Sequential Kernelized Stein Discrepancy

Diego Martinez-Taboada
Carnegie Mellon University

Aaditya Ramdas
Carnegie Mellon University

Abstract

We present a sequential version of the kernelized Stein discrepancy goodness-of-fit test, which allows for conducting goodness-of-fit tests for unnormalized densities that are continuously monitored and adaptively stopped. That is, the sample size need not be fixed prior to data collection; the practitioner can choose whether to stop the test or continue to gather evidence at any time while controlling the false discovery rate. In stark contrast to related literature, we do not impose uniform boundedness on the Stein kernel. Instead, we exploit the potential boundedness of the Stein kernel at arbitrary point evaluations to define test martingales, that give way to the subsequent novel sequential tests. We prove the validity of the test, as well as an asymptotic lower bound for the logarithmic growth of the wealth process under the alternative. We further illustrate the empirical performance of the test with a variety of distributions, including restricted Boltzmann machines.

1 INTRODUCTION

Many statistical procedures heavily rely on the assumptions made about the distribution of the empirical observations, and prove invalid if such conditions are violated. For instance, a high-energy physicist may develop a generative model seeking to produce synthetic observations of particle collisions, motivated by the high cost of experimenting on a real particle collider (Agostinelli et al., 2003; Chekalina et al., 2019). If the model is not accurate, then any conclusion that derives from such synthetic data will lack any scientific interest (Huang et al., 2023; Masserano et al., 2023). The question of whether the data follows a particular distribution may be posed in terms of goodness-of-fit testing (Lehmann et al., 1986; González-Manteiga and Crujeiras, 2011; D'Agostino, 2017). Formally, the simple goodness-of-fit testing problem considers the null hypothesis $H_0 : Q = P$ against the alternative $H_1 : Q \neq P$, for a given P and access to data $X_1, X_2, \ldots \sim Q$. It may also be the case that we have a set of given distributions as candidates and we wish to test if *any* of them accurately models the empirical data (Durbin, 1975; Key et al., 2021). In such a case, the goodness-of-fit problem is posed in terms of a composite null hypothesis $H_0 : Q \in \mathcal{P}$, against the alternative hypothesis $H_1 : Q \notin \mathcal{P}$.

In this work, we focus on a particularly challenging scenario that arises from not having full access to P (or $\mathcal{P}$). We handle classes of distributions whose densities are only known up to normalizing constants. This is a common scenario when working with general energy-based models (LeCun et al., 2006; Du and Mordatch, 2019), such as Ising models (Geman and Geman, 1984; Clifford, 1990) and restricted Boltzmann machines (Ackley et al., 1985; Hinton, 2010), and in Bayesian statistics, where normalizing factors are generally neither known in closed form nor computable (Bolstad and Curran, 2016). Most of the existing works in the literature regarding goodness-of-fit for unnormalized densities have focused on the *batch* or *fixed sample size* setting, where the number of samples n is decided before conducting the analysis, which is later on carried on using observations $X_1, \ldots, X_n$ (Liu et al., 2016; Chwialkowski et al., 2016; Gorham and Mackey, 2017). Nonetheless, this approach comes with several major drawbacks. Because the number of observations in batch tests is determined prior to data collection, there is a risk that too many observations may be allocated to simpler problem instances, thus wasting resources, or that insufficient observations are assigned to more complex instances, which can yield insufficient evidence against the null hypothesis. Furthermore, when test outcomes appear encouraging but not definitive (for instance, if a p-value is marginally higher than a specified significance thresh-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

old), one may be tempted to augment the dataset and undertake the investigation again. Traditional batch testing methods, however, do not allow for this approach.

In contrast, we seek to develop tests that are anytime valid (Ramdas et al., 2023; Grünwald et al., 2024). That is, we can repeatedly decide whether to collect more data based on the current state of the procedure without compromising later assessments, halting the process at any time and for any reason. Formally, a sequential test which is stopped at any given arbitrary moment can be represented by a random stopping time τ taking values in $\{1, 2, \ldots\} \cup \{\infty\}$. The stopping time τ denotes the random time at which the null hypothesis is rejected. A sequential test is called level-α if it satisfies the condition $\mathbb{P}(\tau < \infty) \leq \alpha$ under the null for any stopping time τ.

In this work, we develop level-α sequential goodness-of-fit tests that accommodate for unnormalized densities, building on the concept of kernelized Stein discrepancies. The type I error is controlled even if the tests are continuously monitored and adaptively stopped, hence automatically adapting the sample size to the unknown alternative. In particular, we believe these anytime-valid tests will be of remarkable interest in the following scenarios:

- **Measuring the quality of a proposed distribution:** We may have a specific candidate distribution (or set of candidate distributions) to model the data whose normalizing constant is unknown (e.g., some complex network that models the interaction between genes). Does this distribution accurately model the true data generation process?
- **Measuring sample quality:** We may have obtained an unnormalized density, such as some posterior distribution in Bayesian statistics, that we wish to draw from using Markov Chain Monte Carlo (MCMC) methods. Is the chosen MCMC algorithm yielding samples accurately from such a density?[1]

The anytime-valid guarantees allow for minimizing the sample size required to reject the null if it does not hold, such as the number of expensive gene experiments that we have to run in a lab, or the number of samples that we have to draw from a costly MCMC algorithm.

The work is organized as follows. Section 2 presents the related work. Section 3 introduces preliminary work, with special emphasis on the kernelized Stein discrepancy and testing by betting. These constitute the theoretical foundations of the novel goodness-of-fit tests, presented in Section 4. Section 5 subsequently provides a general derivation for the lower bounds required in the test, alongside different examples. The empirical validity of the tests is presented in Section 6, followed by concluding remarks in Section 7.

2 RELATED WORK

Our contribution falls within the scope of 'sequential, anytime-valid inference', a field which encompasses confidence sequences (Waudby-Smith and Ramdas, 2024) and e-processes (Ramdas et al., 2022; Grünwald et al., 2024). In particular, we exploit the techniques of 'testing by betting', which was recently popularized by Shafer (2021), and is based on applying Ville's maximal inequality (Ville, 1939) to a test (super)martingale (Shafer et al., 2011). Any other approach for constructing confidence sequences can be, in principle, outperformed by applying maximal inequalities to (super)martingale constructions (Ramdas et al., 2020). The interest on this line of research has recently exploded; we refer to reader to Ramdas et al. (2023) and the references therein for a detailed introduction to the field.

