Bet Imaginatively, not Historically in Independent-Data Sequential Testing

Nathaniel Xu
University of British Columbia
xunathan@cs.ubc.ca

Feng Liu
University of Melbourne
feng.liu1@unimelb.edu.au

Danica J. Sutherland
UBC & Amii
dsuth@cs.ubc.ca

Abstract

Sequential testing avoids some of the many controversies and drawbacks related to traditional batch testing with p-values. In particular, the *testing by betting* paradigm has developed rapidly in recent years; it frames the level of evidence against the null as the wealth of a player in a betting game. Naturally, how much a player should bet is a key consideration, and should reflect their confidence in what they know about the next observed data point. Much current work uses standard betting strategies developed for adversarial problems. In the stochastic setting, where the observed data is independent and stationary, these approaches may be suboptimal. We propose a simple betting strategy, ***f**ollow the **l**eave-one-**o**ut **e**power* (FLOE), that exploits the i.i.d. nature of a data stream to estimate the e-power of our current test on the underlying distribution. We average not over the historical evidence collected so far, as generic online learning-based methods do, but instead over the equally-likely possible data sequences we might have seen based on shuffling the data. With an additional principled regularization technique, this simple drop-in method yields substantial improvements in testing performance on a variety of kernel two-sample and independence testing problems. Code is available at https://github.com/xunathan96/floe.

1 Introduction

Hypothesis testing is a fundamental area of statistics and is widely used by practitioners across science, medicine, and other areas as a means for verifying preconceived hypotheses. In the machine learning community, tests have been used for a variety of applications including causal inference (Peters et al., 2016; Heinze-Deml et al., 2018; Pfister et al., 2019; M. Liu et al., 2024), fairness (Romano et al., 2020; Ferrara et al., 2026), domain adaptation (Jang et al., 2022; Kübler et al., 2022), and detecting machine-synthesized content (Lopez-Paz and Oquab, 2017; Sutherland et al., 2017; F. Liu et al., 2020; Chen and J.-K. Wang, 2025; Li et al., 2025; Zhou et al., 2026). Historically, work in hypothesis testing has focused on designing *batch* tests, where the practitioner first collects all available data and then uses it to make a decision. These tests are typically based on the p-value of some test statistic of the underlying data distribution, asking whether the observed statistic value is very unlikely to occur due to chance under the null hypothesis. This testing paradigm is relatively simple, and often bears strong theoretical guarantees.

Batch tests, however, are notoriously prone to subtle misuse. As an example, suppose a clinical trial is verifying the efficacy of a new drug. This can be framed as a two-sample testing problem between the physical condition of users that took the drug to those that took a placebo. If the conditions are distributionally identical, the drug is ineffective; we call this the *null hypothesis*. The trial first collects data from 100 participants and runs a batch two-sample test on this data: this test is inconclusive, i.e., it fails to reject the null hypothesis. Believing the reason it failed due to insufficient data, the experimenters repeat the test a month later with 200 participants. This time the test does reject the null. Because batch tests are only valid at *predetermined* sample sizes, however, the act of testing on

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

new data – even if it is all independent! – invalidates the second test, as can be clearly seen if the experimenter repeats this process fifty times (Ioannidis, 2005; Simmons et al., 2011).

The alternative framework of *sequential* tests directly addresses the above concerns. Rather than collecting all available data at once, these tests are designed to be evaluated each time new data is observed. This approach is naturally more data efficient, as the test can be immediately stopped once enough "evidence" is collected. This evidence is measured by what are known as *e-values* (Vovk and R. Wang, 2021; Shafer, 2021; Ramdas and R. Wang, 2024) – nonnegative variables whose expectation under the null hypothesis is at most one – and are in many ways complementary to classical p-values. Perhaps the biggest advantage is that certain sequences of e-values are naturally closed under multiplication, giving an easy way to aggregate evidence across experiments; this forms the basis of *testing by betting* (Shafer, 2021).

The testing by betting paradigm establishes an equivalence between gathering evidence against the null hypothesis and the wealth of a player in a betting game. This game is constructed so that, if the null hypothesis is true, the game must be fair and the player's wealth is not expected to grow. How much wealth the player has is then directly a measure of how much evidence against the null hypothesis has been accumulated. Thus, a key aspect of this approach is designing betting strategies that maximize the growth rate of a player's wealth. Many of the betting strategies used in practice (Orabona and Pal, 2016; Podkopaev and Ramdas, 2023; Podkopaev et al., 2023; Chugg et al., 2023; Chen and J.-K. Wang, 2025), and in particular the kernel sequential tests of Shekhar and Ramdas (2023), are developed for adversarial online problems where the data is potentially chosen by an adversary. However, in the stochastic regime where data is i.i.d., these strategies fail to take advantage of this additional stationarity – and many tests in this area *do* assume the data is i.i.d. For the betting strategies that do (Waudby-Smith and Ramdas, 2020; Q. Wang et al., 2025; Pandeva et al., 2024), most are based on estimating the expected growth rate, or *e-power*, of the player's wealth.

Contributions. Early on in the betting game, our hypothesis remains uncertain and so e-power is low. As we observe more data, we narrow in on a prediction strategy that lets us more accurately assess our hypothesis (e.g. a stronger test statistic that differentiates between observed distributions); this improved prediction hopefully increases the e-power. For accurate betting, we should therefore make sure to estimate the e-power based on the *current* statistic, not as an average over *previous* statistics.

We do this by exploiting the exchangeability of the data stream to simulate many equally-likely data sequences; each of these sequences form leave-one-out estimates of the current test statistic, which we then use to estimate the e-power of the subsequent step. This approach leads to better e-power approximations, and thus serves as a much better guide for betting. We also prove that this betting strategy eventually recovers the asymptotically optimal bet amount.

Unfortunately, since early e-power estimates (based on scarce data) tend to be noisy, this can lead to overly aggressive early bets. We thus also propose a regularization strategy based on the uncertainty in our e-power estimates, which adaptively allows more aggressive betting in easier problems but is more conservative in highly uncertain problems. Adding this regularization makes our tests more powerful than prior betting strategies on a variety of real and synthetic testing problems.

2 Background

Two-Sample Testing. The batch two-sample testing problem involves determining whether two sets of observed samples $\mathbb{X}_m := (X_1, ..., X_m) \in \mathcal{X}^m$ and $\mathbb{Y}_n := (Y_1, ..., Y_n) \in \mathcal{X}^n$ are drawn from the same distribution. That is, we test the null hypothesis $\mathfrak{H}_0 : \mathbb{P} = \mathbb{Q}$ against the alternative hypothesis $\mathfrak{H}_1 : \mathbb{P} \neq \mathbb{Q}$. A hypothesis test Δ is a $\{0, 1\}$-valued function of the data that rejects the null hypothesis when $\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1$, and does not reject otherwise. Since the samples are random we can never conclude either hypothesis with certainty, and so statistical tests are designed to control the rate of false positives (type-I error) while minimizing the rate of false negatives (type-II error). We say a test is controlled with level $\alpha \in (0, 1)$ if its type-I error rate is at most α, i.e. $\Pr_{\mathfrak{H}_0}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1) \leq \alpha$. Similarly, a test has type-II error rate of $\beta \in (0, 1)$ if $\Pr_{\mathfrak{H}_1}(\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0) \leq \beta$. Among all level-$\alpha$ tests, the one with lowest type-II error β is considered the most powerful; $1 - \beta$ is called the test power. The probability subscripts used above indicate whether samples are drawn under the null or alternative hypotheses, and similar notation will be used for expectations; when none is written, we assume the distribution is under the alternative regime.

Maximum Mean Discrepancy. The *Maximum Mean Discrepancy* (MMD) (Gretton, Borgwardt, et al., 2012) is a measure of distance between distributions, and can be viewed as an integral probability metric over functions in a reproducing kernel Hilbert space (RKHS) $\mathcal{F}$ (Aronszajn, 1950):

$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{F}, \|f\| \leq 1} \mathbb{E}_{X,Y \sim \mathbb{P} \times \mathbb{Q}}[f(X) - f(Y)]. \tag{1}$$

$\mathcal{F}$ is uniquely characterized by a symmetric positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and if this kernel is characteristic (Sriperumbudur et al., 2010) then $\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) = 0$ iff $\mathbb{P} = \mathbb{Q}$. Gaussian kernels $k(x, y) = \exp(-\frac{1}{2}\|x - y\|^2/\gamma^2)$ with lengthscale $\gamma > 0$ are a common such example.

When the MMD is computed between two empirical distributions $\hat{\mathbb{P}}_m$ and $\hat{\mathbb{Q}}_n$, we get a biased but consistent estimate of the population value. If the sample sizes are equal, i.e. $m = n$, this estimate is a one-sample V-statistic on the paired samples $\mathbb{Z}_n = (Z_1, ..., Z_n)$ with $Z_i = (X_i, Y_i)$ and given by

$$\widehat{\mathrm{MMD}}^2_b(\mathbb{Z}_n) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h(Z_i, Z_j), \tag{2}$$

where we define $h(Z, Z') := k(X, X') - k(X, Y') - k(Y, X') + k(Y, Y')$ to be the symmetric V-statistic kernel function. This quadratic-time estimator is straightforward to compute and is a popular statistic for non-parametric tests. In particular, a batch two-sample test can be constructed with finite-sample validity by comparing the p-values of the MMD statistic under permutations of the data (Hemerik and Goeman, 2014). For any fixed sample size n, the test is controlled with level at most α. Like typical p-value tests, however, they are *not* robust to phenomena such as p-hacking; $\Pr_{\mathfrak{H}_0}(\max_{n>0} \Delta(\mathbb{Z}_n) = 1)$ is likely to greatly exceed α.

Sequential Tests. A desirable alternative to batch tests are *anytime valid tests*, which have type-I error controlled uniformly across sample sizes. These *sequential tests* and are particularly suited for online testing problems where data is continuously collected. We define a sequential test based on its stopping time $\tau := \inf\{t \geq 1 : \Delta(\mathbb{Z}_t) = 1\}$, the smallest sample size t such that a batch test on samples $Z_1, ..., Z_t$ rejects the null hypothesis. Here we require time-uniform control: $\Pr_{\mathfrak{H}_0}(\tau < \infty) \leq \alpha$ for some prescribed level $\alpha \in (0, 1)$. The sequential test τ is said to be *consistent* if $\Pr_{\mathfrak{H}_1}(\tau < \infty) = 1$, that is, the process rejects almost surely under the alternative hypothesis.

Recently, the *testing by betting* principle (Shafer, 2021) has been popularized as a blueprint for designing anytime valid sequential tests (Shekhar and Ramdas, 2023; Podkopaev and Ramdas, 2023; Pandeva et al., 2024). The idea is to define a betting game where at every step t we bet on predictions whose odds are sub-fair under the null hypothesis. Then the cumulative wealth at any time characterizes the the amount of evidence against the null up to that point. We formalize this game as follows: a player starts with initial wealth $\mathcal{W}_0 = 1$. At step t, they select a prediction function $\nu_t : \mathcal{Z} \to [-1, \infty)$ such that $\mathbb{E}_{\mathfrak{H}_0}[\nu_t(Z) \mid \mathscr{F}_{t-1}] \leq 0$, where $\mathscr{F}_{t-1} = \sigma(\{Z_1, ..., Z_{t-1}\})$ is the natural filtration, i.e. "everything seen before time t." They also select an $\mathscr{F}_{t-1}$-measurable bet $\lambda_t \in [0, 1)$. Then, after observing sample Z_t, their wealth is updated according to

$$\mathcal{W}_0 = 1, \qquad \mathcal{W}_t = \mathcal{W}_{t-1}\left(1 + \lambda_t \nu_t(Z_t)\right). \tag{3}$$

This wealth process forms a *test supermartingale* under the null (Ramdas and R. Wang, 2024). A supermartingale is a random process where for all t, $\mathbb{E}[\mathcal{W}_t \mid \mathscr{F}_{t-1}] \leq \mathcal{W}_{t-1}$ and $\mathbb{E}|\mathcal{W}_t| < \infty$; a test supermartingale adds the conditions that $\mathcal{W}_0 = 1$ and that $\mathcal{W}_t$ is almost surely nonnegative. A level-α sequential test for $\mathfrak{H}_0$ against $\mathfrak{H}_1$ is then defined by the stopping time

$$\tau = \inf\{t \geq 1 : \mathcal{W}_t \geq 1/\alpha\}, \tag{4}$$

and anytime validity follows from Ville's inequality, a time-uniform generalization of Markov's inequality for nonnegative supermartingales.

