Sequential Nonparametric Two-sample Testing by Betting

IISA Conference, 2022

Shubhanshu Shekhar

Aaditya Ramdas

Department of Statistics and Data Science
Carnegie Mellon University

Sequential Two-Sample Testing

- Given a stream of **paired observations** on $\mathcal{X} \times \mathcal{X}$

$$(X_1, Y_1),\ (X_2, Y_2),\ \ldots \sim P_X \times P_Y \text{ i.i.d.},$$

- decide between the hypotheses:

$$H_0 : P_X = P_Y \quad \text{and} \quad H_1 : P_X \neq P_Y.$$

Goal

For $\alpha \in (0, 1)$, construct a level-α sequential test of power one.

- Under H_0: continue forever w.p. $\geq 1 - \alpha$.
- Under H_1: stop sampling, and reject the null as soon as possible.

Batch Two-Sample Testing

- Here, we have batches of observations: $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ drawn i.i.d.from P_X and P_Y respectively.
- A popular class of batch tests based on statistical distances $d : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$.
  - Define a test statistic $T_{n,m} = d\left(\widehat{P}_{X,n}, \widehat{P}_{Y,m}\right)$.
  - Reject the null, if $T_{n,m}$ is large.
- E.g., χ^2-test, Kolmogorov-Smirnov (KS) test, kernel-MMD test.
- Theoretical and empirical properties have been well studied.
- No such general framework for constructing sequential two-sample tests of power one.

Prior Sequential Nonparametric Tests of Power One

- Darling & Robbins (1968): based on time-uniform DKW inequalitites for univariate observations.
- Balsubramani & Ramdas (2016): based on a confidence sequence (CS) for linear-time kernel-MMD statistic
- Lheritier & Cazals (2017): based on sequential binary classifiers
- Howard & Ramdas (2021): based on CSs using forward supermartingales
- Manole & Ramdas (2021): based on CSs using reverse submartingales.

Prior Sequential Nonparametric Tests of Power One

All existing methods either have strong theoretical guarantees or good empirical performance; but not both.

Prior Sequential Nonparametric Tests of Power One

All existing methods either have strong theoretical guarantees or good empirical performance; but not both.

This Talk

- A fundamentally new framework for designing powerful sequential two-sample tests.
- We take the perspective of a fictitious bettor, repeatedly betting on the observations to disprove the null.
  - **Constraints:** The bets must be fair under H_0, and the bettor cannot borrow money.
- The gain in the bettor's wealth (i.e., W_t/W_0) is a measure of evidence collected against the null.

The Betting Game

Bettor begins with an initial wealth, $W_0 = 1$.

For $t = 1, 2, \ldots$:

- Bettor selects a function $g_t : \mathcal{X} \to [-1/2, 1/2]$.
  - defines a fair payoff function under H_0, $h_t(x, y) = g_t(x) - g_t(y)$.
- Bettor chooses a fraction, $\lambda_t \in [0, 1]$, of his current wealth, W_{t-1}, to gamble.
- The next paired observation, (X_t, Y_t), is revealed.
- Bettor's wealth is updated as follows:

$$\begin{aligned} W_t &= W_{t-1} \times (1 - \lambda_t) + W_{t-1}\lambda_t\big(1 + h_t(X_t, Y_t)\big) \\ &= W_0 \times \prod_{i=1}^{t} \Big(1 + \lambda_i\big(g_i(X_i) - g_i(Y_i)\big)\Big) \end{aligned}$$

From Betting to Sequential Testing

Under H_0, we have $\mathbb{E}[g_t(X_t) - g_t(Y_t) | \mathcal{F}_{t-1}] = 0$. Hence, $\{W_t : t \geq 0\}$ is a **test martingale** — a non-negative martingale with an initial value 1.

From Betting to Sequential Testing

Under H_0, we have $\mathbb{E}[g_t(X_t) - g_t(Y_t) | \mathcal{F}_{t-1}] = 0.$ Hence, $\{W_t : t \geq 0\}$ is a **test martingale** — a non-negative martingale with an initial value 1.