Concurrently, kernel methods have received widespread attention due to their notable empirical performance and theoretical guarantees. Reproducing kernel Hilbert spaces provide the theoretical foundation for these methods (Schölkopf and Smola, 2002; Berlinet and Thomas-Agnan, 2011). We highlight three predominant applications of kernel methods. First, the maximum mean discrepancy (MMD) has given way to kernel-based two sample tests (Gretton et al., 2006, 2012). Second, independence tests have been developed based on the Hilbert Schmidt independence criterion (HSIC) (Gretton et al., 2005, 2007). Third, kernelized Stein discrepancies (KSD) have led to novel goodness-of-fit tests (Liu et al., 2016; Chwialkowski et al., 2016; Gorham and Mackey, 2017). While the analysis of this contribution is tailored to so-called Langevin KSD, the ideas presented herein could potentially be extended to abstract domains including categorical data (Yang et al., 2018), censored data (Fernandez et al., 2020), directional data (Xu and Matsuda, 2020), functional data (Wynne et al., 2025), sequential data (Baum

[1]There exist complexities when sequentially analyzing the fitness of MCMC samples, mainly related to autocorrelation and burn-in stages. In order to minimize the effect of autocorrelation, one could skip samples and only input one every X samples for the sequential test, where X is large enough to considerably lower the correlation between samples. In order to avoid early rejection due to burn-in stage, sequential tests can be initialized after such a burn-in period.

et al., 2023), and point processes (Yang et al., 2019).

At the intersection of testing by betting and kernels, Shekhar and Ramdas (2023) developed a sequential MMD for conducting two sample tests that can be arbitrarily stopped. Similarly, Podkopaev et al. (2023) extended the HSIC to the sequential setting. Recently, Zhou and Liu (2024) proposed a sequential KSD, in a similar spirit to our contribution.[2] These three works rely on the uniform boundedness of the chosen kernel, which is key for the construction of nonnegative martingales. While some kernels commonly used for the MMD and HSIC (such as the RBF or Laplace kernels) are uniformly bounded, this is very rarely the case for the Stein kernel exploited by the KSD. The methodology proposed in this contribution does not assume uniform boundedness of the Stein kernel, which poses a number of theoretical challenges that will be addressed in the subsequent sections.

3 BACKGROUND

3.1 The Kernelized Stein Discrepancy

Throughout this contribution we will draw upon the kernelized Stein discrepancy (KSD), which is a distributional divergence that builds on the concept of reproducing kernel Hilbert space (RKHS) and the general Stein's method (Stein, 1972; Gorham and Mackey, 2015). The KSD may be understood as a kernelized version of score-matching divergence (Hyvärinen and Dayan, 2005).

Reproducing kernel Hilbert spaces (RKHS): Consider $\mathcal{X} \neq \emptyset$ and a Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ of functions $f : \mathcal{X} \to \mathbb{R}$. If there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying (i) $k(\cdot, x) \in \mathcal{H}$ for all $x \in \mathcal{X}$, (ii) $\langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$ for all $x \in \mathcal{X}$ and $f \in \mathcal{H}$, then $\mathcal{H}$ is called an RKHS and k a reproducing kernel. We denote by $\mathcal{H}^d$ the product RKHS containing elements $h := (h_1, \ldots, h_d)$ with $h_i \in \mathcal{H}$ and $\langle h, \tilde{h}\rangle = \sum_{i=1}^d \langle h_i, \tilde{h}_i\rangle_{\mathcal{H}}$.

Kernelized Stein discrepancies (KSD): Assume now that P has density p on $\mathcal{X} \subset \mathbb{R}^d$, and let $\mathcal{H}$ be an RKHS with reproducing kernel k such that $\nabla k(\cdot, x)$ exists for all $x \in \mathcal{X}$. Based on the Stein operator

$$(T_p h)(x) = \sum_{i=1}^d \left(\frac{\partial \log p(x)}{\partial x_i} h_i(x) + \frac{\partial h_i(x)}{\partial x_i} \right), \; h \in \mathcal{H}^d,$$

the KSD is defined as

$$\mathrm{KSD}_{\mathcal{H}}(Q, P) = \sup_{f \in \mathcal{F}_{\mathrm{KSD}}} \mathbb{E}_{X\sim Q}[(T_p f)(X)],$$

where we consider the set $\mathcal{F}_{\mathrm{KSD}} = \{h \in \mathcal{H}^d : \|h\|_{\mathcal{H}^d} \leq 1\}$. Defining $s_p(x) := \nabla_x \log p(x)$ and $\xi_p(\cdot, x) := [s_p(x)k(\cdot, x) + \nabla k(\cdot, x)] \in \mathcal{H}^d$, it follows that

$$\begin{aligned} h_p(x, \tilde{x}) :&= \langle \xi_p(\cdot, x), \xi_p(\cdot, \tilde{x})\rangle_{\mathcal{H}^d} \\ &= \langle s_p(x), s_p(\tilde{x})\rangle_{\mathbb{R}^d} k(x, \tilde{x}) \\ &\quad + \langle s_p(\tilde{x}), \nabla_x k(x, \tilde{x})\rangle_{\mathbb{R}^d} \\ &\quad + \langle s_p(x), \nabla_{\tilde{x}} k(x, \tilde{x})\rangle_{\mathbb{R}^d} + \nabla_x \cdot \nabla_{\tilde{x}} k(x, \tilde{x}) \end{aligned} \tag{1}$$

is a reproducing kernel based on the Moore-Aronszajn theorem. Again, if $\mathbb{E}_{X\sim Q}\sqrt{h_p(X, X)} < \infty$, then $\mathrm{KSD}_{\mathcal{H}}(Q, P) = \|\mathbb{E}_{X\sim Q}[\xi_p(\cdot, X)]\|_{\mathcal{H}^d}$. If k is universal and under certain regularity conditions, then $\mathrm{KSD}_{\mathcal{H}}(Q, P) = 0$ if, and only if, $Q = P$ (Chwialkowski et al., 2016). In the batch setting and simple null hypothesis, a test statistic is generally taken as a V-statistic or U-statistic, and parametric or wild bootstrap is used to calibrate the test. We refer the reader to Key et al. (2021) for the more challenging composite null hypothesis case.[3]

3.2 Testing by Betting

We now present the theoretical foundation of the sequential testing strategy that will be presented in Section 4, which is commonly referred to as *testing by betting*. The key idea behind this general, powerful concept relies on defining a test (super)martingale, and couple it with a betting interpretation (Shafer and Vovk, 2019; Shafer, 2021). Test (super)martingales allow for exploiting Ville's inequality (Ville, 1939) and hence they yield level-α sequential tests. In turn, this enables practitioners to peek at the data at anytime and decide whether to keep gathering data or stop accordingly while preserving theoretical guarantees.