Prediction and Betting Strategies. The power of this test depends on the strategies for selecting the predictors ν_t and bets λ_t. In the kernel two-sample sequential test of Shekhar and Ramdas (2023), they propose using the prediction function $\nu_t(Z) = f_t(X) - f_t(Y)$ for some $\mathscr{F}_{t-1}$-measurable RKHS function f_t; this ensures $\mathbb{E}_{\mathfrak{H}_0}[f_t(X) - f_t(Y) \mid \mathscr{F}_{t-1}] = 0$. To get positive drift under the alternative, a natural choice for f_t is the empirical risk minimizer (ERM), the RKHS function that maximizes Equation (1) for empirical distributions $\hat{\mathbb{P}}_{t-1}$ and $\hat{\mathbb{Q}}_{t-1}$:

$$f(t) \propto \mathbb{E}_{X \sim \hat{\mathbb{P}}_{t-1}}[k(X, t)] - \mathbb{E}_{Y \sim \hat{\mathbb{Q}}_{t-1}}[k(Y, t)].$$

If the RKHS kernel k satisfies $k(x,x) \leq K$, then f_t is normalized to be in a ball of radius $1/(2\sqrt{K})$ so that $f_t(X) \in [-1/2, 1/2]$; this ensures the payoffs $\nu_t(Z)$ are in $(-1, 1)$. The bets can also be chosen in a similar manner, where we set λ_t to be the scalar in $(-1, 1)$ that maximizes the log-wealth at time $t-1$; this is known as GRAPA (Waudby-Smith and Ramdas, 2020) in the literature.

In the context of online learning, both of the aforementioned approaches are *follow-the-leader* (FTL) strategies because they use what would have been the optimal value in the previous round. Instead, Shekhar and Ramdas (2023) opt to use different online strategies, namely *online gradient descent* (OGD) (Zinkevich, 2003) for selecting the critic f_t, and *online Newton step* (ONS) (Hazan et al., 2007) for the bet λ_t. Their motivation for doing so is largely theoretical, as OGD and ONS are shown to respectively achieve $O(\sqrt{T})$ and $O(\log T)$ regret even in the adversarial online setting, whereas FTL strategies may not exhibit sub-linear regret.

Another popular online strategy with strong theoretical guarantees is the μ-weighted *universal portfolio* algorithm (UP) of Cover and Ordentlich (1996), which was originally motivated as an investment strategy for rebalancing m stocks. The authors show that this algorithm achieves minimax-optimal regret bounds when used with a $\text{Dirichlet}(1/2, ..., 1/2)$ prior. Since selecting scalar bets in $[0, 1]$ is equivalent to rebalancing a two-stock portfolio, this simplifies to a $\text{Beta}(1/2, 1/2)$ prior, and numerical integration becomes a practical approach to computing the bets.

The OGD, ONS, and UP algorithms all apply to the general online setting where data could be adversarially selected. In the stochastic regime, however, a test based on FTL payoff and betting strategies is still consistent even if its worst-case adversarial regret is poor. As such, it is not clear which strategy performs better in practice.

3 Proposed Method

3.1 Oracle Test

We start by constructing the oracle sequential two-sample test for witness functions in a bounded reproducing kernel Hilbert space, similar to Shekhar and Ramdas (2023). This test assumes complete knowledge of the underlying data distribution, and involves separate optimal strategies for learning the payoffs and bets. Henceforth, we assume bounded kernels with $\sup_x k(x,x) \leq K$ and $\inf_{x,y} k(x,y) \geq 0$.

The *witness function* f^* is the member in $\mathcal{F}$ that achieves the supremum in Equation (1); f^* "witnesses" the difference between $\mathbb{P}$ and $\mathbb{Q}$. For a version of (1) normalized to a radius-$1/\sqrt{2K}$ ball, which just yields the value $\text{MMD}(\mathbb{P}, \mathbb{Q})/\sqrt{2K}$, this maximizer is simply

$$f^* := \underset{f \in \mathcal{F}, \|f\| \leq \frac{1}{\sqrt{2K}}}{\arg\max} \mathbb{E}[f(X) - f(Y)] = \frac{\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}}{\sqrt{2K}\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|}, \tag{5}$$

where $\mu_{\mathbb{P}} := \mathbb{E}_{\mathbb{P}}[k(\cdot, X)]$ and $\mu_{\mathbb{Q}} := \mathbb{E}_{\mathbb{Q}}[k(\cdot, Y)]$ are known as the *mean embeddings* of $\mathbb{P}$ and $\mathbb{Q}$. We define the *witness predictor* as the function $\nu^*(x,y) := f^*(x) - f^*(y)$, and note that its values are necessarily bounded by one since

$$|\nu^*(x,y)| = |\langle f^*, k(\cdot,x) - k(\cdot,y)\rangle| \leq \|f^*\|\|k(\cdot,x) - k(\cdot,y)\| \leq \frac{1}{\sqrt{2K}}\sqrt{2K} = 1.$$

Its prediction at a paired sample $z = (x,y)$ is, with $h(z,z') = k(x,x') - k(x,y') - k(y,x') + k(y,y')$,

$$\nu^*(z) = f^*(x) - f^*(y) = \frac{\mathbb{E}_Z[h(z,Z)]}{\sqrt{2K\,\mathbb{E}[h(Z,Z')]}}. \tag{6}$$

We denote this predictor evaluated at time t on the sample Z_t to be $V_t^* := \nu^*(Z_t)$, and call this the *witness payoff* at t. One can confirm that $\mathbb{E}_{\mathfrak{H}_0}[V_t^* \mid \mathscr{F}_{t-1}] = 0$, either by noting that f^* does not depend on the data, or by identifying $\mathbb{E}_{\mathfrak{H}_0}[h(Z_t, Z')] = \text{MMD}^2(\mathbb{P}, \mathbb{P}) = 0$.

As for the bets, a natural approach is to, at every step t, pick the value in $[0, 1)$ that maximizes the expected growth in wealth. (Since $V_t^* \geq -1$, the restriction $\lambda < 1$ ensures that $1 + \lambda^* V_t^* > 0$ almost surely.) This strategy is essentially the *Kelly criterion* for portfolio optimization:

$$\lambda_t^* \in \underset{\lambda \in [0,1)}{\arg\max}\, \mathbb{E}[\log(1 + \lambda V_t^*) \mid \mathscr{F}_{t-1}]. \tag{7}$$

The $\mathbb{E}[\log(1 + \lambda V_t^*) \mid \mathscr{F}_{t-1}]$ term is the expected log-growth of the wealth at time t and is also known as the *e-power*. In the i.i.d. setting, the witness payoffs V_t^* are independent of the past, and so this reduces to an unconditional expectation; thus λ_t^* is the same for all time t and is simply denoted as λ^*. Putting everything together, we construct the oracle wealth process as

$$\mathcal{W}_0^* = 1, \qquad \mathcal{W}_t^* = \mathcal{W}_{t-1}^*(1 + \lambda^* V_t^*), \tag{8}$$

and the oracle sequential test as the stopping time $\tau := \inf\{t \geq 1 : \mathcal{W}_t^* \geq 1/\alpha\}$. This test exhibits time-uniform control with level α, and is known to be consistent under the alternative hypothesis.

3.2 Practical Test

Although the oracle test is informative, it is impossible to construct since we do not know the underlying data distribution. Instead, we define our prediction and betting strategies based on estimates of the witness function. To ensure such strategies are valid, we require them to be predictable, i.e. $\mathscr{F}_{t-1}$-measurable; this way the wealth process of (3) forms a test supermartingale.

One standard approach is to compute the empirical risk minimizer for (1) between the empirical distributions $\hat{\mathbb{P}}_{t-1}$ and $\hat{\mathbb{Q}}_{t-1}$. This gives us a predictable critic f_t, and consequently the predictor $\nu_t(z) := f_t(x) - f_t(y)$ and payoff $V_t := \nu_t(Z_t)$. We will stick with this approach. In the stochastic setting, there is little reason to choose otherwise; by symmetry, it cannot help to use another fixed weighting of each data point. Although Stein-style shrinkage estimators of the kernel mean exist (Muandet et al., 2016), the ERM estimator gives minimax-optimal estimates of the MMD (Tolstikhin et al., 2016). It is worth noting as a pointer for future work, however, that this choice does not necessarily maximize the e-power of the test.

If we decide to use the standard wealth process $\mathcal{W}_t = \mathcal{W}_{t-1}(1 + \lambda_t \nu_t(Z_t))$, it remains to choose a predictable bet λ_t. Ideally, we would choose λ_t according to the Kelly criterion (7) for payoffs based on the current critic at time t:

$$\lambda_t \in \operatorname*{arg\,max}_{\lambda \in [0,1)} \mathbb{E}\big[\log(1 + \lambda \nu_t(Z)) \mid \mathscr{F}_{t-1}\big]. \tag{9}$$

However, since our bet must be predictable and ν_t is based on *all* of the available data (i.e., $Z_1, ..., Z_{t-1}$) we have no held-out samples with which to estimate this e-power. The GRAPA betting strategy (Waudby-Smith and Ramdas, 2020), which stands for *growth rate adaptive to the particular alternative*, therefore estimates the e-power based on maximizing $\frac{1}{t-1}\sum_{\ell=1}^{t-1}\log(1+\lambda\nu_\ell(Z_\ell))$, which is equivalent to maximizing the wealth at time $t-1$ for a process using the fixed bet λ (a "follow-the-leader" strategy). In the limit as $\ell \to \infty$, we will typically have $\nu_\ell \to \nu^*$ and λ will converge to λ^* of (7).

As previewed in Section 1, however, the key insight of our work is that $\frac{1}{t-1}\sum_{\ell<t}\log(1+\lambda\nu_\ell(Z_\ell))$ is *not* necessarily a good estimate of the distribution of returns for $V_t = \nu_t(Z_t)$. Hopefully our predictor ν_t has gotten better as we observe more data, but GRAPA weights the historical return of ν_2 equally to how it weights ν_{t-2}! We suggest a simple alternative: while we do not have access to fresh data on which to estimate $\mathbb{E}\log(1+\lambda\nu_t(Z_t))$, in the stochastic setting we can use a *leave-one-out estimate* of the e-power. That is, rather than plugging in an average over the history of our actual critic functions, we can plug in an average over $t-1$ hypothetical sequences in which the last point we observed was Z_1, Z_2, ..., or Z_{t-1}. Because our data is exchangeable, this is an unbiased estimator for the e-power of ν_{t-1}, and hence a good estimate for the e-power of ν_t, the actually relevant quantity. We now detail how this is done.

Critic construction. Write $[t] = \{1, 2, \ldots, t\}$, so that $[t] \setminus \{\ell\}$ is the set of all indices up to t but excluding ℓ. Let $\hat{\mathbb{P}}_{[t]\setminus\{\ell\}}$ and $\hat{\mathbb{Q}}_{[t]\setminus\{\ell\}}$ be the empirical distributions based on Z_i for $i \in [t]\setminus\{\ell\}$. Define the *leave-one-out (loo) critic* $f_{[t]\setminus\{\ell\}}$ as the norm-$1/\sqrt{2K}$ maximizer of the MMD between empirical distributions $\hat{\mathbb{P}}_{[t]\setminus\{\ell\}}$ and $\hat{\mathbb{Q}}_{[t]\setminus\{\ell\}}$, and the corresponding predictor $\nu_{[t]\setminus\{\ell\}}(x, y) = f_{[t]\setminus\{\ell\}}(x) - f_{[t]\setminus\{\ell\}}(y)$. It is straightforward to show that this predictor evaluated on a paired sample $z = (x, y)$ is

$$\nu_{[t]\setminus\{\ell\}}(z) = \frac{\sum_{i\in[t]\setminus\{\ell\}} h(z, Z_i)}{\sqrt{2K \sum_{i\in[t]\setminus\{\ell\}} \sum_{j\in[t]\setminus\{\ell\}} h(Z_i, Z_j)}}. \tag{10}$$

When we evaluate a loo-predictor on its left-out sample, we obtain the *leave-one-out (loo) payoff* $V_\ell^{(t)} := \nu_{[t]\setminus\{\ell\}}(Z_\ell)$. It is worth mentioning that since of course $[t] \setminus \{t\} = [t-1]$, $\nu_{[t]\setminus\{t\}}$ is exactly the standard ERM predictor ν_t, and $V_t^{(t)}$ the ERM payoff V_t. Of the leave-one-out predictors, only $\nu_t = \nu_{[t]\setminus\{t\}}$ is $\mathscr{F}_{t-1}$-measurable, and hence only V_t is a valid payoff at time t.

Betting strategy. Recall that the oracle betting strategy aims to maximize the e-power of the wealth process at time t. Although the exact e-power is impossible to determine, we can construct a $\mathscr{F}_{t-1}$-measurable estimate of it from the loo-payoffs at time $t-1$. Our betting strategy is the maximizer of this estimate in the interval $[0, 1-\epsilon]$ for some small $\epsilon > 0$, i.e.

$$\lambda_t := \operatorname*{arg\,max}_{\lambda \in [0,1-\epsilon]} \frac{1}{t-1} \sum_{\ell=1}^{t-1} \log(1 + \lambda V_\ell^{(t-1)}). \tag{11}$$

We call this betting strategy *follow-the-loo-e-power* (FLOE). As the e-power estimate is strictly[1] concave with respect to $\lambda \in [0, 1)$, a unique maximizer exists. This one-dimensional optimization problem can be done via bisection to ε accuracy in $\log_2(\varepsilon)$ steps; five or ten steps are more than enough for practical purposes. Moreover, computing the e-power estimate at time t can be done in $O(t)$ cost as long as we saved the kernel matrix $H_{ij} := h(Z_i, Z_j)$ of the previous step. This betting strategy also converges to the Kelly criterion bet (Proposition 6).

