Carnegie Mellon University
Statistics & Data Science

MACHINE LEARNING DEPARTMENT

Sequential Predictive Two-Sample and Independence Testing

Aleksandr Podkopaev, Aaditya Ramdas
Carnegie Mellon University

Preliminaries

Two-Sample Testing (2ST). Given iid draws from $P \times Q$, test:

$$H_0 : P = Q \qquad H_1 : P \neq Q$$

Independence Testing (IT). Given iid draws from P_{XY}, test:

$$H_0 : P_{XY} = P_X \times P_Y \qquad H_1 : P_{XY} \neq P_X \times P_Y$$

Issues with batch 2ST/IT. For both problems, even if H_0 is false, it is unknown a priori how much data are needed to reject H_0.

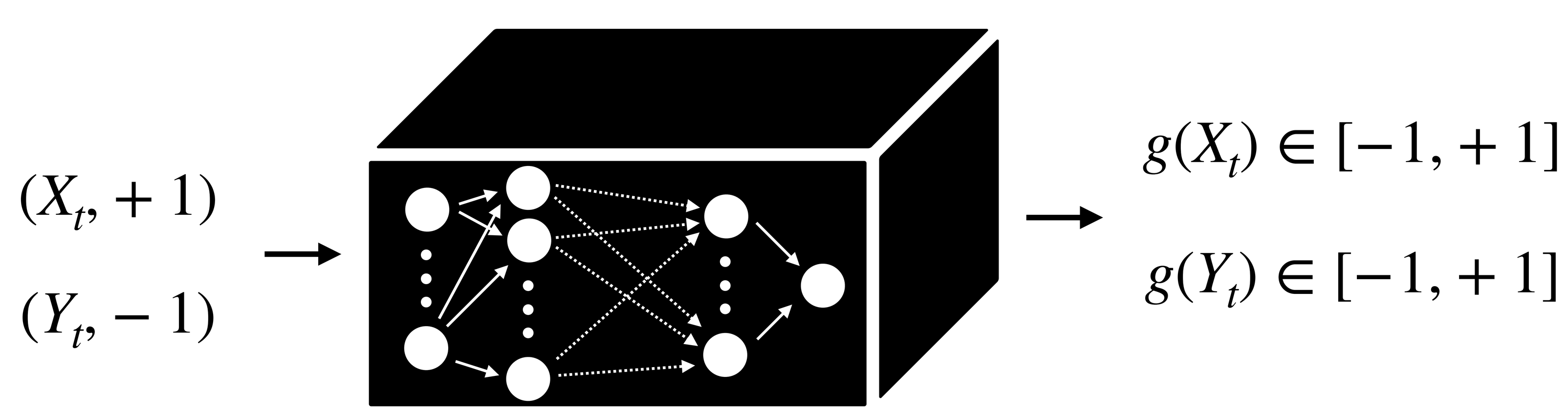
Sequential Test Φ: at time t, uses the first t points to output 0 (collect more data) or 1 (stop and reject H_0).
Stopping time $\tau := \inf\{t \geq 1 : \Phi((X_1, Y_1), \ldots, (X_t, Y_t)) = 1\}$.

$\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$	$\mathbb{P}_{H_1}(\tau < \infty) = 1$
"time-uniform" type-1 error control	"power-one tests" [Darling and Robbins, 1968]

(batch type-1 error control applies to a single sample size t specified before collecting data)

Classification-based 2ST: complements SOTA kernel 2STs for high-dimensional and structured data with localized differences.

1. Label data from P as positive and data from Q as negative.
2. Check whether one can learn a classifier that performs better than chance. If so, reject the null hypothesis.



H_0 **is true:** $g(X_t) \overset{d}{=} g(Y_t)$.

H_1 **is true:** (hopefully) $\mathbb{E}g(X_t) > \mathbb{E}g(Y_t)$.

Sequential nonparametric 2ST by betting

Protocol. Gambler starts with $\mathcal{K}_0 = 1$. At each round t:

1. Gambler selects:
   (a) a **fair** payoff function $f_t : \mathcal{X} \times \mathcal{X} \to [-1, \infty)$:
   $$\mathbb{E}_{H_0}[f_t(X, Y) \mid \mathcal{F}_{t-1}] = 0, \qquad \mathcal{F}_{t-1} = \sigma(\{(X_i, Y_i)\}_{i \leq t-1})$$
   (b) a fraction of wealth: $\lambda_t \in [0,1]$, to bet.
2. Nature reveals $(X_t, Y_t) \sim P \times Q$, and wealth is updated:
   $$\mathcal{K}_t = \mathcal{K}_{t-1} \cdot \left(1 + \lambda_t \cdot f_t(X_t, Y_t)\right)$$

Idea. Use wealth to measure evidence against H_0.