Intuitively, let H_0 be the null hypothesis to be tested. A fictional bettor starts with an initial wealth of $\mathcal{K}_0 = 1$. The fictional bettor then bets sequentially on the outcomes $(X_t)_{t\geq 1}$. To do so, at each round t, the fictional bettor chooses a payoff function $\mathcal{S}_t : \mathcal{X} \to [0, \infty)$ so that $\mathbb{E}_{H_0}[\mathcal{S}_t(X_t)|\mathcal{F}_{t-1}] \leq 1$, where $\mathcal{F}_{t-1} = \sigma(X_1, \ldots, X_{t-1})$ (this ensures a *fair bet* if the null is true). After the outcome X_t is revealed, the bettor's wealth grows or shrinks by a factor of $\mathcal{S}_t(Z_t)$. The bettor's wealth is $\mathcal{K}_t = \mathcal{K}_0 \prod_{i=1}^t \mathcal{S}_i(X_i)$ after t rounds of betting.

Under the null hypothesis, these *fair bets* ensure that the sequence $(\mathcal{K}_t)_{t\geq 0}$ forms a 'test supermartingale' (i.e., a nonnegative supermartingale that

[2] We defer a comprehensive comparison of Zhou and Liu (2024) and our contribution to Appendix B.

[3] In the context of approximating an integral, the KSD also allows for evaluating the realized empirical measure (rather than the marginal distribution). We refer the reader to Kanagawa et al. (2022) for a recent account of this research direction.

starts at 1), and in particular a 'test martingale' if $\mathbb{E}_{H_0}[\mathcal{S}_t(X_t)|\mathcal{F}_{t-1}]$ is 1 almost surely. Based on Ville's inequality, rejecting the null if $\mathcal{K}_t \geq 1/\alpha$, where $\alpha \in (0,1)$ is the desired confidence level, leads to sequential tests that control the type-I error at level α. Under the alternative H_1, the payoff functions $\{\mathcal{S}_t : t \geq 1\}$ should seek a fast growth rate of the wealth, ideally exponentially.

4 SEQUENTIAL GOODNESS-OF-FIT BY BETTING

We now consider a stream of data $X_1, X_2, \ldots \sim Q$. In the context of testing by betting, it is natural to consider test martingales of the form

$$\mathcal{K}_t = \mathcal{K}_{t-1} \times (1 + \lambda_t g_t(X_t)), \quad \mathcal{K}_0 = 1, \tag{2}$$
$$\tau := \min\{t \geq 1 : \mathcal{K}_t \geq 1/\alpha\}, \tag{3}$$

where λ_t and g_t are predictable and

- $\mathbb{E}_{H_0}\left[g_t(X_t)\middle|\mathcal{F}_{t-1}\right] \leq 0$ (to ensure $\mathcal{K}_t$ is a supermartingale under the null hypothesis),
- $g_t \geq -1$ and $\lambda_t \in [0,1]$ (to ensure that $\mathcal{K}_t$ is nonnegative),

so that $\mathcal{K}_t$ is a test supermartingale. Intuitively, $g_t(X_t)$ defines the payoff function, and λ_t the proportion of the wealth that the bettor is willing to risk at round t. Usually, the payoff function g_t is assumed to belong to a set $\mathcal{G}$ that is **uniformly bounded by one in absolute value** and such that $\mathbb{E}_{H_0}\left[g(X)\middle|\mathcal{F}_{t-1}\right] \leq 0$ for all $g \in \mathcal{G}$, and is chosen predictably seeking to maximize $\mathcal{K}_t$ under the alternative.

For the Stein kernel h_p, we note that, if $\mathbb{E}[h_p(X,X)] < \infty$ (Chwialkowski et al., 2016, Lemma 5.1),

$$\begin{aligned}
&\mathbb{E}_{H_0}\left[\frac{1}{t-1}\sum_{i=1}^{t-1} h_p(X_i, X_t)\middle|\mathcal{F}_{t-1}\right] \qquad (4)\\
=&\mathbb{E}_{H_0}\left[\frac{1}{t-1}\sum_{i=1}^{t-1}\langle\xi_p(\cdot,X_i),\xi_p(\cdot,X_t)\rangle_{\mathcal{H}^d}\middle|\mathcal{F}_{t-1}\right]\\
=&\frac{1}{t-1}\sum_{i=1}^{t-1}\langle\xi_p(\cdot,X_i),\mathbb{E}_{H_0}\left[\xi_p(\cdot,X_t)\middle|\mathcal{F}_{t-1}\right]\rangle_{\mathcal{H}^d}\\
=&\frac{1}{t-1}\sum_{i=1}^{t-1}\langle\xi_p(\cdot,X_i),0\rangle_{\mathcal{H}^d}\\
=&0.
\end{aligned}$$

Hence, if h_p was uniformly bounded by one, we could define $g_t(x) = \frac{1}{t-1}\sum_{i=1}^{t-1} h_p(X_i, x)$. Similar payoff functions have been proposed in Shekhar and Ramdas (2023) for two sample testing, and in Podkopaev et al. (2023) for independence testing. In those settings, it is natural to consider kernels that are indeed uniformly bounded by one, and this bound is not too loose[4]. This is the case for the ubiquitous Gaussian (or RBF) and Laplace kernels.

In stark contrast, the Stein kernel h_p need not be uniformly bounded even if built from a uniformly bounded k (or it may be uniformly bounded by an extremely large constant, which would imply a remarkable loss in power if simply normalizing by that constant). Consider the inverse-multi quadratic (IMQ) kernel

$$k_{\mathrm{IMQ}}(x,y) = (1 + \|x-y\|^2)^{-1/2}$$

The IMQ kernel has been extensively employed as the base kernel for constructing the Stein kernel. The success of the IMQ kernel over other common characteristic kernels can be attributed to its slow decay rate (Gorham and Mackey, 2017). For this reason, we build the Stein kernel h_p from k_{IMQ} throughout. Now take $P = \mathcal{N}(0,1)$ to be a standard normal distribution. In such a case, $s_p(x) = -x$, and $h_p(x,x) = x^2 + 1$. This implies that $\sup_{x\in\mathbb{R}} h_p(x,x) = \infty$, i.e., the kernel h_p is not bounded.

Nonetheless, it is very possible that $x \mapsto h_p(x,\tilde{x})$ is bounded for every fixed $\tilde{x}$. We can rewrite

$$\begin{aligned}
h_p(x,\tilde{x}) &= \langle\xi_p(\cdot,x),\xi_p(\cdot,\tilde{x})\rangle_{\mathcal{H}^d}\\
&= \|\xi_p(\cdot,x)\|_{\mathcal{H}^d}\|\xi_p(\cdot,\tilde{x})\|_{\mathcal{H}^d}\cos\beta(x,\tilde{x}),
\end{aligned}$$

where $\cos\beta(x,\tilde{x})$ is the angle between $\xi_p(\cdot,x)$ and $\xi_p(\cdot,\tilde{x})$ in the Hilbert space $\mathcal{H}^d$. Interestingly, while the embeddings $\|\xi_p(\cdot,x)\|_{\mathcal{H}^d}$ may not be uniformly bounded (this is equivalent to h_p not being uniformly bounded), $\cos\beta(x,\tilde{x})$ may decay to zero faster than $\|\xi_p(\cdot,x)\|_{\mathcal{H}}$ approaches infinity (for any given $\tilde{x}$). From now on, we assume that we have access to a function $M_p : \mathcal{X} \mapsto \mathbb{R}_{\geq 0}$ such that

$$M_p(\tilde{x}) \geq -\inf_{x\in\mathcal{X}} h_p(x,\tilde{x}).$$

The upper bound M_p plays a key role in the proposed test, as it allows for defining martingales that are nonnegative. We defer to Section 5 a general method to derive such an upper bound, as well as specific examples of such derivation.