Ville's Inequality (1939)

For any test martingale $\{W_t : t \geq 0\}$ and an $\alpha \in (0, 1]$, we have

$$\mathbb{P}\left(\exists t \geq 0 : W_t \geq \frac{1}{\alpha}\right) \leq \alpha.$$

JEAN VILLE

Étude critique de la notion de collectif

Thèses de l'entre-deux-guerres, 1939

From Betting to Sequential Testing

Under H_0, we have $\mathbb{E}[g_t(X_t) - g_t(Y_t) | \mathcal{F}_{t-1}] = 0.$ Hence, $\{W_t : t \geq 0\}$ is a **test martingale** — a non-negative martingale with an initial value 1.

Ville's Inequality (1939)

For any test martingale $\{W_t : t \geq 0\}$ and an $\alpha \in (0, 1]$, we have

$$\mathbb{P}\left(\exists t \geq 0 : W_t \geq \frac{1}{\alpha}\right) \leq \alpha.$$

- Define the test (i.e., a stopping time):

$$\tau := \min\{t \geq 1 : W_t \geq 1/\alpha\}.$$

- For arbitrary (predictable) sequences $\{(g_t, \lambda_t) : t \geq 1\}$, Ville's inequality implies

$$\mathbb{P}(\tau < \infty) \leq \alpha, \quad \text{under } H_0.$$

Performance of the test under H_1

Under H_1, we require $\{W_t : t \geq 0\}$ to grow rapidly to infinity.

Performance of the test under H_1

Under H_1, we require $\{W_t : t \geq 0\}$ to grow rapidly to infinity.

Faster growth of $W_t \implies$ Stronger statistical properties of τ

Performance of the test under H_1

Under H_1, we require $\{W_t : t \geq 0\}$ to grow rapidly to infinity.

Faster growth of $W_t \implies$ Stronger statistical properties of τ

- **Consistency.**

$$\mathbb{P}\left(\exists n \geq 1 : W_n \geq 1/\alpha\right) = 1 \quad \Rightarrow \quad \mathbb{P}(\tau < \infty) = 1.$$

Performance of the test under H_1

Under H_1, we require $\{W_t : t \geq 0\}$ to grow rapidly to infinity.

Faster growth of $W_t \Longrightarrow$ Stronger statistical properties of τ

- **Consistency.**

$$\mathbb{P}\left(\exists n \geq 1 : W_n \geq 1/\alpha\right) = 1 \quad \Rightarrow \quad \mathbb{P}(\tau < \infty) = 1.$$

- **Exponential consistency.**

$$\liminf_{n\to\infty} \frac{-1}{n} \log\left(\mathbb{P}(W_n < 1/\alpha)\right) > 0 \quad \Rightarrow \quad \liminf_{n\to\infty} \frac{-1}{n} \log\left(\mathbb{P}(\tau > n)\right) > 0.$$

Performance of the test under H_1

Under H_1, we require $\{W_t : t \geq 0\}$ to grow rapidly to infinity.

Faster growth of $W_t \Longrightarrow$ Stronger statistical properties of τ

- **Consistency.**

$$\mathbb{P}\left(\exists n \geq 1 : W_n \geq 1/\alpha\right) = 1 \quad \Rightarrow \quad \mathbb{P}(\tau < \infty) = 1.$$

- **Exponential consistency.**

$$\liminf_{n \to \infty} \frac{-1}{n} \log\left(\mathbb{P}(W_n < 1/\alpha)\right) > 0 \quad \Rightarrow \quad \liminf_{n \to \infty} \frac{-1}{n} \log\left(\mathbb{P}(\tau > n)\right) > 0.$$

- **Finite Expected Stopping Time.**

$$\sum_{n \geq 0} \mathbb{P}\left(W_n < \frac{1}{\alpha}\right) < \infty \quad \Rightarrow \quad \mathbb{E}[\tau] = \sum_{n=0}^{\infty} \mathbb{P}\left(\tau > n\right) < \infty.$$

