fact that $\tau(p)$ is affine over the maximal intervals of $p$ such that $\tau(p)$ is constant, so the integral is the same if $\tau(p)$ is replaced over each such interval by its average over the interval, and (b) follows from the fact that if $U$ is a random variable uniformly distributed on the interval $[0, 1)$, then the CDF of $\tau(U)$ is $F_0$ because for any $c \geq 0$, $\mathbb{P}\{\tau(U) > c\} = \mathbb{P}\{U \leq F_0(c)\} = F_0(c)$. This establishes (6) and (7).

(Proof of 4) This follows from (6) and the fact the CDF of $R$ and $R'$ satisfies $dF(r) = (\alpha + (1 - \alpha)r) \, dF_0(r)$ over $[0, \infty)$ and $F(\{\infty\}) = (1 - \alpha)F_1(\{\infty\})$. 

\[\square\]