[4]Shekhar and Ramdas (2023) proposed an extension to unbounded kernels. They exploit the symmetry (around zero) of their payoff function under the null, which follows from the nature of two sample testing. Such a symmetry does not hold for us.

4.1 Simple Null Hypothesis

We are now ready to present the novel sequential goodness-of-fit tests. Let P be a distribution which is known up to its normalizing constant, and let us consider the null hypothesis $H_0 : Q = P$ against the alternative $H_1 : Q \neq P$. Note that the scores $s_p(x)$ of P do not depend on such normalizing factors and thus the reproducing kernel h_p is computable.

Again, we emphasize that defining a wealth process based on the payoff function $\frac{1}{t-1}\sum_{i=1}^{t-1} h_p(X_i, x)$ does not, in principle, yield a nonnegative martingale under the null. Nonetheless, the normalized payoff function

$$g_t(x) = \frac{1}{\frac{1}{t-1}\sum_{i=1}^{t-1} M_p(X_i)} \left(\frac{1}{t-1} \sum_{i=1}^{t-1} h_p(X_i, x) \right). \tag{5}$$

is lower bounded by -1. Hence, as long as the betting strategy λ_t is nonnegative and bounded above by 1, we are able to define a wealth process that forms a test martingale. The following theorem formally establishes such a result; the proof is deferred to Appendix C.

Theorem 1 (Validity under null.)**.** *Assume that $\mathbb{E}_{H_0}[h_p(X, X')] = 0$, and let $\lambda_t \in [0, 1]$ be predictable. The wealth process*

$$\mathcal{K}_t = \mathcal{K}_{t-1} \times (1 + \lambda_t g_t(X_t)), \quad \mathcal{K}_0 = 1, \tag{6}$$

where g_t is defined as in (5), is a test martingale. The stopping time

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

defines a level-α sequential test.

Algorithm 1 summarizes the procedure introduced in this section. Given that the complexity of computing g_t is $O(t)$ for each t, the complexity of Algorithm 1 for T rounds is $O(T^2)$. Note that $\mathbb{E}[h_p(X, X')] = 0$ under the null if $\mathbb{E}[h_p(X, X)] < \infty$, and thus in that case Theorem 1 establishes the validity of Algorithm 1 for any betting strategy that is predictable. However, the power of the test will heavily depend on the chosen betting strategy. For instance, if we take $\lambda_t = 0$ for all t (this is, we never bet any money), then the wealth will remain constant as $\mathcal{K}_t = 1$, and so the test is powerless. There exists a variety of betting strategies that have been studied in the literature. In this contribution, we focus on aGRAPA ('approximate GRAPA') and LBOW ('Lower-Bound On the Wealth'). We refer the reader to Waudby-Smith and Ramdas (2024) to a detailed presentation of different betting strategies.

Definition 1 (aGRAPA strategy)**.** *Let $(g_i(X_i))_{t\geq 1} \in [-1, \infty)^{\mathbb{N}}$ denote a sequence of outcomes. Initialize $\lambda_1^{aGRAPA} = 0$. For each round $t = 1, 2, \ldots$, observe payoff $g_i(X_i)$ and update*

$$\lambda_{t+1}^{aGRAPA} = 1 \wedge \left(0 \vee \frac{\frac{1}{t-1}\sum_{i=1}^{t-1} g_i(X_i)}{\frac{1}{t-1}\sum_{i=1}^{t-1} g_i^2(X_i)} \right). \tag{7}$$

Definition 2 (LBOW strategy)**.** *Let $(g_i(X_i))_{t\geq 1} \in [-1, \infty)^{\mathbb{N}}$ denote a sequence of outcomes. Initialize $\lambda_1^{LBOW} = 0$. For each round $t = 1, 2, \ldots$, observe payoff $g_i(X_i)$ and update*

$$\lambda_{t+1}^{LBOW} = 0 \vee \frac{\frac{1}{t-1}\sum_{i=1}^{t-1} g_i(X_i)}{\frac{1}{t-1}\sum_{i=1}^{t-1} g_i(X_i) + \frac{1}{t-1}\sum_{i=1}^{t-1} g_i^2(X_i)}. \tag{8}$$

We have introduced versions of the betting strategies where we force λ_t to be nonnegative. This need not always be the case. The idea of not allowing for negative bets was exploited by Podkopaev et al. (2023) as well, motivated by the fact that positive payoffs are expected under the alternative. The motivation is double in this work, as we also expect positive payoffs under the alternative, but the nonnegativity of λ_t allows us to only having to lower bound the payoff function g_t, instead of lower and upper bounding it. While the aGRAPA strategy shows better empirical performance, LBOW allows for providing theoretical guarantees of the asymptotic wealth under the alternative.

Theorem 2 (E-power under alternative)**.** *Let $(X_i)_{i\geq 1}$ be independent and identically distributed copies of X. Assume $\mathbb{E}[h_p(X, X)] < \infty$ and $\mathbb{E}[M_p(X)] < \infty$. Denote $g^*(x) := \mathbb{E}[h_p(X, x)]/\mathbb{E}[M_p(X)]$. Under H_1, if $\mathbb{E}[g^*(X)] > 0$, the LBOW betting strategy yields*

$$\liminf_{t\to\infty} \frac{\log \mathcal{K}_t}{t} \geq \frac{\left(\mathbb{E}[g^*(X)]\right)^2/2}{\mathbb{E}[g^*(X)] + \mathbb{E}[(g^*(X))^2]} := r^*.$$

It follows that $\mathcal{K}_t \overset{a.s.}{\to} \infty$ and $\mathbb{P}_{H_1}(\tau < \infty) = 1$.

Informally, the above theorem states that under the alternative, $\mathcal{K}_t \geq \exp(r^* t(1 - o(1)))$, meaning that up to asymptotically negligible terms, the wealth grows exponentially fast (in the number of data points t) at the rate r^*. Note that under the null, $\mathbb{E}[h_p(X, X')] = 0$, implying $\mathbb{E}[g^*(X)] = 0$, and thus $r^* = 0$, which accords with our claim that under the null, the wealth is a nonnegative martingale (whose expectation stays constant with t) and thus does not grow with t. This then implies that the stopping time of the test, which is the time at which the wealth exceeds $1/\alpha$, is (up to leading order) given by the expression $\log(1/\alpha)/r^*$.