H_0 **is true:** Using our strategy, a gambler is not expected to gain any money.

$(\mathcal{K}_t)_{t \geq 0}$ is a nonnegative martingale starting at 1 for any $(f_t)_{t \geq 1}$ and $(\lambda_t)_{t \geq 1}$ that satisfy the above constraints.

For $\tau := \inf\{t \geq 1 : \mathcal{K}_t \geq 1/\alpha\}$, Ville's inequality implies:

$$\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha.$$

H_1 **is true:** Using our strategy makes gambler's wealth grow exponentially: $\mathbb{E}[\log(\mathcal{K}_t)/t] > 0$.

Payoff Functions. $f(X_t, Y_t, g_t) = \frac{1}{2}\left(g_t(X_t) - g_t(Y_t)\right)$

Betting Fractions. Follow the best $\lambda_\star$ in hindsight via Online Newton step [Hazan et al., 2007].

IT. At round t, bet on pair $(X_{2t-1}, Y_{2t-1}), (X_{2t}, Y_{2t}) \sim P_{XY}$ and use external randomization to produce instances from the product of the marginal distributions: $(X_{2t-1}, Y_{2t}), (X_{2t}, Y_{2t-1}) \sim P_X \times P_Y$.

Power and adaptivity to the complexity

H_1 **is true** (+ suppose that $g_\star$ has a positive margin):
$\mathcal{K}_t \overset{\text{a.s.}}{\to} +\infty$, which implies consistency:

$$\mathbb{P}_{H_1}(\tau < \infty) = 1$$

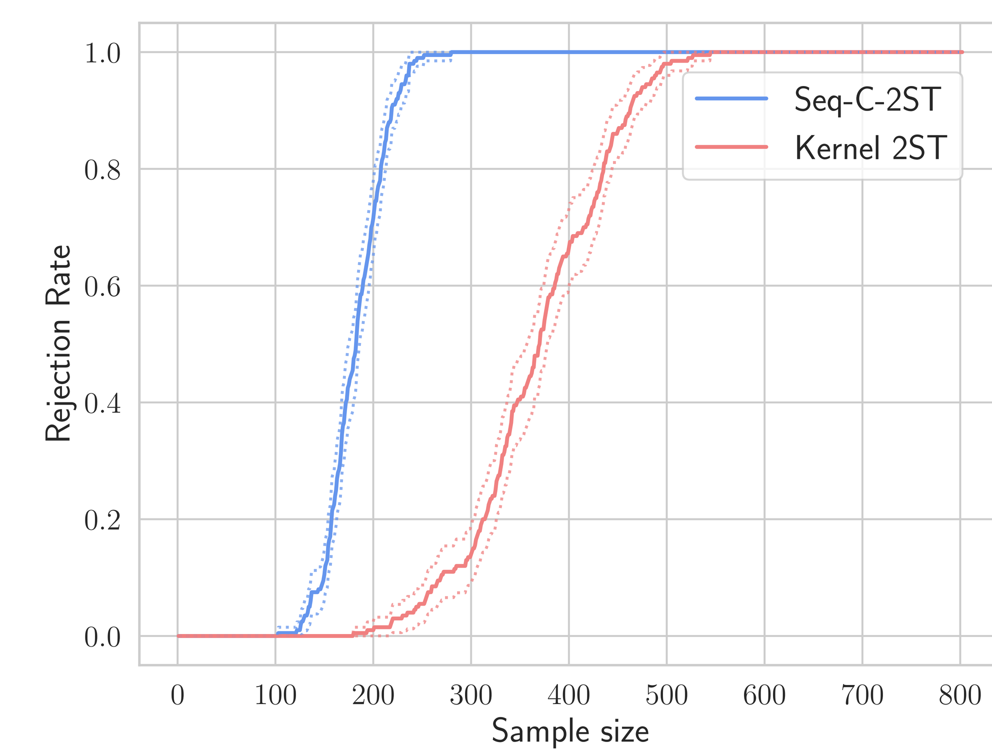
Wealth (proxy for power) grows exponentially:

$$\liminf_{t \to \infty} \frac{1}{t} \log \mathcal{K}_t \overset{\text{a.s.}}{\geq} \frac{1}{4} \mathbb{E} f(X, Y, g_\star)$$

KDEF dataset. [Lundqvist et al., 1998]

Instances from P:
Happy, neutral, surprised

Instances from Q:
Afraid, angry, disgusted



Also in the paper

- Oracle upper bounds on growth rate.
- Variance-adaptivity properties.
- Relationship to SPRT.
- Regression-based IT.