Summary so far

- We defined a sequential test: $\tau = \min\{t \geq 1 : W_t \geq 1/\alpha\}$.
- $\{W_t : t \geq 1\}$ is the wealth of a fictitious bettor, betting on the observations in a repeated game with $W_0 = 1$.
- **Under H_0**, for arbitrary predictable $\{(g_t, \lambda_t) : t \geq 1\}$, we have $\mathbb{P}(\tau < \infty) \leq \alpha$.
- **Under H_1**, statistical properties of τ depend on how quickly W_t grows to infinity.
  - this depends strongly on the choice of $\{(\lambda_t, g_t) : t \geq 1\}$.
- Rest of the talk: A principled approach for selecting $\{(\lambda_t, g_t) : t \geq 1\}$.

Overview of our approach

- **Step 1:** Select an appropriate function class $\mathcal{G}$
  - Or equivalently, an Integral Probability Metric (IPM)
- **Step 2:** Design an “Oracle Test”
  - Uses terms, g^* and λ^*, depending on the unknown P_X and P_Y
- **Step 3:** Design a practical sequential test
  - Uses a sequence of predictable estimates of g^* and λ^*

Step 1 – Select a function class $\mathcal{G}$

- For simplicity, we assume that $\mathcal{G}$ consists of functions taking values in $[-1/2, 1/2]$.
- Can define

$$d_{\mathcal{G}}(P_X, P_Y) = \max_{g \in \mathcal{G}} \mathbb{E}_{P_X}[g(X)] - \mathbb{E}_{P_Y}[g(Y)], \tag{1}$$

which is a metric if $\mathcal{G}$ is rich enough.

Step 1 – Select a function class $\mathcal{G}$

- For simplicity, we assume that $\mathcal{G}$ consists of functions taking values in $[-1/2, 1/2]$.
- Can define

$$d_{\mathcal{G}}(P_X, P_Y) = \max_{g \in \mathcal{G}} \ \mathbb{E}_{P_X}[g(X)] - \mathbb{E}_{P_Y}[g(Y)], \tag{1}$$

  which is a metric if $\mathcal{G}$ is rich enough.
- *Witness function*

$$g^* \in \arg\max_{g \in \mathcal{G}} \ \mathbb{E}_{P_X}[g(X)] - \mathbb{E}_{P_Y}[g(Y)]. \tag{2}$$

Step 1 – Select a function class $\mathcal{G}$

- For simplicity, we assume that $\mathcal{G}$ consists of functions taking values in $[-1/2, 1/2]$.
- Can define

$$d_{\mathcal{G}}(P_X, P_Y) = \max_{g \in \mathcal{G}} \mathbb{E}_{P_X}[g(X)] - \mathbb{E}_{P_Y}[g(Y)], \tag{1}$$

  which is a metric if $\mathcal{G}$ is rich enough.
- *Witness function*

$$g^* \in \arg\max_{g \in \mathcal{G}} \mathbb{E}_{P_X}[g(X)] - \mathbb{E}_{P_Y}[g(Y)]. \tag{2}$$

- g^* provides the maximum contrast between P_X and P_Y

Step 1 – Select a function class $\mathcal{G}$

- For simplicity, we assume that $\mathcal{G}$ consists of functions taking values in $[-1/2, 1/2]$.
- Can define

$$d_{\mathcal{G}}(P_X, P_Y) = \max_{g \in \mathcal{G}} \mathbb{E}_{P_X}[g(X)] - \mathbb{E}_{P_Y}[g(Y)], \tag{1}$$

  which is a metric if $\mathcal{G}$ is rich enough.
- *Witness function*

$$g^* \in \arg\max_{g \in \mathcal{G}} \mathbb{E}_{P_X}[g(X)] - \mathbb{E}_{P_Y}[g(Y)]. \tag{2}$$

- g^* provides the maximum contrast between P_X and P_Y
- If $P_X = P_Y$, then g^* is an arbitrary element of $\mathcal{G}$