The proof of Theorem 2 is deferred to Appendix C. We point out that the unboundedness of the Stein kernel

does not allow for easily extending the arguments presented in Podkopaev et al. (2023), which rely on a different betting strategy whose guarantees stem from uniformly bounded payoff functions. Furthermore, we highlight the mildness of the assumptions in Theorem 2. Assumption $\mathbb{E}[M_p(X)] < \infty$ only requires the first moment of the upper bounds to exist. This condition usually reduces to the existence of a specific moment of the original distribution, which is often easily verifiable. Assumption $\mathbb{E}[h_p(X, X)] < \infty$ is an ubiquitous assumption in the KSD theory; it is equivalent to the existence of the second moment of $\|\xi_p(\cdot, X)\|_{\mathcal{H}^d}$, which is precisely the theoretical object that the KSD builds on. Finally, note that $P \neq Q$ does not necessarily imply $\mathbb{E}[g^*(X)] > 0$. However, there exist sufficient conditions for this implication to hold, such as C_0-universality of k and a moment condition on $\nabla \log(p(X)/q(X)$ (Chwialkowski et al., 2016, Theorem 2.2.) or a root exponential growth of $\|s_p\|$ for a rich family of base kernels (Barp et al., 2024, Application 2). These conditions have already been extensively studied in the batch setting (Chwialkowski et al., 2016; Liu et al., 2016; Gorham and Mackey, 2017; Barp et al., 2024), and equally apply to our setting.

Algorithm 1 Sequential KSD

Input: Significance level α; data stream $X_1, X_2, \ldots \sim Q$, score function s_p, base kernel k.
Define h_p from s_p and k as in (1).
for $t = 1, 2, \ldots$ **do**
  Observe X_t;
  Compute $g_t(X_t)$ using (5) and $\mathcal{K}_t$ using (6);
  if $\mathcal{K}_t \geq 1/\alpha$ **then**
    Reject H_0 and stop;
  else
    Compute $\lambda_{t+1} \in [0, 1]$ following (7) or (8);
  end if
end for

4.2 Composite Null Hypothesis

Let now $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a set of distributions with unknown normalizing constants parametrized by θ. By a slight abuse of notation, we denote $s_\theta = s_{p_\theta}$, $h_\theta = h_{p_\theta}$, $M_\theta = M_{p_\theta}$, and so on. Consider the null hypothesis $H_0 : Q \in \mathcal{P}$ against $H_1 : Q \notin \mathcal{P}$. We define

$$g_t^\theta(x) = \frac{1}{\frac{1}{t-1}\sum_{i=1}^{t-1} M_\theta(X_i)} \left(\frac{1}{t-1} \sum_{i=1}^{t-1} h_\theta(X_i, x) \right)$$

for $\theta \in \Theta$. We note that, under the null hypothesis, there exists $\theta_0 \in \Theta$ such that $Q = P_{\theta_0}$. Inspired by universal inference, we propose to consider the wealth process

$$\mathcal{K}_t^C = \min_{\theta \in \Theta} \mathcal{K}_t^\theta.$$

Theorem 3. *$\mathcal{K}_t^C$ is dominated by a test supermartingale, and τ is a level-α sequential test.*

The minimizer may be computed differently depending on the nature of the problem. If Θ is finite, then it is obtained as the minimum of a discrete set. For arbitrary Θ, it may be computed using numerical optimisation algorithms.

5 DERIVATION OF SENSIBLE BOUNDS

5.1 A General Approach

We highlight that the derivation of the bounds $M_p(x)$ is key on this approach. While these bounds heavily depend on the distribution P through its score function s_p, we explore general ways of deriving bounds $M_p(x)$. We focus on deriving bounds for the IMQ. Nonetheless, we highlight that similar derivations would follow for other kernels, such as the RBF or Laplace kernels. We start by noting that, for the IMQ kernel,

$$\begin{aligned} \nabla_x k(x, \tilde{x}) &= -(1 + \|x - \tilde{x}\|^2)^{-\frac{3}{2}}(x - \tilde{x}), \\ \nabla_x \cdot \nabla_{\tilde{x}} k(x, \tilde{x}) &= -3(1 + \|x - \tilde{x}\|^2)^{-\frac{5}{2}}\|x - \tilde{x}\|^2 \\ &\quad + d(1 + \|x - \tilde{x}\|^2)^{-\frac{3}{2}}. \end{aligned}$$

Hence, for a fixed $\tilde{x}$,

- $k(x, \tilde{x})$ is $O(\|x - \tilde{x}\|^{-1})$,
- $\|\nabla_x k(x, \tilde{x})\|$ is $O(\|x - \tilde{x}\|^{-2})$,
- $\nabla_x \cdot \nabla_{\tilde{x}} k(x, \tilde{x}) \geq \min(-3 + d, 0)$.

In order to explicitly obtain a bound, we work with each of the terms (i) $|\langle s_p(x), s_p(\tilde{x})\rangle_{\mathbb{R}^1} k(x, \tilde{x})| \leq \|s_p(x)\|\|s_p(\tilde{x})\|k(x, \tilde{x})|$, (ii) $\langle s_p(\tilde{x}), \nabla_x k(x, \tilde{x})\rangle_{\mathbb{R}^d} + \langle s_p(x), \nabla_{\tilde{x}} k(x, \tilde{x})\rangle_{\mathbb{R}^d}$, (iii) $\nabla_x \cdot \nabla_{\tilde{x}} k(x, \tilde{x})$ separately. For term (i), we upper bound $\|s_p(x)\|$ by $\gamma(\|x - \tilde{x}\|) + \gamma'(\tilde{x})$, where γ and γ' are appropriate functions. For term (ii), we note that

$$\begin{aligned} &\langle s_p(\tilde{x}), \nabla_x k(x, \tilde{x})\rangle_{\mathbb{R}^d} + \langle s_p(x), \nabla_{\tilde{x}} k(x, \tilde{x})\rangle_{\mathbb{R}^d} \\ =&\langle s_p(\tilde{x}) - s_p(x), -\nabla_x k(x, \tilde{x})\rangle_{\mathbb{R}^d} \\ =&\langle s_p(\tilde{x}) - s_p(x), (1 + \|x - \tilde{x}\|^2)^{-\frac{3}{2}}(x - \tilde{x})\rangle_{\mathbb{R}^d} \\ =&(1 + \|x - \tilde{x}|^2)^{-\frac{3}{2}}\langle s_p(\tilde{x}) - s_p(x), x - \tilde{x}\rangle_{\mathbb{R}^d}, \end{aligned}$$

so it suffices to upper bound work with $s_p(\tilde{x}) - s_p(x)$ and upper bound $\|s_p(\tilde{x}) - s_p(x)\|$ by $\Gamma(\|x - \tilde{x}\|) + \Gamma'(\tilde{x})$,

where Γ and Γ' are again appropriate functions. Note that (iii) has already been lower bounded.