Putting everything together, we use bets λ_t chosen according to (11) and ERM payoffs $V_t = \nu_t(Z_t)$, where ν_t is the norm-$1/\sqrt{2K}$ witness predictor between $\hat{\mathbb{P}}_{[t-1]}$ and $\hat{\mathbb{Q}}_{[t-1]}$. The test martingale is

$$\mathcal{W}_0 = 1, \qquad \mathcal{W}_t = \mathcal{W}_{t-1}(1 + \lambda_t V_t). \tag{12}$$

The corresponding sequential level-α test is the stopping time $\tau := \inf\{t \geq 1 : \mathcal{W}_t \geq 1/\alpha\}$, which under the null is finite with probability no greater than α.

Test Consistency. Although the test is guaranteed to be anytime valid with level-α, we would also like it to be consistent: it should eventually reject the null under the alternative. If not, one could simply construct a valid, but uninformative, sequential test like $\mathcal{W}_t = 1$. The following theoretical result confirms the wealth process of Equation (12) is consistent while maintaining finite-sample type-I control. Its proof is provided in Appendix A.1.

Theorem 1. *If $\tau := \min\{t \geq 1 : \mathcal{W}_t \geq 1/\alpha\}$ with $\mathcal{W}_t$ from (12) with a characteristic kernel, then*

$$\Pr_{\mathfrak{H}_0}(\tau < \infty) \leq \alpha \quad \textit{and} \quad \Pr_{\mathfrak{H}_1}(\tau < \infty) = 1.$$

3.3 Bet Regularization

The finite sample performance of any betting strategy is subtly tied to the uncertainty of predictor outputs. At sufficiently large t, we can expect the ERM predictors to be good approximations of the witness predictor, and so we can be confident their evaluated payoffs are accurate measures of evidence against the null hypothesis. On the other hand, we should be more skeptical of payoffs when t is small, as predictions are less accurate in this regime. An ideal betting strategy should be aggressive when we are confident in the payoffs and conservative when we are not. Figure 1 illustrates what happens when uncertainty is not encoded into the betting strategy. When t is small, the ERM payoffs are more frequently negative, which causes wealth to initially decrease. This phenomenon motivates using an uncertainty-based regularizer for the bets.

There are several ways we can quantify this uncertainty. Perhaps the most direct approach is with high probability bounds on either the predictor ν_t or its payoff V_t. But, while informative, these bounds alone do not justify particular betting values. We address this issue by defining a regularization paradigm based on the expected return $\mathbb{E}[\lambda V_t \mid \mathscr{F}_{t-1}]$, which directly connects to the choice of bets.

Definition 2. *Let $\alpha, \beta \in (0, 1)$, and $\mathbb{Z}_{t-1} = (Z_1, \ldots, Z_{t-1})$. We call $\Lambda_t^{(\alpha,\beta)}(\mathbb{Z}_{t-1}) \subseteq [0, 1)$ an (α, β)-regularized set of bets for the ERM predictor at time t if*

$$\Pr_{\mathbb{Z}_{t-1}} \left(\inf_{\lambda \in \Lambda_t^{(\alpha,\beta)}(\mathbb{Z}_{t-1})} \mathbb{E}[\lambda V_t \mid \mathscr{F}_{t-1}] > -\alpha \right) \geq 1 - \beta.$$

[1] Provided at least one payoff is non-zero, which is almost surely true for non-degenerate distributions.

An (α, β)-regularized set of bets is an interval of bets that guarantees an expected return of at least $-\alpha$ with probability no less than $1 - \beta$. This definition inherently accounts for the uncertainty of the prediction function, as we need to be confident the expected profit will not drop too low. It is worth pointing out that bets in $\Lambda_t^{(\alpha,\beta)}$ also indirectly regularize the e-power $\mathbb{E}[\log(1 + \lambda V_t) \mid \mathscr{F}_{t-1}]$ through Lemma 4, though this bound is certainly not sharp.

The next theorem (proved in Appendix A.2) constructs one possible (α, β)-regularized set. It is based on a martingale version of a Bennett-type concentration inequality, which uses variance information to provide sharper bounds. In this case, it uses the variance of the first-order projection of the V-statistic kernel, which we denote by $\zeta(\mathbb{P} \times \mathbb{Q}) := \mathrm{Var}_Z[\mathbb{E}_{Z'}[h(Z, Z')]]$. This term is essentially the asymptotic variance of the MMD estimator in Equation (2); estimating this is required for several notable kernel two-sample testing approaches (Sutherland et al., 2017; F. Liu et al., 2020; Sutherland and Deka, 2022; Wei et al., 2026). The typical biased estimate is

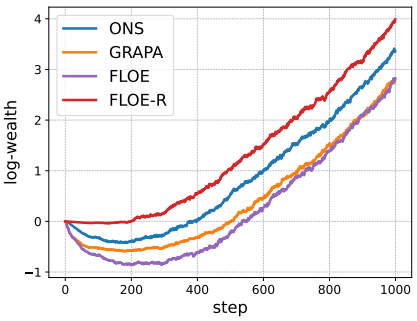


Figure 1: Wealth of several betting strategies (ONS, GRAPA, FLOE) for the x-Normal problem described in Section 4. Poor initial estimates often lead to negative returns, whereas a regularized betting strategy (FLOE-R) minimizes losses by betting more conservatively at the start.

$$\hat{\zeta}_b(\mathbb{Z}_t) := \frac{1}{t}\sum_{i=1}^{t}\left(\frac{1}{t}\sum_{j=1}^{t} h(Z_i, Z_j) - \frac{1}{t^2}\sum_{k=1}^{t}\sum_{j=1}^{t} h(Z_k, Z_j)\right)^2. \tag{13}$$

Theorem 3. *Let $\mathbb{Z}_t = (Z_1, ..., Z_t)$ be i.i.d. samples from $\mathbb{P} \times \mathbb{Q}$ and suppose we have tail bounds on the estimates in Equation (2) and Equation (13),*

$$\Pr\left(\widehat{\mathrm{MMD}}_b^2(\mathbb{Z}_t) - \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) \le \varepsilon_t^{(1)}\right) \ge 1 - \delta_1$$
$$\Pr\left(\zeta(\mathbb{P}\times\mathbb{Q}) - \hat{\zeta}_b(\mathbb{Z}_t) \le \varepsilon_t^{(2)}\right) \ge 1 - \delta_2,$$

with non-negative errors $\varepsilon_t^{(1)} = o(1)$ and $\varepsilon_t^{(2)} = o(1)$. For brevity, we omit arguments $\mathbb{P}$, $\mathbb{Q}$, $\mathbb{Z}_t$ and define the penalized estimates $\widetilde{\mathrm{MMD}}_b^2 := \widehat{\mathrm{MMD}}_b^2 - \varepsilon_t^{(1)}$ and $\widetilde{\zeta}_b := \hat{\zeta}_b + \varepsilon_t^{(2)}$. Then for all $\alpha \in (0, 1)$

$$\lambda_{t+1} \in \left[0, \min\left\{1, \frac{\alpha\sqrt{2K}\,\widehat{\mathrm{MMD}}_b}{\left[\frac{1}{4}\widetilde{\zeta}_b\phi_+^{-1}\left(\frac{16\log 1/\delta_0}{t\,\widetilde{\zeta}_b}\right) - \widetilde{\mathrm{MMD}}_b^2\right]_+}\right\}\right)$$

is a $(\alpha, \delta_0 + \delta_1 + \delta_2)$-regularized set of bets for the ERM predictor at time $t + 1$, where ϕ_+^{-1} is the positive inverse of $\phi(u) = (1 + u)\log(1 + u) - u$.

The first term in the denominator is expected to decrease with time, causing the bet interval to increase. We illustrate an example of this interval for a real wealth process in Figure C.3a of the appendix, and point the reader back to Figure 1 for its impact on the wealth. This regularization can also be applied to any betting strategy, not just for FLOE, as long as we use an ERM prediction strategy. We find it helps FLOE more than ONS, however, as the latter naturally takes smaller increments. Comparisons between the two are given in Appendix C.

Finally, we emphasize that this regularization does not increase asymptotic computational complexity. Computing its bound at time t can be done in $O(t)$ time using the saved kernel gram matrix H_{ij} from the previous step. Cumulatively over T steps, this amounts to $O(T^2)$ complexity in both compute and memory, the same as any kernel sequential test based on ERM or OGA prediction strategies. The complete kernel sequential test with this regularization procedure is outlined in Algorithm 1.

Practical Considerations. Theorem 3 gives a general way to regularize the bets, but it also depends on how tight the tail bounds on $\widehat{\mathrm{MMD}}_b^2(\mathbb{Z}_t)$ and $\hat{\zeta}_b(\mathbb{Z}_t)$ are. If these bounds are too loose, then the regularization set becomes too conservative, i.e. the interval remains small even after large time t.

Algorithm 1 Sequential test with the regularized FLOE betting strategy

Input: Data stream $Z_1, Z_2, ..., Z_T$ sampled independently from $\mathbb{P} \times \mathbb{Q}$.
Parameters: Test level $\eta \in (0, 1)$. Regularization strength $\alpha, \delta_0, \delta_1, \delta_2 \in (0, 1)$.
Output: $\{0, 1\}$ where 0 indicates 'do not reject' and 1 is 'reject' the null hypothesis $\mathfrak{H}_0 : \mathbb{P} = \mathbb{Q}$.

```
W_0 ← 1
for t = 1, 2, ..., T do
    Compute ERM payoff V_t = ν_t(Z_t)
    Compute bet λ_t according to Equation (11)
    Construct the interval Λ_t^{(α,δ_0+δ_1+δ_2)} described in Theorem 3
    λ_t ← Project λ_t onto Λ_t^{(α,δ_0+δ_1+δ_2)}
    W_t ← W_{t-1}(1 + λ_t V_t)
    if W_t ≥ 1/η then
        return 1
return 0
```

For the MMD estimator this is not a major issue, as there exist $O(t^{-1/2})$ error bounds with reasonable constants; for example, we know $\varepsilon_t^{(1)} = 2/t + 4\sqrt{[\log 1/\delta_1]/t}$ from a straightforward application of McDiarmid's inequality. The variance estimator $\hat{\zeta}_b$ is more problematic. Not only do we need a practical bound with small enough constants, but this bound should be on the square root of the estimator $\sqrt{\hat{\zeta}_b}$, because $\hat{\zeta}_b \phi_+^{-1}(1/\hat{\zeta}_b)$ simplifies to a leading term depending on $\sqrt{\hat{\zeta}_b}$ once we take the Taylor expansion. A peeling argument or entropy method could be used to do this, but neither would give practical bounds. Therefore, we omit penalizing $\hat{\zeta}_b$ in practice and simply use the unpenalized variance estimate when constructing the regularized set. Although this sacrifices any finite-sample guarantees, we expect it to still approximate the theoretical interval well for large enough t.

Another simplification we employ in practice is to replace $\phi_+^{-1}(u)$ with $h^{-1}(u) := u/3 + \sqrt{2u}$, which is the relaxation used when deriving Bernstein's inequality from Bennett's. This still ensures the correct confidence level of the regularization set, since $\phi_+^{-1}(u) \leq h^{-1}(u)$ for all $u \geq 0$.

4 Experiments

We evaluate our proposed sequential test across a number of synthetic and real-life datasets. Its performance with regularization (FLOE-R) and without (FLOE) is compared to a number of baselines including ONS (Hazan et al., 2007), GRAPA (Waudby-Smith and Ramdas, 2020), UP (Cover and Ordentlich, 1996; Orabona and Jun, 2024) with a Beta(1/2,1/2) prior, and the kernel sequential test of Shekhar and Ramdas (2023) (labelled SR) which uses OGD for the prediction strategy and ONS for the bets. All tests are controlled with level-0.05, and the regularization strength for FLOE-R is set to $\alpha = \delta_0 = \delta_1 = 0.1$ while omitting the variance penalty. We use Gaussian kernels for all problems, and select the lengthscale to optimize the performance of ONS; tuning for FLOE(-R) might lead to slightly larger improvements. Results are illustrated in Figure 2. Further experiment details and additional results on the Blobs and Higgs datasets are given in Section C.

Sinusoid. This problem, adapted from Sejdinovic et al. (2012), involves differentiating between sinusoidal and uniform distributions in a compact space. We sample $\mathbb{P}$ from the standard uniform on $[-\pi, \pi]^2$, and $\mathbb{Q}$ according to $\propto 1 + \sin(\omega x)\sin(\omega y)$ on the same support. Higher frequencies ω produce subtler departures from the uniform distribution, resulting in a harder two-sample problem. We use $\omega = 3.5$ and visualize its density in Figure C.3b.

x-Normal. We distinguish a 2D Gaussian mixture forming an "X" shape (Figure C.3c) from an isotropic Gaussian. Each mixture component of the X is constructed with a positive and negative correlation coefficient of equal magnitude; the lower the correlation, the harder it is to distinguish from a standard Gaussian. We use a correlation magnitude of 0.65. This problem is essentially the HDGM problem of F. Liu et al. (2020) in two dimensions.

