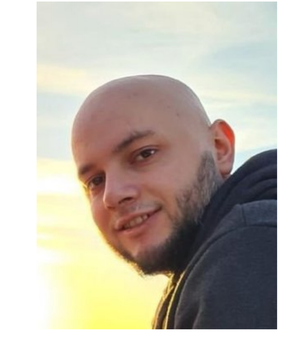
Carnegie Mellon University
Statistics & Data Science

Sequential Kernelized Independence Testing

(to appear at ICML 2023)

Aleksandr Podkopaev[1], Patrick Blöbaum[2],
Shiva Kasiviswanathan[2], Aaditya Ramdas[1,2]
[1]Carnegie Mellon University, [2]AWS

tl;dr consistent sequential nonparametric independence testing.

Preliminaries

Independence Testing (IT). Given iid draws $(X_1, Y_1), (X_2, Y_2), \ldots$ from P_{XY}, construct a test for:

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

1. X and Y need not take values in the same space.
2. No parametric assumptions on distributions.

Issue. 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, outputs 0 (collect more data) or 1 (reject H_0 and stop) based on first t points.

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: ***prespecified*** sample size t.

Kernel Measures of Dependence. Let $\mathcal{G}$ (and $\mathcal{H}$) be an RKHS with positive-definite kernel k (and l) and canonical feature map φ (and ψ) defined on $\mathcal{X}$ (and $\mathcal{Y}$).

$$\mathrm{HSIC}(P_{XY}; \mathcal{G}, \mathcal{H}) = \| \mu_{XY} - \mu_X \otimes \mu_Y \|^2$$

$$\mu_{XY} = \mathbb{E}_{P_{XY}}[\varphi(X) \otimes \psi(Y)] \quad \mu_X = \mathbb{E}_{P_X}[\varphi(X)] \quad \mu_Y = \mathbb{E}_{P_Y}[\psi(Y)]$$

$$\|\mu_{XY} - \mu_X \otimes \mu_Y\| = \sup_{g:\|g\| \leq 1} \langle g, \mu_{XY} - \mu_X \otimes \mu_Y \rangle$$

$$g_\star = \frac{\mu_{XY} - \mu_X \otimes \mu_Y}{\|\mu_{XY} - \mu_X \otimes \mu_Y\|}$$

witness function (notices maximum discrepancy)

- For 1-d and linear kernel, $\mathrm{HSIC}(P_{XY}; \mathcal{G}, \mathcal{H}) = (\mathrm{Cov}(X, Y))^2$.
- For common kernels, characteristic condition holds:

$$\mathrm{HSIC}(P_{XY}; \mathcal{G}, \mathcal{H}) = 0 \text{ iff } H_0 \text{ is true } (> 0 \text{ otherwise})$$

Sequential nonparametric IT by betting

Protocol. (Bet on two observations from P_{XY})

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

1. Gambler selects:
   (a) a **fair** payoff function $f_t : (\mathcal{X} \times \mathcal{Y})^2 \to [-1, \infty)$:
   $$\mathbb{E}_{H_0}[f_t((X, Y), (X', Y')) \mid \mathcal{F}_{t-1}] = 0, \qquad \mathcal{F}_{t-1} = \sigma(\{(X_i, Y_i)\}_{i \leq 2t})$$
   (b) a fraction of wealth: $\lambda_t \in [-1, 1]$, to bet.
2. Nature reveals two points from P_{XY}, and wealth is updated:
   $$\mathcal{K}_t = \mathcal{K}_{t-1} \cdot \left(1 + \lambda_t \cdot f_t((X_{2t+1}, Y_{2t+1}), (X_{2t+2}, Y_{2t+2}))\right)$$

Idea. Use wealth to measure evidence against H_0.

$$\tau := \inf\{t \geq 1 : \mathcal{K}_t \geq 1/\alpha\}$$

H_0 is true: $(\mathcal{K}_t)_{t \geq 0}$ is a nonnegative martingale for any $(f_t)_{t \geq 1}$ and $(\lambda_t)_{t \geq 1}$ that satisfy the above constraints.

By Ville's inequality

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

Goal. Pick $(f_t)_{t \geq 1}$, $(\lambda_t)_{t \geq 1}$ to guarantee wealth growth under H_1.

Payoff Functions. (replace terms in HSIC with estimators)

plug-in witness function
computed A $\{(X_i, Y_i)\}_{i \leq 2t}$

$$f_t((X, Y), (X', Y')) = \left\langle \hat{g}_t, \frac{1}{2}\left(\varphi(X') - \varphi(X)\right) \otimes \left(\psi(Y') - \psi(Y)\right) \right\rangle$$

unbiased estimator of $\mu_{XY} - \mu_X \otimes \mu_Y$
computed from $(X, Y), (X', Y')$

(computation requires linear in t kernel evaluations)

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

Power and adaptivity to the complexity

H_1 is true: $\mathcal{K}_t \xrightarrow{\text{a.s.}} +\infty$, which implies consistency:

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

Wealth (proxy for power) grows exponentially:

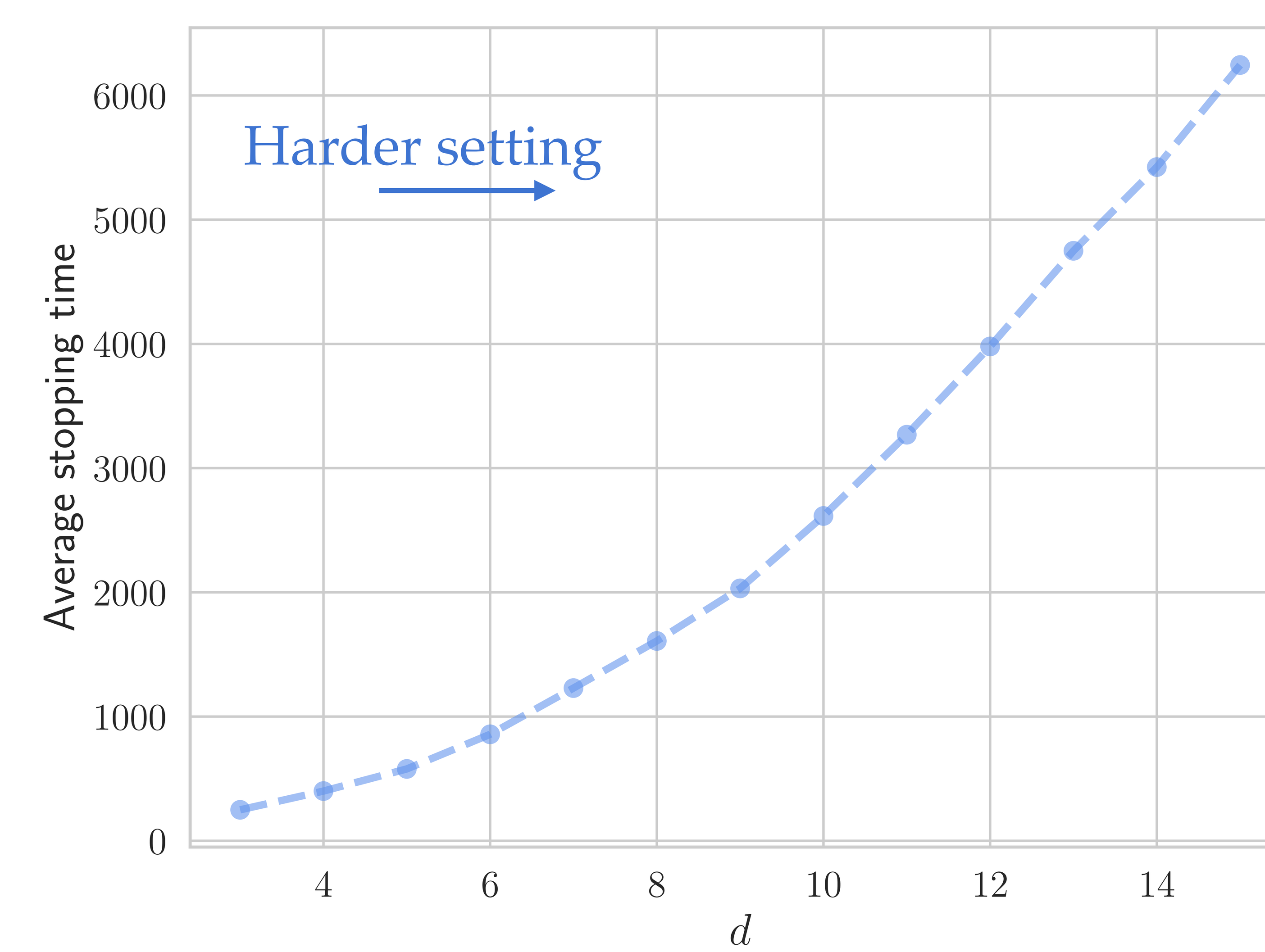
$$\liminf_{t \to \infty} \frac{1}{t} \log \mathcal{K}_t \overset{\text{a.s.}}{\geq} \frac{M_1}{4} \cdot \left(\frac{M_1}{M_2} \wedge 1\right)$$

$$M_1 = \mathbb{E}\, f_\star((X, Y), (X', Y')) = \sqrt{\mathrm{HSIC}(P_{XY}; \mathcal{G}, \mathcal{H})}$$

$$M_2 = \mathbb{E}\, f_\star^2((X, Y), (X', Y')) \leq 1$$

Corollary.
$$\liminf_{t \to \infty} \frac{1}{t} \log \mathcal{K}_t \overset{\text{a.s.}}{\geq} \frac{1}{4} \mathrm{HSIC}(P_{XY}; \mathcal{G}, \mathcal{H})$$

$$(X_t, Y_t) = (U_t^{(1)}, U_t^{(2)}),\ U_t \sim \mathrm{Unif}(\mathbb{S}^d)$$



Also in the paper

- IT beyond the iid case & testing instantaneous independence
- Alternative kernel measures of dependence (COCO, KCC)
- Extensions to unbounded kernels (via reduction to testing symmetry)