The choices of $\gamma, \gamma', \Gamma, \Gamma'$ will become clear in the next subsection, where we provide specific examples. We highlight that there should exist sensible choices as soon as $\|s_p(x)\|k(x,\tilde{x})$ and $\|s_p(x) - s_p(\tilde{x})\|\|\nabla k(x,\tilde{x})\|$ are both $O(1)$ for fixed $\tilde{x}$, i.e. $\|s_p(x)\| = O(\|x\|)$ in the case of the IMQ kernel. In addition to the cases considered in the next subsection, which include Gaussian distributions and restricted Boltzmann machines, a variety of models may fulfill this condition. In particular, exponential families with canonical parameters η_i and canonical observations T_i of the form $p_\theta(x) \propto \exp(\sum_{i \leq m} \eta_i(\theta) T_i(x))$ fall under this umbrella if $\|\nabla T_i(x)\| = O(\|x\|)$ for all $i \leq m$. For instance, exponential graphical models with linear interactions between nodes, which have found applications in protein signaling networks (see, e.g., Yang et al. (2015, Section 4.2.2.)) and spatial models (see, e.g., Besag (1986, Section 4.2.2.) and Besag (1974, Section 4.1.)) among others, attain such an assumption. Kernel exponential family models (Canu and Smola, 2006) are another prominent example, which fulfill the condition as long as the gradients of the basis functions of a finite-rank approximation (as considered in Matsubara et al. (2022)) are $O(\|x\|)$ (which is precisely the case in Matsubara et al. (2022)).

5.2 Specific Examples of Bound Derivations

We present specific instances of bound derivations in order to elucidate the use of the general approach. The bounds obtained here will be later used in the experiments displayed in Section 6.

Gaussian distribution: Let us consider $P_\theta = \mathcal{N}(\theta, 1)$, i.e., a normal distribution with mean θ and unit variance, and k to be the inverse multi-quadratic kernel. Given that $s_\theta(x) = -(x - \theta)$, we first derive

$$\begin{aligned}|\langle s_\theta(x), s_\theta(\tilde{x})\rangle k(x,\tilde{x})| &\leq |x-\theta||\tilde{x}-\theta|k(x,\tilde{x}) \\ &\leq (|x-\tilde{x}| + |\tilde{x}-\theta|)\,|\tilde{x}-\theta| \\ &\quad \times k(x,\tilde{x}) \\ &\leq |\tilde{x}-\theta|\,(1+|\tilde{x}-\theta|)\,,\end{aligned}$$

where the last inequality can be easily derived by separately considering the two cases $|x - \tilde{x}| \leq 1$ and $|x - \tilde{x}| > 1$. Secondly,

$$\begin{aligned}&(1+|x-\tilde{x}|^2)^{-\frac{3}{2}}\langle s_p(\tilde{x}) - s_p(x), x - \tilde{x}\rangle \\ =& -(1+|x-\tilde{x}|^2)^{-\frac{3}{2}}|x-\tilde{x}| \in [-1, 0].\end{aligned}$$

For a fixed $\tilde{x}$, it thus follows that i) $|\langle s_\theta(x), s_\theta(\tilde{x})\rangle k(x,\tilde{x})| \leq |\tilde{x} - \theta|\{1 + |\tilde{x} - \theta|\}$, ii) $\langle s_\theta(\tilde{x}), \nabla_x k(x,\tilde{x})\rangle + \langle s_\theta(x), \nabla_{\tilde{x}} k(x,\tilde{x})\rangle \in [-1, 0]$, iii) $\nabla_x \cdot \nabla_{\tilde{x}} k(x,\tilde{x}) \geq -2$.

Consequently, it suffices to define $M_\theta(\tilde{x}) = |\tilde{x} - \theta|\{1 + |\tilde{x} - \theta|\} + 3$. Note that $\mathbb{E}[M_\theta(X)] < \infty$ given that the first two moments of a Gaussian distribution are finite.

Intractable model: Following the examples in Liu et al. (2019) and Matsubara et al. (2022), we consider the intractable model with density $p_\theta(y) \propto \exp(\theta_1 \tanh x_1 + \theta_2 \tanh x_2 - 0.5\|x\|^2)$, where $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$, $x \in \mathbb{R}^3$. For $\theta = 0$, we recover the density of a Gaussian distribution $\mathcal{N}(0, I_3)$. Given that

$$s_\theta(x) = \left(\theta_1(1-\tanh^2 x_1), \theta_2(1-\tanh^2 x_2), 0\right)^T - x,$$

we first derive

$$\begin{aligned}&\|\left(\theta_1(1-\tanh^2 x_1), \theta_2(1-\tanh^2 x_2), 0\right)^T - x\| \\ \leq& \|\theta\| + \|\tilde{x} - x\| \\ &+ \|\tilde{x} - \left(\theta_1(1-\tanh^2 \tilde{x}_1), \theta_2(1-\tanh^2 \tilde{x}_2), 0\right)^T\| \\ \leq& \|\theta\| + \|s_\theta(\tilde{x})\| + \|\tilde{x} - x\|,\end{aligned}$$

and so $|\langle s_\theta(x), s_\theta(\tilde{x})\rangle k(x,\tilde{x})|$ is upper bounded by

$$\begin{aligned}&\|s_\theta(x)\|\|s_\theta(\tilde{x})\|k(x,\tilde{x}) \\ \leq& \left(\|\theta\| + \|s_\theta(\tilde{x})\| + \|\tilde{x} - x\|\right)\|s_\theta(\tilde{x})\|k(x,\tilde{x}) \\ \leq& \left(\|\theta\| + \|s_\theta(\tilde{x})\| + 1\right)\|s_\theta(\tilde{x})\|.\end{aligned}$$

Secondly, $\|s_p(\tilde{x}) - s_p(x)\| \leq \|\theta\| + \|x - \tilde{x}\|$, and so

$$\begin{aligned}&\left|(1+\|x-\tilde{x}\|^2)^{-\frac{3}{2}}\langle s_p(\tilde{x}) - s_p(x), x - \tilde{x}\rangle\right| \\ \leq& (1+\|x-\tilde{x}\|^2)^{-\frac{3}{2}}\|x-\tilde{x}\|\left(\|\theta\| + \|x-\tilde{x}\|\right) \\ \leq& \|\theta\| + 1.\end{aligned}$$

For a fixed $\tilde{x}$, it thus follows that i) $|\langle s_\theta(x), s_\theta(\tilde{x})\rangle k(x,\tilde{x})| \leq (\|\theta\| + \|s_\theta(\tilde{x})\| + 1)\|s_\theta(\tilde{x})\|$, ii) $|\langle s_\theta(\tilde{x}), \nabla_x k(x,\tilde{x})\rangle + \langle s_\theta(x), \nabla_{\tilde{x}} k(x,\tilde{x})\rangle| \leq \|\theta\| + 1$, iii) $\nabla_x \cdot \nabla_{\tilde{x}} k(x,\tilde{x}) \geq 0$.