CIFAR 10 vs 10.1. This problem involves discerning between images from the CIFAR-10 (Krizhevsky and Hinton, 2009) and CIFAR-10.1 (Torralba et al., 2008; Recht et al., 2018) test sets. The latter dataset was created entirely from new samples as an alternative test set for models

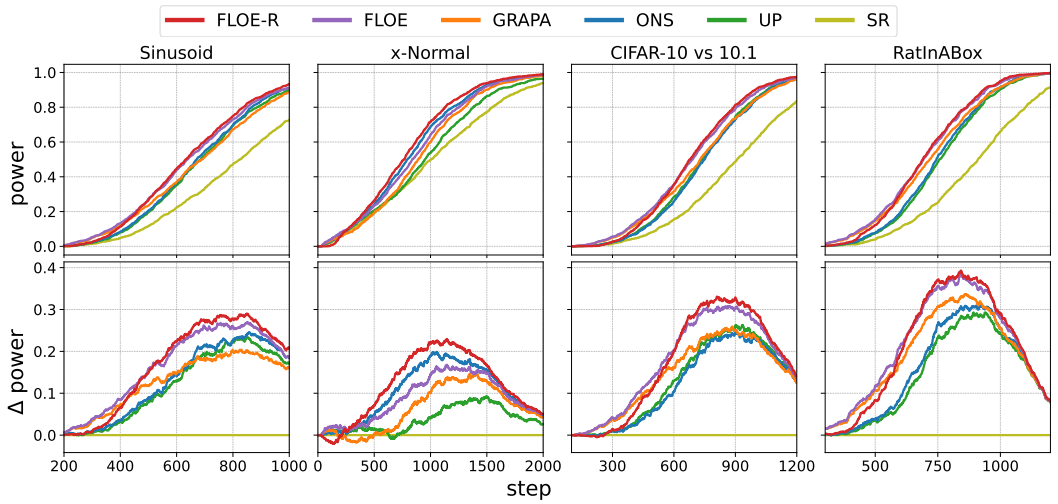


Figure 2: Mean test power on a variety of datasets; each test is run 1,000 times. The second row subtracts the performance of SR from each point for easier visualization.

trained on CIFAR-10, and models trained on it generally exhibit a drop in test accuracy. We compare the low-level features of each image obtained from a ResNet-50 model (He et al., 2016) pretrained on ImageNet.

RatInABox (George et al., 2024)**.** We simulate the neural activity of a rat moving inside slightly different boxed environments: one box has a protruding wall, while the other does not. This discrepancy in geometry should result in slightly different place cell (neural cells that are activated in the vicinity of certain locations) firing rates for each environment. We sample from 10 fixed place cells in the rat.

All of the considered tests are significantly better than the test of Shekhar and Ramdas (2023): this difference is due to using an ERM prediction strategy rather than OGD, since we find the ONS baseline substantially outperforms SR. As previously mentioned, their motivation for using OGD is theoretical, as there exist known regret bounds for OGD but not for ERM-based strategies.

Among the ERM-based tests, however, FLOE-R performs as well as, and in most cases better than, every competitor. Particularly, it improves finite sample performance against ONS and GRAPA by over 5% for some problems. Even without the regularizer FLOE is enough to beat these other strategies in all benchmarks except x-Normal, and is only slightly worse than FLOE-R in most cases.

In general, including the regularizer sacrifices early power for faster wealth growth, hence higher later power; the largest boost to power we observed is almost 10% in x-Normal. However, there are situations where including this regularization might be detrimental: if the problem is trivial enough to be solved in only a few steps. In this case, its performance hinges on how sharp the bound in Definition 2 is.

FLOE is also noticeably better than Universal Portfolio, which was shown by Cover and Ordentlich (1996) to achieve minimax optimal regret against the best in-hindsight constant betting strategy. The optimal strategy, however, is not constant, and so non-constant algorithms like FLOE may outperform them. Importantly, the UP betting strategy is invariant to the order of observations $Z_1, ..., Z_t$ (Orabona and Jun, 2024), which suggests it weights V_1 and V_t identically. However, as mentioned in our motivation for FLOE, we should avoid this; late payoffs are based on a far better understanding of the distribution than early ones.

Hyperparameter Sensitivity. We would like to confirm that our regularizer is robust to the choice of hyperparameters $\alpha, \delta_0, \delta_1 \in (0, 1)$. To compartmentalize the ablation we fix $\alpha = 0.1$ and vary $\delta := \delta_0 = \delta_1$ between two extremes: 0.01 and 0.5. Figure 3 illustrates the difference in power between FLOE at each level δ and ONS. In all problems only the extreme choice $\delta = 0.01$ results in lower power than ONS. The other regularization hyperparameter to consider is α. For this ablation we fix $\delta_0 = \delta_1 = 0.1$ and vary α between 0.05 and 0.5; these results are shown in Figure C.5 of the appendix. Overall, we find that our choice of parameters $\alpha = \delta_0 = \delta_1 = 0.1$ is a good middle ground for regularization strength, and that parameter choices around 0.1 and above all tend to be fairly

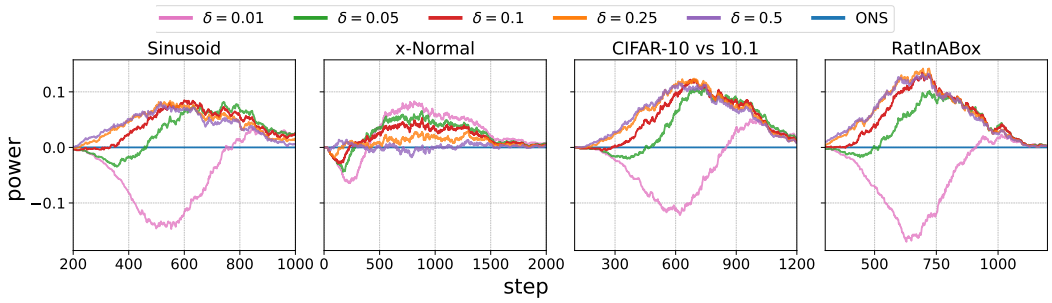


Figure 3: Power of FLOE-R relative to ONS at varying regularization levels δ and fixed $\alpha = 0.1$.

robust in terms of performance. When α and δ are very large (e.g. > 0.5), however, the regularizer impact is minimal.

FLOE for Sequential Kernel Independence Testing. Nothing about FLOE is specific to two-sample testing. The generality of the leave-one-out approach can easily be applied to learning critic functions for other problems, such as independence testing. We illustrate this by applying FLOE to the sequential kernel independence test (SKIT) of (Podkopaev et al., 2023). Similar to the kernel two-sample case, the leave-one-out payoffs for SKIT can be computed in cumulative quadratic time and memory. Exact expressions and implementation details are provided in Section B.

We evaluate FLOE on independence-testing versions of the Sinusoid, x-Normal, and RatInABox problems. For the former two, we sample X and Y jointly from the 2D distribution containing component-wise dependency. For RatInABox, we compare the firing rates of 6 place cells against 6 boundary vector cells. Detailed problem and test settings are given in Section B.1. The performance of FLOE against GRAPA, ONS, and Universal Portfolio with a Beta$(1/2, 1/2)$ prior is plotted in Figure 4. Results show FLOE outperforms these other baselines even without the use of regularization.

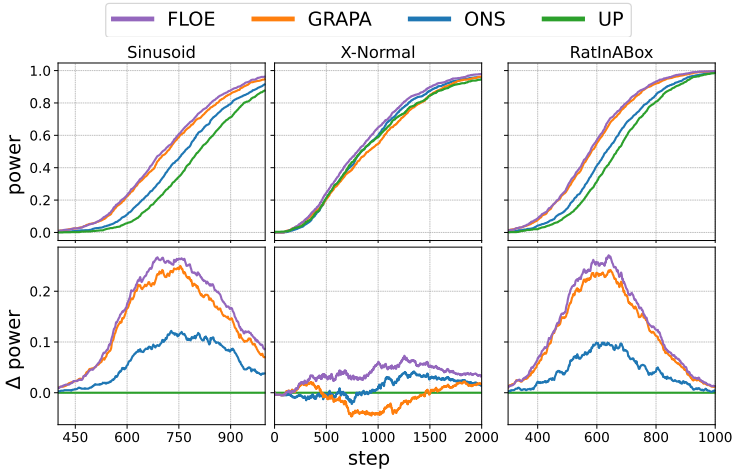


Figure 4: Mean test power on a variety of datasets; each test is run 1,000 times. The second row subtracts the performance of UP from each point for easier visualization.

5 Conclusion

Existing betting strategies in sequential anytime-valid testing default to a handful of classical online learning algorithms with nice theoretical adversarial-regime guarantees. In the stochastic setting, however, standard FTL algorithms may perform better in practice. Therefore, we framed the problem of selecting betting strategies in online testing in such a way as to very naturally motivate following a better estimate of the e-power. Our proposed method, FLOE, is natural, simple to implement, computationally efficient for kernel tests, and leads to substantial improvements in test power over prior work at very little computational cost.

Acknowledgments and Disclosure of Funding

The authors would like to thank Zheng He for productive discussions, and the anonymous reviewers for insights that improved the paper.

This work was enabled in part by the Natural Sciences and Engineering Research Council of Canada, the Canada CIFAR AI Chairs program, the Digital Research Alliance of Canada, and Calcul Québec. Feng Liu is supported by the ARC with grant numbers DP230101540, DE240101089 and DP260103996.

References

Nathan Aronszajn (1950). "Theory of Reproducing Kernels." *Transactions of the American Mathematical Society* 68, pages 337–404.

P. Baldi, P. Sadowski, and D. Whiteson (July 2014). "Searching for exotic particles in high-energy physics with deep learning." *Nature Communications* 5.1.

Peter L. Bartlett and Shahar Mendelson (Mar. 2003). "Rademacher and Gaussian complexities: risk bounds and structural results." *J. Mach. Learn. Res.* 3, pages 463–482.

Can Chen and Jun-Kun Wang (2025). "Online Detection of LLM-Generated Texts via Sequential Hypothesis Testing by Betting." *Forty-second International Conference on Machine Learning*.

Ben Chugg, Santiago Cortes-Gomez, Bryan Wilder, and Aaditya Ramdas (2023). "Auditing Fairness by Betting." *Thirty-seventh Conference on Neural Information Processing Systems*.

Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton (2015). "Fast Two-Sample Testing with Analytic Representations of Probability Measures." *Advances in Neural Information Processing Systems*.

Thomas M. Cover (1991). "Universal Portfolios." *Mathematical Finance* 1.1, pages 1–29.

Thomas M. Cover and Erik Ordentlich (1996). "Universal portfolios with side information." *IEEE Transactions on Information Theory* 42.2, pages 348–363.

Antonio Ferrara, Francesco Cozzi, Alan Perotti, André Panisson, and Francesco Bonchi (2026). "Size-adaptive Hypothesis Testing for Fairness." *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Tom M George, Mehul Rastogi, William de Cothi, Claudia Clopath, Kimberly Stachenfeld, and Caswell Barry (Feb. 2024). "RatInABox, a toolkit for modelling locomotion and neuronal activity in continuous environments." *eLife* 13. Edited by Mackenzie W Mathis, Michael J Frank, and Antonio Fernandez-Ruiz, e85274.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola (2012). "A Kernel Two-Sample Test." *Journal of Machine Learning Research* 13.25, pages 723–773.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola (2007). "A Kernel Statistical Test of Independence." *Advances in Neural Information Processing Systems*. Edited by J. Platt, D. Koller, Y. Singer, and S. Roweis. Volume 20. Curran Associates, Inc.

Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K. Sriperumbudur (2012). "Optimal kernel choice for large-scale two-sample tests." *Advances in Neural Information Processing Systems*. Volume 25.

Elad Hazan, Amit Agarwal, and Satyen Kale (Dec. 2007). "Logarithmic regret algorithms for online convex optimization." *Mach. Learn.* 69.2–3, pages 169–192.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (June 2016). "Deep Residual Learning for Image Recognition." *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. CVPR '16. Las Vegas, NV, USA: IEEE, pages 770–778.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen (2018). "Invariant Causal Prediction for Nonlinear Models." *Journal of Causal Inference* 6.2, page 20170016.

Jesse Hemerik and Jelle Goeman (Nov. 2014). "Exact testing with random permutations." *TEST* 27.4, pages 811–825.

Wassily Hoeffding (1961). *The Strong Law of Large Numbers for U-Statistics*. Technical report Institute of Statistics Mimeo Series 302. North Carolina State University, Department of Statistics.