Step 2 – Oracle Test

- Construct the 'oracle' process $\{W_t^* : t \geq 0\}$, with $W_0^* = 1$, and

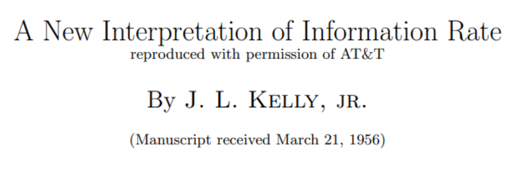
$$W_t^* = W_{t-1}^* \times \left(1 + \lambda^* \left(g^*(X_t) - g^*(Y_t)\right)\right),$$

- Construct the 'oracle' process $\{W_t^* : t \geq 0\}$, with $W_0^* = 1$, and

$$W_t^* = W_{t-1}^* \times \left(1 + \lambda^* \left(g^*(X_t) - g^*(Y_t)\right)\right),$$

- where λ^* is the **log-optimal betting fraction**:

$$\lambda^* \in \underset{\lambda \in (-1,1)}{\arg\max}\, \mathbb{E}\left[\log(1 + \lambda(g^*(X) - g^*(Y)))\right].$$


A New Interpretation of Information Rate
reproduced with permission of AT&T

By J. L. KELLY, JR.

(Manuscript received March 21, 1956)

Step 2 – Oracle Test

- Construct the 'oracle' process $\{W_t^* : t \geq 0\}$, with $W_0^* = 1$, and

$$W_t^* = W_{t-1}^* \times \left(1 + \lambda^*\left(g^*(X_t) - g^*(Y_t)\right)\right),$$

- where λ^* is the **log-optimal betting fraction**:

$$\lambda^* \in \operatorname*{arg\,max}_{\lambda \in (-1,1)} \mathbb{E}\left[\log(1 + \lambda(g^*(X) - g^*(Y)))\right].$$

- Define the 'oracle test': $\tau^* = \min\left\{t \geq 1 : W_t^* \geq \frac{1}{\alpha}\right\}$.

Step 2 – Oracle Test

- Construct the 'oracle' process $\{W_t^* : t \geq 0\}$, with $W_0^* = 1$, and

$$W_t^* = W_{t-1}^* \times \left(1 + \lambda^* \left(g^*(X_t) - g^*(Y_t)\right)\right),$$

- where λ^* is the **log-optimal betting fraction**:

$$\lambda^* \in \operatorname*{arg\,max}_{\lambda \in (-1,1)} \mathbb{E}\left[\log(1 + \lambda(g^*(X) - g^*(Y)))\right].$$

- Define the 'oracle test': $\tau^* = \min\left\{t \geq 1 : W_t^* \geq \frac{1}{\alpha}\right\}$.
- The test τ^* is exponentially consistent, and has a finite expected stopping time.

Step 3 – Practical Test

- g^* and λ^* in τ^* are not known $\Rightarrow$ Use data-driven estimates.
- A **prediction strategy** ($\mathcal{A}_P$) to select $\{g_t : t \geq 1\} \approx g^*$.
  - Specific choice of $\mathcal{A}_P$ will depend on $\mathcal{G}$.
- A **betting strategy** ($\mathcal{A}_B$) to select $\{\lambda_t : t \geq 1\} \approx \lambda^*$.
  - Existing methods, such as Online Newton Step (ONS), are sufficient for our purposes.
- Construct the wealth process

$$W_t = W_{t-1} \times \left(1 + \lambda_t\big(g_t(X_t) - g_t(Y_t)\big)\right).$$

- Define the level-α test: $\tau = \min\left\{t \geq 1 : W_t \geq \frac{1}{\alpha}\right\}$

Summary: Steps of our sequential test

Initialization:

- A function class $\mathcal{G}$
- a prediction strategy ($\mathcal{A}_P$) to select $\{g_t : t \geq 1\}$
- ONS betting strategy ($\mathcal{A}_B$) to select $\{\lambda_t : t \geq 1\}$
- $W_0 = 1$

For $t = 1, 2, \ldots$:

- Get the next g_t from the prediction strategy, $\mathcal{A}_P$.
- Get the next λ_t from the betting strategy, $\mathcal{A}_B$.
- Observe the next pair (X_t, Y_t).
- Update $W_t = W_{t-1} \times \big(1 + \lambda_t\big(g_t(X_t) - g_t(Y_t)\big)\big)$.
- Reject H_0, if $W_t \geq 1/\alpha$.