We thus define $M_\theta(\tilde{x}) = (\|\theta\| + \|s_\theta(\tilde{x})\| + 1)\|s_\theta(\tilde{x})\| + \|\theta\| + 1$. Note that $\mathbb{E}[M_\theta(X)] < \infty$ given that the distribution has Gaussian type tails, and so its first two moments are finite.

Gaussian-Bernoulli Restricted Boltzmann Machine: We consider P to be a Gaussian-Bernoulli Restricted Boltzmann Machine, following related contributions in the literature (Liu et al., 2016; Schrab et al., 2022). It is a graphical model that includes a binary hidden variable h, taking values in $\{1, -1\}^{d_h}$, and a continuous observable variable x within $\mathbb{R}^d$. These variables are linked by the joint density function

$$p(x,h) = \frac{1}{Z}\exp\left(\frac{1}{2}x^T B h + b^T x + c^T h - \frac{1}{2}\|x\|_2^2\right),$$

where Z is the normalizing constant. It follows that the density p of x is

$$p(x) = \sum_{h \in \{-1,1\}^{d_h}} p(x, h).$$

The computation of p for large dimension d_h is intractable; nonetheless, the score function is computable as

$$s_p(x) = b - x + \frac{B}{2}\phi(\frac{B^T x}{2} + c), \quad \phi(y) = \frac{e^{2y} - 1}{e^{2y} + 1}.$$

We first derive

$$\left\| \frac{B}{2}\phi(\frac{B^T x}{2} + c) - \frac{B}{2}\phi(\frac{B^T \tilde{x}}{2} + c) \right\| \overset{(i)}{\leq} \|B\|_{op}\sqrt{d_h},$$

where (i) is obtained given that $\phi(y) \in [-1, 1]^{d_h}$ for all y. Thus $|\langle s_p(x), s_p(\tilde{x})\rangle k(x, \tilde{x})| \leq \|s_p(x)\|\|s_p(\tilde{x})\|k(x, \tilde{x})$ is upper bounded by $\left(\|s_p(\tilde{x})\| + \|\tilde{x} - x\| + \left\|\frac{B}{2}\phi(\frac{B^T x}{2} + c) - \frac{B}{2}\phi(\frac{B^T \tilde{x}}{2} + c)\right\|\right)$ $\times\|s_p(\tilde{x})\|k(x, \tilde{x})$, which is again upper bounded by

$$\left(\|s_p(\tilde{x})\| + 1 + \|B\|_{op}\sqrt{d_h}\right)\|s_p(\tilde{x})\|.$$

Secondly, $\|s_p(\tilde{x}) - s_p(x)\| \leq \|B\|_{op}\sqrt{d_h} + \|x - \tilde{x}\|$, and so $\left|(1 + \|x - \tilde{x}\|^2)^{-\frac{3}{2}}\langle s_p(\tilde{x}) - s_p(x), x - \tilde{x}\rangle\right|$ is upper bounded by

$$\begin{aligned}&(1 + \|x - \tilde{x}\|^2)^{-\frac{3}{2}}\|x - \tilde{x}\|\left(\|B\|_{op}\sqrt{d_h} + \|x - \tilde{x}\|\right)\\ \leq&\|B\|_{op}\sqrt{d_h} + 1.\end{aligned}$$

Lastly, we have that $\langle \nabla_x k(\cdot, x), \nabla_{\tilde{x}} k(\cdot, \tilde{x})\rangle_{\mathcal{H}^d} \geq 0$. Hence it suffices to define $M_p(\tilde{x}) = \left(\|s_p(\tilde{x})\| + 1 + \|B\|_{op}\sqrt{d_h}\right)\|s_p(\tilde{x})\| + \|B\|_{op}\sqrt{d_h} + 1$. We can further upper bound $\|B\|_{op} \leq \|B\|_{fr}$, with the Frobenius norm being easily computable. Note that $\mathbb{E}[M_p(X)] < \infty$ given that the distribution has Gaussian type tails, and so its first two moments are finite.

6 EXPERIMENTS

We consider the distributions introduced in the previous section, using the upper bounds derived therein. **Gaussian distribution:** We take $\theta_0 = 0$ under the null. **Intractable model:** We take $\theta_0 = (0, 0)$ under the null. We defer the restricted Boltzmann machine example to Appendix A.1. The code may be found here.

Figure 1 exhibits the performance of the proposed test with $\alpha = 0.05$, under the alternatives $\theta_1 = 1$ and $\theta_1 = (1, 1)$ respectively. We emphasize the exponential growth of the wealth process. This behaviour is expected, in view of Theorem 2 and the fact that all these examples fulfil regularity conditions so that $\mathbb{E}_{H_1}[g^*(X)] > 0$; an actual empirical verification of the lower bound derived in Theorem 2 is deferred to Appendix A.6. Furthermore, we stress that aGRAPA empirically outperforms LBOW. This is due to the fact that $\lambda_{t+1}^{\text{aGRAPA}} > \lambda_{t+1}^{\text{LBOW}}$, so aGRAPA bets more aggressively. We defer the illustration of the type-I error control at the desired level 0.05 to Appendix A.2, a discussion on how the tightness of the bounds M_p empirically affects the power of the test to Appendix A.5, and examples of testing composite null hypotheses to Appendix A.4.

We emphasize once again that, in contrast to sequential procedures, batch setting algorithms implicitly commit to a sample size that is chosen prior to running the experiment, which may lead to substantially suboptimal choices of sample sizes. We exhibit such a phenomenon in Figure 2, which displays the proportion of rejections for the Gaussian distribution under different alternatives and illustrates the convenience of employing anytime valid tests. If using the classical kernelized Stein discrepancy test (left plot), 20 observations lead to rejection rates lower than 0.50, and 400 observations show rejection rates of 1. This implies that 20 observations are not enough to obtain a powerful test, and probably less than 500 observations would have sufficed to yield high power (resulting in a cheaper experiment). However, this cannot be known beforehand. In stark contrast, all the null hypotheses are eventually rejected by the proposed test (right plot), with the empirical powers being significantly high already after 150 observations, and always being able to collect more data and keep running the experiment with anytime validity. We defer a longer comparison between the batch and sequential tests to Appendix A.3.