John P. A. Ioannidis (2005). "Why Most Published Research Findings Are False." *PLoS Medicine* 2.8, e124.

Sooyong Jang, Sangdon Park, Insup Lee, and Osbert Bastani (July 2022). "Sequential Covariate Shift Detection Using Classifier Two-Sample Tests." *Proceedings of the 39th International Conference on*

Machine Learning. Volume 162. Proceedings of Machine Learning Research. PMLR, pages 9845–9880.

Alex Krizhevsky and Geoffrey Hinton (2009). *Learning multiple layers of features from tiny images*. Technical report. Toronto, Ontario: University of Toronto.

Jonas M. Kübler, Vincent Stimper, Simon Buchholz, Krikamol Muandet, and Bernhard Schölkopf (2022). "AutoML Two-Sample Test." *Advances in Neural Information Processing Systems*.

Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J. Su (2025). "A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules." *The Annals of Statistics* 53.1, pages 322–351.

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J. Sutherland (2020). "Learning deep kernels for non-parametric two-sample tests." *Proceedings of the 37th International Conference on Machine Learning*. ICML'20.

Mingzhou Liu, Xinwei Sun, Yu Qiao, and Yizhou Wang (2024). "Causal discovery via conditional independence testing with proxy variables." *Proceedings of the 41st International Conference on Machine Learning*. ICML'24. Vienna, Austria: JMLR.org.

David Lopez-Paz and Maxime Oquab (2017). "Revisiting Classifier Two-Sample Tests." *International Conference on Learning Representations*.

Krikamol Muandet, Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, and Bernhard Schölkopf (2016). "Kernel Mean Shrinkage Estimators." *Journal of Machine Learning Research* 17.48, pages 1–41.

Francesco Orabona and Kwang-Sung Jun (2024). "Tight Concentrations and Confidence Sequences From the Regret of Universal Portfolio." *IEEE Transactions on Information Theory* 70.1, pages 436–455.

Francesco Orabona and David Pal (2016). "Coin Betting and Parameter-Free Online Learning." *Advances in Neural Information Processing Systems*.

Teodora Pandeva, Patrick Forré, Aaditya Ramdas, and Shubhanshu Shekhar (2024). "Deep anytime-valid hypothesis testing." *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*. Volume 238. Proceedings of Machine Learning Research. PMLR, pages 622–630.

Adam Paszke et al. (2019). "PyTorch: an imperative style, high-performance deep learning library." *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen (2016). "Causal inference by using invariant prediction: identification and confidence intervals." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.5, pages 947–1012.

Niklas Pfister, Peter Bühlmann, and Jonas Peters (2019). "Invariant Causal Prediction for Sequential Data." *Journal of the American Statistical Association* 114.527, pages 1264–1276.

Aleksandr Podkopaev, Patrick Blöbaum, Shiva Kasiviswanathan, and Aaditya Ramdas (July 2023). "Sequential Kernelized Independence Testing." *Proceedings of the 40th International Conference on Machine Learning*. Volume 202. Proceedings of Machine Learning Research. PMLR, pages 27957–27993.

Aleksandr Podkopaev and Aaditya Ramdas (2023). "Sequential Predictive Two-Sample and Independence Testing." *Advances in Neural Information Processing Systems*.

Aaditya Ramdas and Ruodu Wang (July 2024). "Hypothesis Testing with E-values." *Foundations and Trends in Statistics* 1.1–2, pages 1–390.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar (2018). *Do CIFAR-10 Classifiers Generalize to CIFAR-10?* arXiv: 1806.00451 [cs.LG].

Yaniv Romano, Stephen Bates, and Emmanuel Candes (2020). "Achieving Equalized Odds by Resampling Sensitive Attributes." *Advances in Neural Information Processing Systems*. Volume 33, pages 361–371.

Dino Sejdinovic, Arthur Gretton, Bharath Sriperumbudur, and Kenji Fukumizu (2012). *Hypothesis testing using pairwise distances and associated kernels (with Appendix)*. arXiv: 1205.0411 [cs.LG].

Glenn Shafer (2021). "Testing by betting: A strategy for statistical and scientific communication." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 184.2, pp. 407–431.

Shubhanshu Shekhar and Aaditya Ramdas (2023). "Nonparametric Two-Sample Testing by Betting." *IEEE Transactions on Information Theory* 70.2, pages 1178–1203.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn (2011). "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22.11, pages 1359–1366.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet (2010). "Hilbert Space Embeddings and Metrics on Probability Measures." *Journal of Machine Learning Research* 11.50, pages 1517–1561.

Danica J. Sutherland and Namrata Deka (2022). *Unbiased estimators for the variance of MMD estimators*. arXiv: 1906.02104.

Danica J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton (2017). "Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy." *International Conference on Learning Representations*.

Ilya O Tolstikhin, Bharath K. Sriperumbudur, and Bernhard Schölkopf (2016). "Minimax Estimation of Maximum Mean Discrepancy with Radial Kernels." *Advances in Neural Information Processing Systems*.

Antonio Torralba, Rob Fergus, and William T. Freeman (2008). "80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.11, pages 1958–1970.

Václav Voráček and Francesco Orabona (2025). "STAR-Bets: Sequential TArget-Recalculating Bets for Tighter Confidence Intervals." *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Vladimir Vovk and Ruodu Wang (June 2021). "E-values: Calibration, combination and applications." *The Annals of Statistics* 49.3.

Qiuqi Wang, Ruodu Wang, and Johanna Ziegel (June 2025). "E-backtesting." *Management Science* 72.6, pages 4952–4973.

Ian Waudby-Smith and Aaditya Ramdas (2020). "Estimating means of bounded random variables by betting." *Journal of the Royal Statistical Society Series B: Statistical Methodology*.

Aaron Wei, Milad Jalali, and Danica J. Sutherland (2026). "Maximum Mean Discrepancy with Unequal Sample Sizes via Generalized U-Statistics." *Transactions on Machine Learning Research*. ISSN: 2835-8856.

Hongyi Zhou, Jin Zhu, Pingfan Su, Kai Ye, Ying Yang, Shakeel Gavioli-Akilagun, and Chengchun Shi (2026). "AdaDetectGPT: Adaptive Detection of LLM-Generated Text with Statistical Guarantees." *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Martin A. Zinkevich (2003). "Online Convex Programming and Generalized Infinitesimal Gradient Ascent." *International Conference on Machine Learning*.

A Proofs

Lemma 4 (Indirect E-Power Regularization)**.** *Any (α, β)-regularized set of bets $\Lambda^{(\alpha,\beta)} \subseteq [0, 1-\epsilon]$ for some $\epsilon \in (0,1)$ also regularizes the e-power according to*

$$\Pr_{\mathbb{Z}_{t-1}}\left(\inf_{\lambda\in\Lambda^{(\alpha,\beta)}} \mathbb{E}[\log(1+\lambda V_t) \mid \mathscr{F}_{t-1}] > \log(1-\alpha) - \frac{1}{2\epsilon^2}\operatorname{Var}[V_t \mid \mathscr{F}_{t-1}]\right) \geq 1-\beta.$$

Proof. The idea is to apply Taylor's theorem with a second-order Lagrange remainder term to $\log(1+\lambda V_t)$ about $\mathbb{E}[\lambda V_t \mid \mathscr{F}_{t-1}]$. For brevity, we will assume all expectations are conditioned on the filtration $\mathscr{F}_{t-1}$, even if not explicitly written.

The Taylor expanded expression for the e-power is

$$\mathbb{E}[\log(1+\lambda V_t)] = \log(1+\mathbb{E}[\lambda V_t]) - \mathbb{E}\left[\frac{(\lambda V_t - \mathbb{E}[\lambda V_t])^2}{2(1+\zeta)^2}\right],$$

for some random variable ζ between λV_t and $\mathbb{E}[\lambda V_t]$. Since $|\lambda V_t| \leq 1-\epsilon$, it follows that

$$\mathbb{E}[\log(1+\lambda V_t)] \geq \log(1+\mathbb{E}[\lambda V_t]) - \frac{\operatorname{Var}[V_t]}{2\epsilon^2}.$$

Finally, taking the infimum over the regularization set $\lambda \in \Lambda^{(\alpha,\beta)}$ and applying a union bound with $\Pr(\inf_{\lambda\in\Lambda^{(\alpha,\beta)}} \mathbb{E}[\lambda V_t] > -\alpha) \geq 1-\beta$ yields, with probability $1-\beta$,

$$\inf_{\lambda\in\Lambda^{(\alpha,\beta)}} \mathbb{E}[\log(1+\lambda V_t)] \geq \log(1-\alpha) - \frac{\operatorname{Var}[V_t]}{2\epsilon^2},$$

as desired. □

A.1 Test Consistency

Lemma 5. *Under the alternative hypothesis with a characteristic kernel, the leave-one-out predictor $\nu_\ell^{(t)} := \nu_{[t]\setminus\{\ell\}}$ defined in Equation (10) converges uniformly to the witness predictor ν^* defined in Equation (6) for any left-out sample $\ell \leq t$.*

$$\sup_{z\in\mathcal{Z}} \left|\nu_\ell^{(t)}(z) - \nu^*(z)\right| \xrightarrow{a.s.} 0 \qquad \forall \ell \leq t.$$

Proof. Since the RKHS $\mathcal{F}$ is a strong Glivenko-Cantelli class (Gretton, Borgwardt, et al., 2012; Bartlett and Mendelson, 2003), so too is $\mathcal{H} = \{h(z,\cdot) : z \in \mathcal{Z}\}$ as it is a linear composition of $\mathcal{F}$. This means $\mathcal{H}$ satisfies the uniform law of large numbers

$$\sup_{z\in\mathcal{Z}} \left|\frac{1}{t}\sum_{i=1}^{t} h(z, Z_i) - \mathbb{E}[h(z,Z)]\right| \xrightarrow{a.s.} 0$$

If we let $g_t := \frac{1}{t}\sum_{i=1}^t h(\cdot, Z_i)$ and $g := \mathbb{E}_Z[h(\cdot, Z)]$, this is equivalent to saying $\|g_t - g\|_\infty \xrightarrow{a.s.} 0$.

Let $\eta_t := \frac{1}{t^2}\sum_{i=1}^t\sum_{j=1}^t h(Z_i, Z_j)$ and $\eta := \mathbb{E}[h(Z,Z')]$. The strong law of large numbers for U-Statistics (Hoeffding, 1961) tells us $|\eta_t - \eta| \xrightarrow{a.s.} 0$. Also note that η is just the population MMD which, under the alternative hypothesis with a characteristic kernel, is positive.

It follows that $(g_t, \eta_t) \xrightarrow{a.s.} (g, \eta)$ in the product space, and so applying the continuous mapping theorem with $\Phi = (g,\eta) \mapsto g/\sqrt{2K\eta}$ yields

$$\left\|\frac{g_t}{\sqrt{2K\eta_t}} - \frac{g}{\sqrt{2K\eta}}\right\|_\infty \xrightarrow{a.s.} 0.$$

Notice that $\nu^* = g/\sqrt{2K\eta}$ and $\nu_\ell^{(t)} = g_t/\sqrt{2K\eta_t} + o(1)$ for any $\ell \leq t$, and so $\max_{\ell\leq t}\|\nu_\ell^{(t)} - \nu^*\|_\infty \xrightarrow{a.s.} 0$. □

Proposition 6. *Let $V_\ell^{(t)}$ be the loo payoffs and f^* be the witness function of Equation (5) with a characteristic kernel. Also, suppose the bets λ are in a bounded set $[0, 1-\epsilon]$ for some $\epsilon > 0$. Then, under the alternative hypothesis, we have*

$$\sup_{\lambda \in [0,1-\epsilon]} \left| \frac{1}{t} \sum_{\ell=1}^{t} \log(1 + \lambda V_\ell^{(t)}) - \mathbb{E}\left[\log\left(1 + \lambda\left(f^*(X) - f^*(Y)\right)\right)\right] \right| \xrightarrow{a.s.} 0$$

Proof. Let $\psi_\lambda(V) := \log(1 + \lambda V)$ and recall that we defined $V_\ell^* := \nu^*(Z_\ell)$. Since $\lambda < 1$ and $V_\ell^{(t)} \in [-1, 1]$ the function ψ_λ is L-Lipschitz for some $L > 0$. Define the sup-norm $\|\psi_\lambda(V)\|_\infty = \sup_\lambda |\psi_\lambda(V)|$ with respect to λ, and apply the decomposition

$$\left\| \frac{1}{t} \sum_{\ell=1}^{t} \psi_\lambda(V_\ell^{(t)}) - \mathbb{E}[\psi_\lambda(V_1^*)] \right\|_\infty \leq \left\| \frac{1}{t} \sum_{\ell=1}^{t} \psi_\lambda(V_\ell^{(t)}) - \psi_\lambda(V_\ell^*) \right\|_\infty + \left\| \frac{1}{t} \sum_{\ell=1}^{t} \psi_\lambda(V_\ell^*) - \mathbb{E}[\psi_\lambda(V_1^*)] \right\|_\infty.$$