Performance Guarantees

Smaller Regret of $\mathcal{A}_P$ $\Rightarrow$ Faster growth of W_t $\Rightarrow$ Stronger properties of the test τ.

Performance Guarantees

Regret of $\mathcal{A}_P$

$$\mathcal{R}_n(\mathcal{A}_P) = \sup_{g \in \mathcal{G}} \left[\left(\sum_{t=1}^{n} g(X_t) - g(Y_t) \right) - \left(\sum_{t=1}^{n} g_t(X_t) - g_t(Y_t) \right) \right].$$

Performance Guarantees

Regret of $\mathcal{A}_P$

$$\mathcal{R}_n(\mathcal{A}_P) = \sup_{g \in \mathcal{G}} \left[\left(\sum_{t=1}^{n} g(X_t) - g(Y_t) \right) - \left(\sum_{t=1}^{n} g_t(X_t) - g_t(Y_t) \right) \right].$$

Regret-Power Connections under H_1 (Informal)

- If $\lim_{n\to\infty} \mathcal{R}_n/n$ is smaller than $d_{\mathcal{G}}(P_X, P_Y)$ w.p. 1, the test τ is consistent.
- If $\mathcal{R}_n/n \to 0$ with sufficiently large probability, then $\mathbb{E}[\tau]$ is finite.
- If the $\mathcal{R}_n/n \to 0$ w.p. 1, then the test τ is exponentially consistent.

Application 1: Sequential KS Test

- $\mathcal{X} = \mathbb{R}$, and $\mathcal{G} = \{\mathbf{1}_{(-\infty,u]} - 0.5 : u \in \mathbb{R}\}$.
- Plug-in prediction strategy ($\mathcal{A}_{\text{plug-in}}$): $g_t = \mathbf{1}_{(-\infty,u_t]} - 0.5$, where

$$u_t \in \operatorname*{arg\,max}_{u \in \mathbb{R}} \; \widehat{F}_{X,t-1}(u) - \widehat{F}_{Y,t-1}(u).$$

- For any $n \geq 1$, $\mathcal{R}_n(\mathcal{A}_{\text{plug-in}})/n = O(1/\sqrt{n})$, w.p. $1 - 1/n^2$.
- Hence, the resulting test is consistent, and satisfies $\mathbb{E}[\tau] = \mathcal{O}\left(1/d_{\text{KS}}^2(P_X, P_Y)\right)$ under H_1.
- There exist distributions on which any test, τ', must have $\mathbb{E}[\tau'] = \Omega\left(1/d_{\text{KS}}^2(P_X, P_Y)\right)$.

Application 2: Sequential Kernel MMD Test

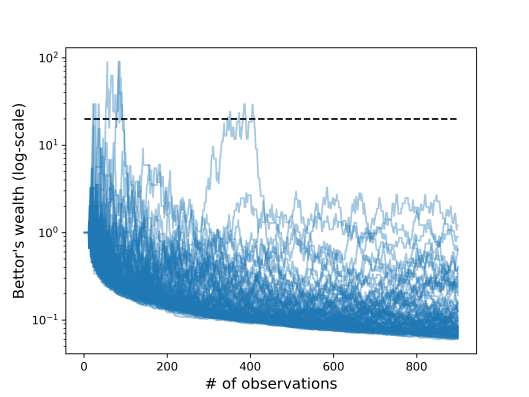
- General $\mathcal{X}$, and $\mathcal{G} = \{g \in \text{RKHS}(k) : \|g\|_k \leq 1\}$.
- Projected Gradient Ascent prediction strategy ($\mathcal{A}_{\text{PGA}}$)
- $\mathcal{A}_{\text{PGA}}$ satisfies $\mathcal{R}_n(\mathcal{A}_{\text{PGA}})/n = \mathcal{O}(1/\sqrt{n})$, w.p. 1.
- Hence, the resulting test is exponentially consistent, and satisfies $\mathbb{E}[\tau] = \mathcal{O}\left(1/d^2_{\text{MMD}}(P_X, P_Y)\right)$ under H_1.
- There exist distributions on which any test, τ', satisfies $\mathbb{E}[\tau'] = \Omega\left(1/d^2_{\text{MMD}}(P_X, P_Y)\right)$, under H_1.