For the first term, we can use the Lipschitz property to write

$$\begin{aligned}
\left\| \frac{1}{t} \sum_{\ell=1}^{t} \psi_\lambda(V_\ell^{(t)}) - \psi_\lambda(V_\ell^*) \right\|_\infty &\leq \frac{1}{t} \sum_{\ell=1}^{t} \left\| \psi_\lambda(V_\ell^{(t)}) - \psi_\lambda(V_\ell^*) \right\|_\infty \leq \frac{1}{t} \sum_{\ell=1}^{t} \sup_\lambda \left| \psi_\lambda(V_\ell^{(t)}) - \psi_\lambda(V_\ell^*) \right| \\
&\leq \frac{1}{t} \sum_{\ell=1}^{t} L \left| V_\ell^{(t)} - V_\ell^* \right| \\
&\leq L \max_{i \leq t} \left| V_i^{(t)} - V_i^* \right| \\
&\leq L \max_{i \leq t} \sup_{z \in \mathcal{Z}} \left| \nu_i^{(t)}(z) - \nu^*(z) \right|.
\end{aligned}$$

By Lemma 5 this must also converge a.s. to zero. The second term converges a.s. to zero by the uniform law of large numbers since $\{\psi_\lambda : \lambda \in [0, 1-\epsilon]\}$ is a strong Glivenko-Cantelli class when the domain is bounded in $[-1, 1]$. This is easily verified by noting that the Rademacher complexity of this function class is sub-linear. □

Theorem 1. *If $\tau := \min\{t \geq 1 : \mathcal{W}_t \geq 1/\alpha\}$ with $\mathcal{W}_t$ from (12) with a characteristic kernel, then*

$$\Pr_{\mathfrak{H}_0}(\tau < \infty) \leq \alpha \quad \textit{and} \quad \Pr_{\mathfrak{H}_1}(\tau < \infty) = 1.$$

Proof. It suffices to show that the e-power converges to some positive value, i.e.

$$\lim_{t \to \infty} \mathbb{E}[\log(1 + \lambda_t V_t^{(t)}) \mid \mathscr{F}_{t-1}] \overset{a.s.}{>} 0,$$

then consistency of the test directly follows Proposition 4.2 of Pandeva et al. (2024).

Since the bets are bounded away from 1 in an open interval, they must be a subset of $[0, 1-\epsilon]$ for some $\epsilon \in (0, 1)$. Note that λ_t exists and is unique on the set $[0, 1-\epsilon]$ by concavity of the log function. Thus, it follows from the uniform convergence result of Proposition 6 that

$$\lambda_t \xrightarrow{a.s.} \lambda^* := \operatorname*{arg\,max}_{\lambda \in [0,1-\epsilon]} \mathbb{E}[\log(1 + \lambda(f^*(X) - f^*(Y)))].$$

Since the data is i.i.d., the conditional e-power is equivalent to $\mathbb{E}_Z[\log(1 + \lambda_t \cdot v_t^{(t)}(Z))]$. We know from above that $\lambda_t \xrightarrow{a.s.} \lambda^*$, and Lemma 5 tells us $v_t^{(t)}(Z) \xrightarrow{a.s.} v^*(Z)$. As such, applying the continuous mapping theorem and dominated convergence theorem tells us

$$\mathbb{E}[\log(1 + \lambda_t V_t^{(t)}) \mid \mathscr{F}_{t-1}] \xrightarrow{a.s.} \mathbb{E}[\log(1 + \lambda^*(f^*(X) - f^*(Y)))].$$

This limit is just $\max_{\lambda \in [0,1-\epsilon]} \mathbb{E}[\log(1 + \lambda(f^*(X) - f^*(Y)))]$. Taking the Taylor expansion of $\log(1 + \lambda V)$ about 0 gives

$$\lambda \mathbb{E}[V^*] - \frac{\lambda^2 \mathbb{E}[(V^*)^2]}{2!} + \frac{\lambda^3 \mathbb{E}[(V^*)^3]}{3!} - \cdots$$

which, under the alternative hypothesis, is positive for some small enough $\lambda > 0$ since $\mathbb{E}[V^*] = \mathrm{MMD}(\mathbb{P}, \mathbb{Q}) > 0$. □

A.2 Bet Regularization

Lemma 7 (Bennett tail bound)**.** *If $X - \mathbb{E}[X] \leq b$ and $\sigma^2 := \mathrm{Var}[X] < \infty$ then*

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq \exp(\lambda^2\sigma^2 g(\lambda b)) \quad \forall \lambda > 0,$$

where $g = u \mapsto (e^u - u - 1)/u^2$.

Proof. Take the Taylor expansion of $e^{\lambda X}$ and use the fact $g(u) = \sum_{k=2}^{\infty} \frac{u^{k-2}}{k!}$

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq 1 + \mathbb{E}[\lambda^2(X - \mathbb{E}[X])^2 g(\lambda b)] \leq 1 + \lambda^2\sigma^2 g(\lambda b).$$

The result follows because $1 + x \leq e^x$. □

Proposition 8. *Let $\tilde{V}_{t+1} = \frac{1}{t}\sum_{i=1}^{t} h(Z_i, Z_{t+1})$ be the unnormalized ERM-payoff at step $t+1$ for i.i.d. data $Z_1, Z_2, ...,$ and let $\mathscr{F}_t = \sigma(\{Z_1, ..., Z_t\})$. Then for any error rate $\delta \in (0, 1)$*

$$\Pr\left(\mathbb{E}[\tilde{V}_{t+1} \mid \mathscr{F}_t] > \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) - \frac{\zeta}{4}\phi_+^{-1}\left(\frac{16\log 1/\delta}{t\zeta}\right)\right) \geq 1 - \delta,$$

where ϕ_+^{-1} is the positive inverse of $\phi(u) = (1+u)\log(1+u) - u$, and $\zeta = \mathrm{Var}_Z[\mathbb{E}_{\tilde{Z}}[h(Z, \tilde{Z})]]$.

Proof. Let $U_i := \mathbb{E}[\tilde{V}_{t+1} \mid \mathscr{F}_i]$ and note the sequence $(U_0, U_1, ..., U_t)$ forms a Doob martingale. First, the Martingale differences $U_i - U_{i-1}$ are bounded below by $-4/t$ since

$$\begin{aligned} U_i - U_{i-1} &= \frac{1}{t}\left(\mathbb{E}[h(Z_i, Z_{t+1}) \mid \mathscr{F}_i] - \mathbb{E}[h(Z_i, Z_{t+1}) \mid \mathscr{F}_{i-1}]\right) \\ &= \frac{1}{t}\left(\mathbb{E}[h(z_i, Z)] - \mathbb{E}[h(Z, Z')]\right) \geq -\frac{4}{t}. \end{aligned}$$

Also, its conditional variance is

$$\mathrm{Var}[U_i - U_{i-1} \mid \mathscr{F}_{i-1}] = \frac{1}{t^2}\mathrm{Var}_Z[\mathbb{E}_{\tilde{Z}}[h(Z, \tilde{Z})]] = \frac{\zeta}{t^2}.$$

Hence the conditional differences satisfy a Bennett tail bound, a fact we now use to derive the result.

For any $\epsilon \in \mathbb{R}$, we have the following Chernoff bound on the lower tail:

$$\Pr(U_0 - U_t \geq \epsilon) \leq \inf_{\lambda > 0} e^{-\lambda\epsilon}\mathbb{E}[e^{\lambda(U_0 - U_t)}].$$

The expectation in the bound can be unrolled into a composition of martingale differences, each of which is bounded using Lemma 7 with $b := 4/t$ and $\sigma^2 := \zeta/t^2$ to produce

$$\begin{aligned} \mathbb{E}[e^{\lambda(U_0-U_t)}] &= \mathbb{E}[e^{\lambda\sum_{i=1}^{t} U_{i-1}-U_i}] \\ &= \mathbb{E}[e^{\lambda\sum_{i=1}^{t-1} U_{i-1}-U_i}\mathbb{E}[e^{\lambda(U_{t-1}-U_t)} \mid \mathscr{F}_{t-1}]] \\ &\leq \exp\left(\lambda^2 t\sigma^2 g(\lambda b)\right). \end{aligned}$$

Plugging this back into our Chernoff bound yields

$$\Pr(U_0 - U_t \geq \epsilon) \leq \inf_{\lambda > 0} \exp(-\lambda\epsilon + \lambda^2 t\sigma^2 g(\lambda b)) = \exp\left(-\frac{t\sigma^2}{b^2}\phi\left(\frac{b\epsilon}{t\sigma^2}\right)\right),$$

where the minimizer can be derived by noticing that the expression in the exponent is convex in λ; it is $\lambda^* = \frac{1}{b}\log(1 + \frac{b\epsilon}{t\sigma^2})$. This satisfies $\lambda^* > 0$ for any $\epsilon > 0$. Finally, inverting this tail bound, noticing $U_0 = \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q})$ and substituting in the values for b and σ^2 gives us the desired result. □

Theorem 3. *Let $\mathbb{Z}_t = (Z_1, ..., Z_t)$ be i.i.d. samples from $\mathbb{P} \times \mathbb{Q}$ and suppose we have tail bounds on the estimates in Equation (2) and Equation (13),*

$$\Pr\left(\widehat{\mathrm{MMD}}_b^2(\mathbb{Z}_t) - \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) \leq \varepsilon_t^{(1)}\right) \geq 1 - \delta_1$$

$$\Pr\left(\zeta(\mathbb{P}\times\mathbb{Q}) - \hat{\zeta}_b(\mathbb{Z}_t) \leq \varepsilon_t^{(2)}\right) \geq 1 - \delta_2,$$

with non-negative errors $\varepsilon_t^{(1)} = o(1)$ *and* $\varepsilon_t^{(2)} = o(1)$. *For brevity, we omit arguments* $\mathbb{P}$, $\mathbb{Q}$, $\mathbb{Z}_t$ *and define the penalized estimates* $\widetilde{\mathrm{MMD}}_b^2 := \widehat{\mathrm{MMD}}_b^2 - \varepsilon_t^{(1)}$ *and* $\widetilde{\zeta}_b := \hat{\zeta}_b + \varepsilon_t^{(2)}$. *Then for all* $\alpha \in (0,1)$

$$\lambda_{t+1} \in \left[0, \min\left\{1, \frac{\alpha\sqrt{2K}\,\widehat{\mathrm{MMD}}_b}{\left[\frac{1}{4}\widetilde{\zeta}_b\phi_+^{-1}\left(\frac{16\log 1/\delta_0}{t\,\widetilde{\zeta}_b}\right) - \widetilde{\mathrm{MMD}}_b^2\right]_+}\right\}\right)$$

is a $(\alpha, \delta_0+\delta_1+\delta_2)$*-regularized set of bets for the ERM predictor at time* $t+1$, *where* ϕ_+^{-1} *is the positive inverse of* $\phi(u) = (1+u)\log(1+u) - u$.

Proof. Let $\Lambda(\mathbb{Z}_t)$ be the proposed regularized bet set. We want to show that with confidence $1-(\delta_0+\delta_1+\delta_2)$

$$\inf_{\lambda\in\Lambda(\mathbb{Z}_t)} \mathbb{E}[\alpha + \lambda V_{t+1} \mid \mathscr{F}_t] > 0.$$

To make things more readable, we omit dependencies on $\mathbb{P}\times\mathbb{Q}$ and $\mathbb{Z}_t$, and we let $\eta_t := \widehat{\mathrm{MMD}}_b^2 = \frac{1}{t^2}\sum_{i=1}^t\sum_{j=1}^t h(Z_i, Z_j)$ be the biased MMD estimator.

Notice that for all $\lambda \in \Lambda(\mathbb{Z}_t)$

$$-\frac{\alpha\sqrt{2K\eta_t}}{\lambda} < \widetilde{\mathrm{MMD}}_b^2 - \frac{\widetilde{\zeta}_b}{4}\phi_+^{-1}\left(\frac{16\log 1/\delta_0}{\widetilde{\zeta}_b t}\right),$$

which is easily verified by solving this bound for λ. The term on the right is is increasing with $\widetilde{\mathrm{MMD}}_b^2$ and decreasing with $\widetilde{\zeta}_b$, so by taking a union bound with the events $\{\widetilde{\mathrm{MMD}}_b^2 \le \mathrm{MMD}^2\}$ and $\{\widetilde{\zeta}_b \ge \zeta\}$ we get with confidence $1-(\delta_1+\delta_2)$ that

$$-\frac{\alpha\sqrt{2K\eta_t}}{\lambda} < \mathrm{MMD}^2 - \frac{\zeta}{4}\phi_+^{-1}\left(\frac{16\log 1/\delta_0}{\zeta t}\right) \quad \forall\lambda\in\Lambda(\mathbb{Z}_t).$$

By Proposition 8 the term on the right is upper bounded by $\mathbb{E}[\tilde{V}_{t+1} \mid \mathscr{F}_t]$ with probability $1-\delta_0$, and so a union bound with this event gives us with confidence $1-(\delta_0+\delta_1+\delta_2)$ that

$$-\frac{\alpha\sqrt{2K\eta_t}}{\lambda} < \mathbb{E}[\tilde{V}_{t+1} \mid \mathscr{F}_t] \quad \forall\lambda\in\Lambda(\mathbb{Z}_t).$$

Rearranging terms gives us the desired result. □

B FLOE for Sequential Kernel Independence Testing

We give an overview of the sequential kernel independence test, and refer readers to (Podkopaev et al., 2023) for a more detailed explanation.

Oracle Critic. Similar to the kernel two-sample test, the oracle function for the independence test is the supremum over bounded Hilbert-Schmidt operators from RKHS $\mathcal{H}_X$ to RKHS $\mathcal{H}_Y$.

Let k and l be the kernels associated with $\mathcal{H}_X$ and $\mathcal{H}_Y$, respectively, and let $\phi(X) := k(\cdot, X)$ and $\varphi(Y) := l(\cdot, Y)$ be their canonical feature maps. Also suppose $\sup_{x,x'} k(x, x') \leq K_X$ and $\sup_{y,y'} l(y, y') \leq K_Y$. The Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2007) is

$$\sqrt{\text{HSIC}(X, Y)} = \sup_{\|F\| \leq 1} \langle F, C_{YX} \rangle_{\text{HS}(\mathcal{H}_X, \mathcal{H}_Y)}$$

where $C_{YX} = \mathbb{E}[\varphi(Y) \otimes \phi(X)]$ is the covariance operator defined with $[\varphi(Y) \otimes \phi(X)](f) = \langle f, \phi(X) \rangle \varphi(Y)$. Thus, the oracle critic is the normalized covariance operator $C_{YX}/\|C_{YX}\|$.

Suppose we have $2n$ i.i.d. samples $(X_1, Y_1), ..., (X_{2n}, Y_{2n})$. For convenience, we rename the latter n samples by $(X'_1, ..., X'_n)$ so that we now have two sets of samples $(X_i, Y_i)_{i=1..n}$ and $(X'_i, Y'_i)_{i=1..n}$.

A test martingale constructed with this oracle critic has payoffs given by

$$V_t = \frac{1}{2K_X K_Y} \left\langle \frac{C_{YX}}{\|C_{YX}\|}, (\varphi(Y_t) - \varphi(Y'_t)) \otimes (\phi(X_t) - \phi(X'_t)) \right\rangle,$$

which is easily verified by checking $|V_t| \leq 1$ and $\mathbb{E}_{\mathfrak{H}_0}[V_t \mid \mathscr{F}_{t-1}] = 0$.

Loo Payoffs. Let $\hat{C}_{YX}^{[t]\setminus\{\ell\}}$ be an estimate of the covariance operator from all samples indexed by $1, ..., t$ excluding the samples (X_ℓ, Y_ℓ) and (X'_ℓ, Y'_ℓ). For brevity, we will denote $\hat{\mathbb{E}}^t_{i\neq\ell}[a_i] = \frac{1}{2(t-1)} \sum_{i=1; i\neq\ell}^t a_i$. Then,

$$\begin{aligned}\hat{C}_{YX}^{[t]\setminus\{\ell\}} = \hat{\mathbb{E}}^t_{i\neq\ell}[\varphi(Y_i) \otimes \phi(X_i) + \varphi(Y'_i) \otimes \phi(X'_i)] \\ - \hat{\mathbb{E}}^t_{k\neq\ell}[\varphi(Y_k) + \varphi(Y'_k)] \otimes \hat{\mathbb{E}}^t_{j\neq\ell}[\phi(X_j) + \phi(X'_j)].\end{aligned}$$

It follows that the loo-payoffs $V_\ell^{(t)}$, which is the loo-critic $\hat{C}_{YX}^{[t]\setminus\{\ell\}}$ evaluated at its held-out samples (X_ℓ, Y_ℓ) and (X'_ℓ, Y'_ℓ), is

$$\begin{aligned}V_\ell^{(t)} &= \frac{1}{2K_X K_Y \|\hat{C}_{YX}^{[t]\setminus\{\ell\}}\|} \left\langle \hat{C}_{YX}^{[t]\setminus\{\ell\}}, (\varphi(Y_\ell) - \varphi(Y'_\ell)) \otimes (\phi(X_\ell) - \phi(X'_\ell)) \right\rangle \\ &= \frac{1}{2K_X K_Y \|\hat{C}_{YX}^{[t]\setminus\{\ell\}}\|} \Big\{ \hat{\mathbb{E}}^t_{i\neq\ell} \left[\big(k(X_i, X_\ell) - k(X_i, X'_\ell)\big) \big(l(Y_i, Y_\ell) - l(Y_i, Y'_\ell)\big) \right] \\ &\qquad + \hat{\mathbb{E}}^t_{i\neq\ell} \left[\big(k(X'_i, X_\ell) - k(X'_i, X'_\ell)\big) \big(l(Y'_i, Y_\ell) - l(Y'_i, Y'_\ell)\big) \right] \\ &\qquad - \hat{\mathbb{E}}^t_{i\neq\ell} \left[\big(k(X_i, X_\ell) - k(X_i, X'_\ell) + k(X'_i, X_\ell) - k(X'_i, X'_\ell)\big) \right] \\ &\qquad \times \hat{\mathbb{E}}^t_{i\neq\ell} \left[\big(l(Y_i, Y_\ell) - l(Y_i, Y'_\ell) + l(Y'_i, Y_\ell) - l(Y'_i, Y'_\ell)\big) \right] \Big\}.\end{aligned}$$

Note that $\|\hat{C}_{YX}^{[t]\setminus\{\ell\}}\|$ is exactly the root biased HSIC estimator from samples $1..t$ excluding ℓ.

Implementation Details. Similarly to the kernel two-sample test of Section 3.2, this independence test can be done in linear time per step as long as we store the kernel matrices $K_{ij} = k(X_i, X_j)$ and $L_{ij} = l(Y_i, Y_j)$ of the previous step. Cumulatively, the test is quadratic in both compute and memory. Maintaining the latter is possible by subtracting the sums where the left-out index ℓ is included. For example, in $\|\hat{C}_{YX}^{[t]\setminus\{\ell\}}\|$ we can simply compute the HSIC estimate and then subtract the terms that include index ℓ in them. Since HSIC is composed of nested sums, the principle of inclusion-exclusion is used.

B.1 Experiments

We compare the performance of FLOE against GRAPA (Waudby-Smith and Ramdas, 2020), ONS (Hazan et al., 2007), and universal portfolio with Beta(1/2,1/2) prior (UP) (Cover, 1991; Orabona and Jun, 2024) betting strategies, and report the results in Figure 4. We use Gaussian kernels for both $\mathcal{H}_X$ and $\mathcal{H}_Y$ and select the bandwidths to optimize the performance of ONS. We find FLOE performs the best on our considered datasets.

Sinusoid (Sejdinovic et al., 2012)**.** We test for dependencies between X and Y sampled jointly from $\propto 1 + \sin(\omega x)\sin(\omega y)$ on the support $[-\pi, \pi]^2$. Higher frequencies ω produce subtler dependencies, resulting in a harder independence test. We use $\omega = 5$.

x-Normal. We jointly sample X and Y from the xNormal distribution used in our two-sample testing experiments. Smaller correlation coefficients result in harder-to-detect dependencies. We use a coefficient of ± 0.5 for its normal components.

RatInABox (George et al., 2024)**.** We compare the firing rates of place cells (neurons that are activated in the vicinity of certain locations) to those of boundary vector cells (neurons that fire when the rat is at a specific distance from a wall). These cell activation locations are not disjoint, and so there should be dependencies between their firing rates. We use a standard 1x1 box environment with a protruding wall, and sample from 6 place cells and 6 boundary vector cells.

C Additional Experiments and Details

Blobs. We perform an additional evaluation on the Blobs dataset (Chwialkowski et al., 2015; Gretton, Sejdinovic, et al., 2012), which is a benchmark synthetic problem composed of a 3x3 uniform grid of bivariate Gaussians. These Gaussian 'blobs' are isotropic in $\mathbb{P}$ and anisotropic in $\mathbb{Q}$ with each blob randomly assigned an eigenvalue ratio and rotation angle. The difficulty lies in the fact the discrepancy between distributions are encoded at a much smaller frequency than the overall data. Results are plotted in Figure C.1 for the average power over 1,000 tests. FLOE-R performs better than FLOE at moderate to large sample sizes, and both are better than ONS and GRAPA.

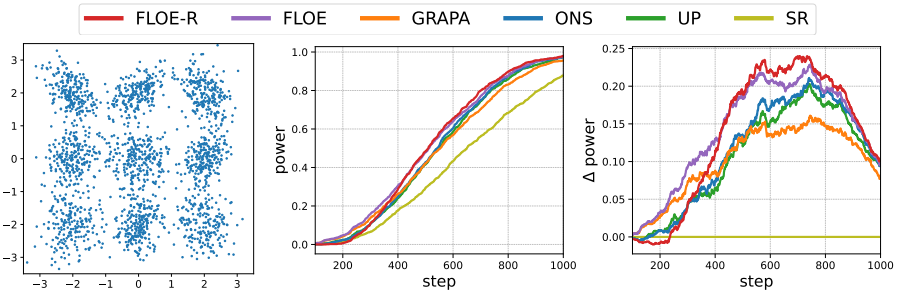


Figure C.1: Blob dataset: On the left are samples from $\mathbb{Q}$. In the middle is the mean power at time t. On the right is the power relative to SR.

Higgs (Baldi et al., 2014)**.** We aim to identify if a Higgs boson has been created by comparing the signals of processes that produce it to background processes that do not. In particular, we focus on the φ-momenta of four particle jets, which hold very little discriminating information for whether a Higgs boson was birthed.

Visualized Bet Interval. It will be informative to see how the bet interval of Theorem 3 evolves with time. Asymptotically this interval is expected to be $[0, 1)$ as both the MMD and variance estimators converge to positive values while the first term in the denominator decreases to zero. That said, this interval is not necessarily monotonic, though we expect it to roughly be the case on average. We plot this interval along with the actual bet values used by FLOE (without regularization) in Figure C.3a. This figure makes it clear why the regularizer is necessary; initial bets are often far more aggressive than they should be.

CIFAR-10 vs 10.1 Features. The featurizer we use is a ResNet-50 model (He et al., 2016) pretrained on ImageNet and provided by the torchvision library (Paszke et al., 2019). In particular we use the

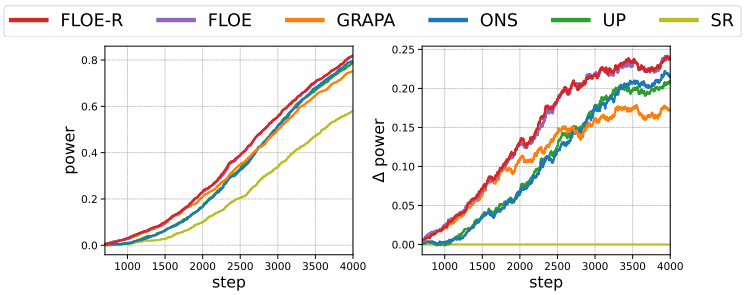
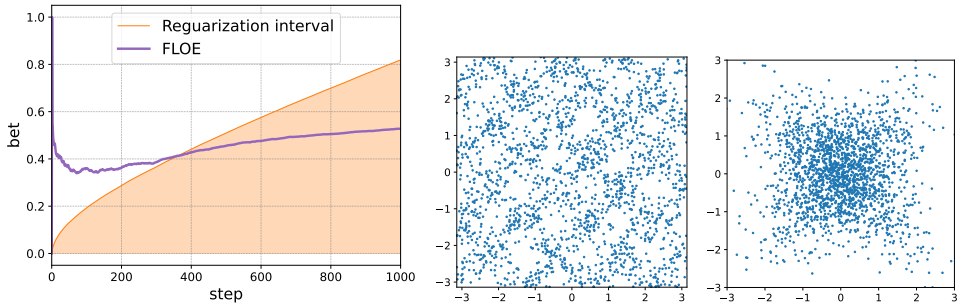


Figure C.2: Higgs dataset. The plot on the right subtracts the power of SR from each baseline for easier visualization.



(a) The average regularization interval used by FLOE-R for the xNormal dataset.

(b) Sinusoid

(c) X-Normal

Figure C.3: (a) Visualization of FLOE's bet value and the regularization interval used for FLOE-R. (b)-(c) Samples from the toy datasets.

weights corresponding to "IMAGENET1K_V2", and make a few modifications to the architecture: the maxpool layer is removed, and the last fully-connected layer is replaced with a linear layer that projects the number of features down to 100.