An Example

Under H_0, the wealth process of the bettor rarely exceeds the level $1/\alpha$.

$$P_X = N(0, 1)$$

$$P_Y = N(0, 1)$$

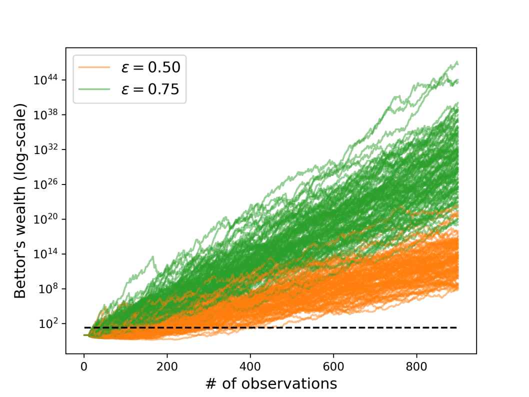


An Example

Under H_1, the wealth process with the plug-in prediction strategy, grows at an exponential rate.

$$P_X = N(0, 1)$$

$$P_Y = N(\varepsilon, 1)$$



Extension to time-varying distributions

Our ideas easily extend to the following case:

For $t = 1, 2, \ldots$:

- Bettor selects g_t and λ_t.
- Adversary selects distributions $P_{X,t}$ and $P_{Y,t}$.
- The pair, $(X_t, Y_t) \sim P_{X,t} \times P_{Y,t}$ is revealed.
- Update the wealth: $W_t = W_{t-1} \times \big(1 + \lambda_t\big(g_t(X_t) - g_t(Y_t)\big)\big)$.
- Reject the null if $W_t \geq 1/\alpha$.

Under some mild assumptions on $\mathcal{G}$, the test defined above is consistent.

Other Extensions and Generalizations

- Relaxing the assumption of paired observations.
- Relaxing the boundedness assumption on the functions in $\mathcal{G}$.
- A general problem unifying several tasks such as two-sample testing, independence testing, and symmetry testing.

Thank you.

Details of Regret-Power Result

- $\limsup_{n\to\infty} \frac{\mathcal{R}_n}{n} < d_{\mathcal{G}}(P_X, P_Y)$ a.s. $\Rightarrow \mathbb{P}_{P_{XY}}(\tau < \infty) = 1$.
- For a sequence $r_n \to 0$, define $E_n = \{\mathcal{R}_n/n \leq r_n\}$. Then,

$$\sum_{n\geq 1} \mathbb{P}_{P_{XY}}(E_n^c) < \infty \;\Rightarrow\; \mathbb{E}_{P_{XY}}[\tau] < \infty.$$

- If $\mathbb{P}_{P_{XY}}(E_n^c) = 0$ for some $r_n \to 0$, then we have

$$\liminf_{n\to\infty} \frac{-1}{2n} \mathbb{P}_{P_{XY}}(\tau > n) = \beta^*. \quad \text{(optimal exponent)}$$

Testing invariance to an operator

- Given a stream of observations: $U_1, U_2, \ldots$ on $\mathcal{U}$, drawn i.i.d.from P_U.
- Let $T : \mathcal{U} \to \mathcal{U}$ be a known operator.
- Consider the problem:

$$H_0 : P_U = P_U \circ T^{-1}, \quad \text{versus} \quad H_1 : P_U \neq P_U \circ T^{-1}.$$

- This formulation unifies several problems such as two-sample testing, independence testing, and symmetry testing.
- For two-sample testing:

$$\mathcal{U} = \mathcal{X} \times \mathcal{X}, \quad U = (X, Y), \quad P_U = P_X \times P_Y$$
$$T : \mathcal{X} \times \mathcal{X} \to \mathcal{X} \times \mathcal{X}, \quad \text{such that} \quad T(x, y) = (y, x).$$