Regularizing ONS. Our regularizer not only applies to FLOE, but to any general betting strategy as long as the prediction strategy is ERM. Thus, we could also regularize ONS this way. It is worth mentioning that ONS is a "parameter-free" algorithm, meaning it will converge no matter the choice of learning rate. However, in finite samples these hyperparameters still play a factor. Our experiments use essentially the same hyperparameters as those of (Shekhar and Ramdas, 2023), with the only difference being the choice of the inverse step size γ. Typically γ is set to be $\gamma = 2(b - \log(1+b))/b^2$ where b is an upper bound on $\langle \nabla(-\log(1+\lambda V)), \lambda - \lambda' \rangle$. This ensures suitable regret bounds. For the test of (Shekhar and Ramdas, 2023) they clip the bets to $[-1/2, 1/2]$, and for this interval $b = 2$ which leads to their choice of γ. In our case the bets are in $[0, 1)$ which, in the limit as $V_t \to V_\infty > 0$, gives a bound $b = 1$ and consequently $\gamma = 2(1 - \log(2))$. As for the initial Hessian approximation a_0, we do the same as them and set it to 1.0.

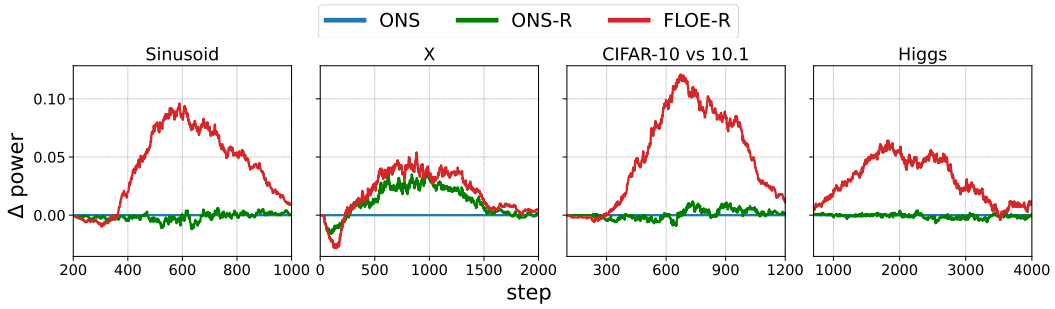


Figure C.4: Test power of regularization strategies relative to standard ONS.

Figure C.4 plots power of ONS with our regularizer (ONS-R) relative to the performance of ONS without the regularizer. The relative power of FLOE-R is also included for reference. The regularization strength is set with $\delta_0 = \delta_1 = 0.1$ and $\alpha = 0.2$, which is slightly weaker than what we used for our other experiments. This is because ONS seems to perform better with a weaker regularizer than FLOE, as its step sizes are naturally smaller. These smaller step sizes could also explain why regularization appears less impactful. For Sinusoid, Higgs, and CIFAR, there is essentially no difference between ONS and ONS-R; ONS-R does seem to help in the X-dataset.

Hyperparameter Sensitivity. We assess the sensitivity of the regularization parameter α for the practical $(\alpha, \delta_0 + \delta_1)$-regularizer constructed based on Theorem 3. We set $\delta_0 = \delta_1 = 0.1$ and vary α between 0.05 and 0.5. The results are presented in Figure C.5 as the power relative to ONS. In general, it seems α is slightly more sensitive than δ_0, δ_1 for small extremes, but still robust overall when considering values around 0.1 and larger. Our recommendation is to keep $\alpha = \delta_0 = \delta_1 = 0.1$ and if necessary, tweak the error rates δ_0 and δ_1.

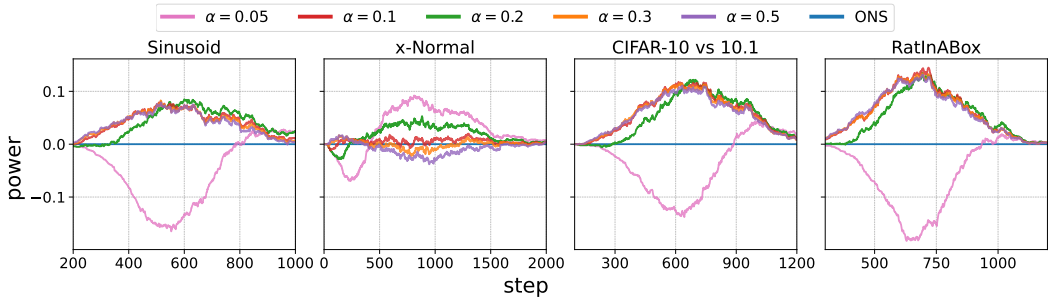


Figure C.5: Power of FLOE-R relative to ONS at varying regularization strengths α and fixed $\delta = 0.1$.

For completeness, we also include in Figure C.6 the raw power plots corresponding to the relative power plots of Figure 3.

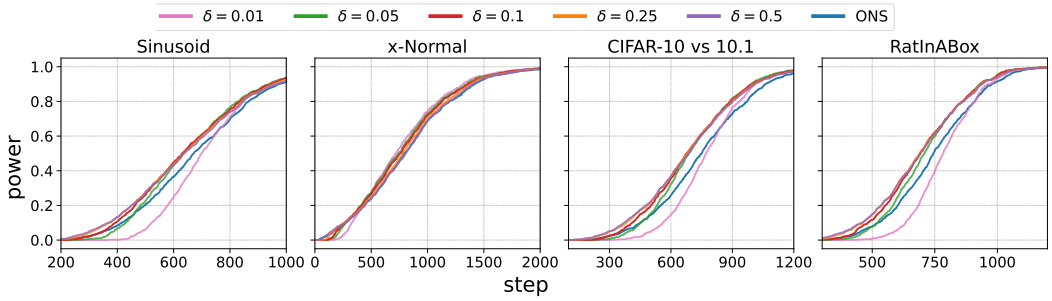


Figure C.6: Power of FLOE-R relative to ONS at varying regularization strengths δ and fixed $\alpha = 0.1$.

Kernel Bandwidth. All our experiments use the Gaussian kernel $k(x, y) = \exp(-\frac{1}{2}\|x - y\|^2/\gamma^2)$ at problem-specific bandwidths; these bandwidths were selected to approximately maximize the power of the *ONS betting strategy* (rather than FLOE or FLOE-R) on a validation set. This mean power was computed over 100 tests at time $t = 1000$ for the x-Normal, CIFAR, and RatInABox problems, time $t = 800$ for the Sinusoid and Blobs problems, and $t = 3500$ for the Higgs dataset. This results in the following bandwidths: the Sinusoid problem uses $\gamma = 0.3$; the Blobs dataset uses $\gamma = 0.3$; the x-Normal and CIFAR datasets uses $\gamma = 0.5$; the RatInABox dataset uses $\gamma = 0.1$; the Higgs dataset uses $\gamma = 0.4$. The same procedure was used for the sequential kernel independence test, yielding $\gamma_X = \gamma_Y = 0.18$ for Sinusoid, $\gamma_X = \gamma_Y = 0.9$ for x-Normal, and $\gamma_X = 0.35, \gamma_Y = 0.63$ for RatInABox.

Kernel Selection. Since the only requirement on the prediction strategy is that it must be predictable, we can select the kernel bandwidth of the critic at time t based on any criterion computed from the samples up to time $t - 1$. One approach is to select the bandwidth that maximizes the wealth at time $t - 1$; this is the same as GRAPA except applied to the model parameters instead of the bets. Alternatively, we could also optimize the e-power of the current step, which can be done by using the

loo e-power estimates, i.e.

$$\gamma_t = \arg\max_{\gamma} \sum_{\ell=1}^{t-1} \log(1 + \lambda_t^{(\gamma)} V_\ell^{(t-1,\gamma)}).$$

This strategy is essentially FLOE applied to the kernel bandwidths γ. Figure C.7 compares the GRAPA and FLOE kernel selection approaches with the median heuristic; the reported power is averaged over 1,000 tests and each baseline uses the FLOE betting strategy. To speed up the optimization, we consider a discrete set of 20 evenly-spaced bandwidths between $[0.1, 0.7]$ so that each e-power estimate can be computed in parallel. Overall, FLOE outperforms GRAPA in most problems. Even in Sinusoid, its worst performing problem, FLOE is only slightly worse.

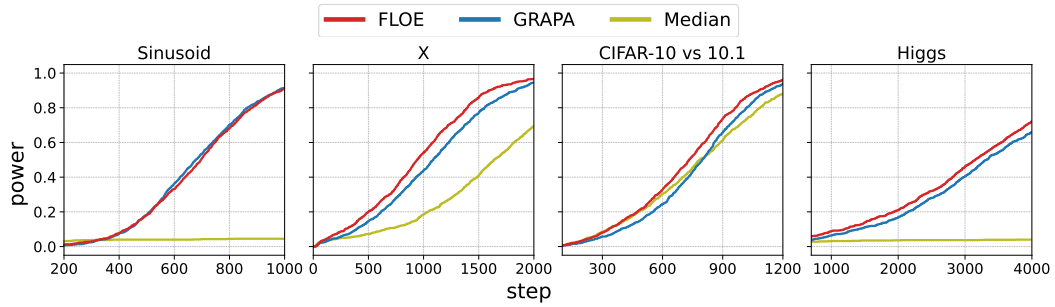


Figure C.7: Kernel selection strategies on a variety of problems.

Performance against STaR-Bets. As an additional baseline, we compare FLOE to the STaR-Bets betting strategy of Voráček and Orabona (2025) for the bounded mean testing problem. To translate their approach (Algorithm 4) to our kernel two-sample testing process, we replace their payoffs $X_t - m$ with our kernel critic payoffs $f_t(X_t) - f_t(X_t)$ and use $\alpha = \delta = 0.05$. Results are reported in Figure C.8.

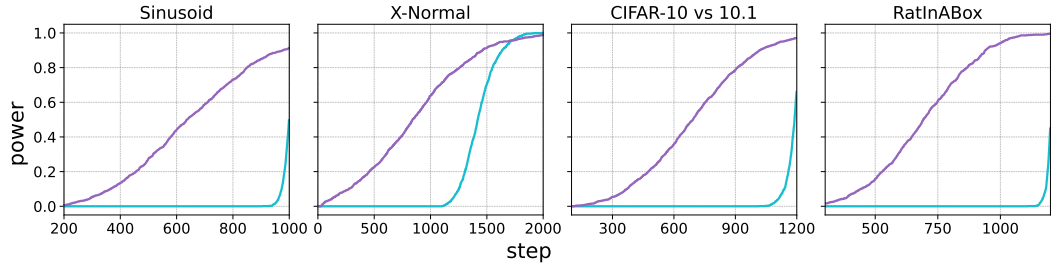


Figure C.8: Power of FLOE (purple) against STaR-Bets (neon blue).

We find that FLOE significantly outperforms STaR-Bets for the two-sample problem. STaR-Bets' burn-in period is quite long, and by the time it starts accumulating wealth FLOE already has near perfect power. That said, after burn-in STaR-Bets' power increases at a far more rapid rate than FLOE. We also point out that FLOE does not apply to the bounded mean testing problem, as there is no critic function to be learned. It's payoffs are simply the centred samples $X_t - m$. If we were to define $f_t(X_t) = X_t - m$, FLOE would simply reduce to GRAPA.

Assets and Licenses. We describe some of the assets and their licenses used in our work:

- Higgs Dataset (Baldi et al., 2014) available from the UCI Machine Learning Repository under the CC BY 4.0 license.
- CIFAR-10 (Krizhevsky and Hinton, 2009) distributed via the UCI Machine Learning Repository under CC BY 4.0.
- CIFAR-10.1 (Recht et al., 2018) released under the MIT license.
- A pretrained ResNet-50 model (He et al., 2016) provided by Torchvision (Paszke et al., 2019) and released under the BSD-3-Clause license.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We touch on our main contributions in the abstract, and provide more details in the introduction (particularly the sub-section titled "contributions").

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The "practical considerations" of section 3.3 discusses limitations of our regularization strategy and proposes a possible workaround. Section 4 also touches on a potential situation where including this regularization is not ideal, as well as including ablation on the sensitivity of the regularizer hyperparameters. Computation complexity is briefly discussed in Section 3.2 and 3.3.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Full proofs for each theorem are provided in the appendix and, to the best of our knowledge, are correct.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our proposed test procedure is described in detail and, apart from the inherent randomness of the data, all dataset settings are described in Section 4 and Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code we used for all experiments is provided as supplementary material. However, only partial instructions are currently given.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All hyperparameters pertaining to our datasets and method are described in section 4 and Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We considered plotting empirical confidence intervals on the power plots of our experiments but found this made results harder to read and were not very informative. Each of our experiments are run 1,000 times and so the mean power is relatively accurate.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The compute resources used is not imperative to our approach. A GPU is helpful, but not required, and memory usage is small enough so that an everyday computer can run it.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research conforms to the NeurIPS code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: In the introduction, we briefly discussed positive societal impacts in terms of mitigating p-hacking. We did not address any negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: There are no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have included a section in Appendix B detailing the assets and licenses used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets are introduced.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or human-based experiments are used.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No crowdsourcing or human-based experiments are used.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Our proposed method was not developed with LLM use.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.