7 CONCLUSION

We have developed a sequential version of the kernelized Stein discrepancy goodness-of-fit test, which gives way to goodness-of-fit tests that can handle distributions with unknown normalizing constants and can be continuously monitored and adaptively stopped. We have done so by combining tools from testing by betting with RKHS theory, while avoiding assuming uniform boundedness of the Stein reproducing kernel. We have proved the validity of the test, as well as exponential growth of the wealth process under the alternative and mild regularity conditions. Our experiments have exhibited the empirical performance of the test in a variety of scenarios.

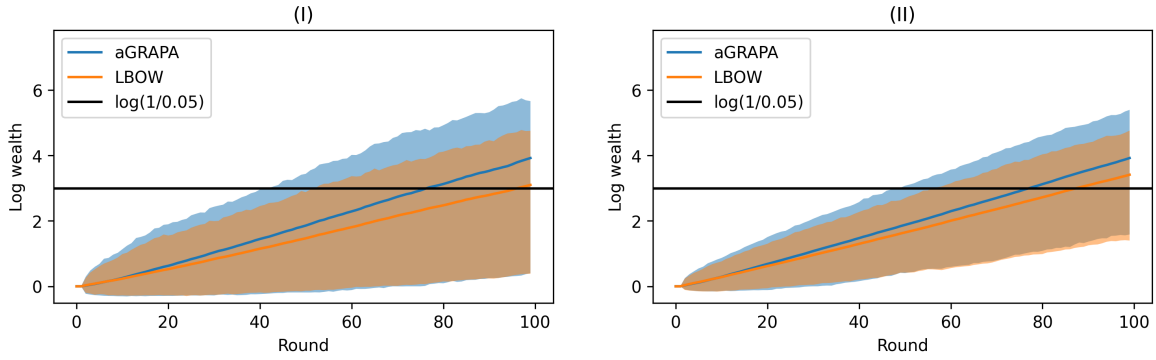

Figure 1: Average logarithmic wealth alongside 95% empirical confidence intervals for 1000 simulations under the alternative for (I) the Gaussian distribution, (II) the intractable model. We highlight the exponential growth of the wealth.

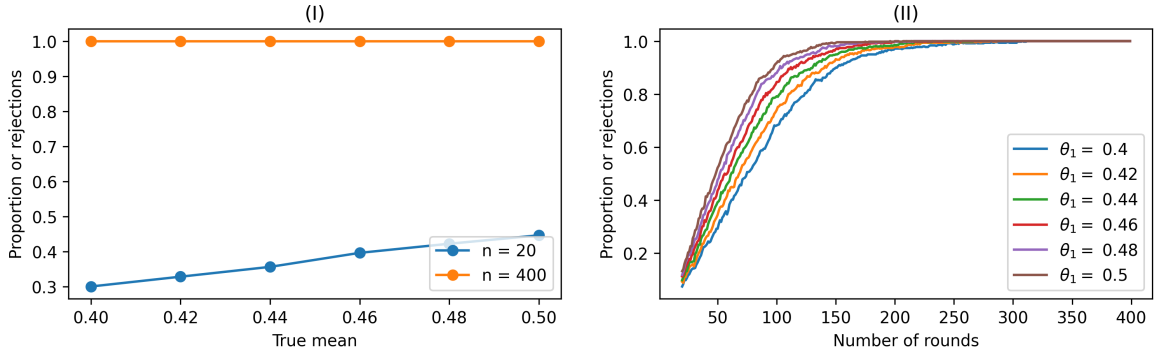

Figure 2: Proportion of rejections for the Gaussian distribution considered in Section 6 under the alternatives $\theta_1 \in \{0.40, 0.42, 0.44, 0.46, 0.48, 0.50\}$ for (I) the (classical) batch setting kernelized Stein discrepancy with sample size n, (II) the (proposed) sequential kernelized Stein discrepancy. The (proposed) sequential test always ends up rejecting the null hypotheses, while the batch test will not do so if the original sample size is too small.

In this contribution, we have presented a novel martingale construction that does neither exploit nor need uniform boundedness of the kernel. While the theory presented here has been developed for and motivated by the Stein kernel, such a martingale construction may be exploited by any kernel that is either unbounded or uniformly bounded by a constant that is too loose. This opens the door to develop sequential two sample and independence tests with kernels that are currently unexplored, which conforms an exciting direction of research.

Acknowledgements

DMT gratefully acknowledges that the project that gave rise to these results received the support of a fellowship from 'la Caixa' Foundation (ID 100010434). The fellowship code is LCF/BQ/EU22/11930075. AR was funded by NSF grant DMS-2310718. The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions.

References

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169.

Agostinelli, S., Allison, J., Amako, K. a., Apostolakis, J., Araujo, H., Arce, P., Asai, M., Axen, D., Banerjee, S., Barrand, G., et al. (2003). Geant4—a simulation toolkit. *Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250–303.

Barp, A., Simon-Gabriel, C.-J., Girolami, M., and Mackey, L. (2024). Targeted separation and convergence with kernel discrepancies. *Journal of Machine Learning Research*, 25(378):1–50.

Baum, J., Kanagawa, H., and Gretton, A. (2023). A kernel stein test of goodness of fit for sequential models. In *International Conference on Machine Learning*, pages 1936–1953. PMLR.

Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 48(3):259–279.

Bolstad, W. M. and Curran, J. M. (2016). *Introduction to Bayesian statistics*. John Wiley & Sons.

Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*, volume 149. Springer Science & Business Media.

Canu, S. and Smola, A. (2006). Kernel methods and the exponential family. *Neurocomputing*, 69(7-9):714–720.

Chekalina, V., Orlova, E., Ratnikov, F., Ulyanov, D., Ustyuzhanin, A., and Zakharov, E. (2019). Generative models for fast calorimeter simulation: the lhcb case. In *EPJ Web of Conferences*, volume 214, page 02034. EDP Sciences.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In *International Conference on Machine Learning*, pages 2606–2615. PMLR.

Clifford, P. (1990). Markov random fields in statistics. *Disorder in physical systems: A volume in honour of John M. Hammersley*, pages 19–32.

D'Agostino, R. B. (2017). *Goodness-of-fit-techniques*. Routledge.

Du, Y. and Mordatch, I. (2019). Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32.

Durbin, J. (1975). Kolmogorov-smirnov tests when parameters are estimated with applications to tests of exponentiality and tests on spacings. *Biometrika*, 62(1):5–22.

Fernandez, T., Rivera, N., Xu, W., and Gretton, A. (2020). Kernelized stein discrepancy tests of goodness-of-fit for time-to-event data. In *International Conference on Machine Learning*, pages 3112–3122. PMLR.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741.

González-Manteiga, W. and Crujeiras, R. M. (2011). A general view of the goodness-of-fit tests for statistical models. In *Modern Mathematical Tools and Techniques in Capturing Complexity*, pages 3–16. Springer.

Gorham, J. and Mackey, L. (2015). Measuring sample quality with Stein's method. *Advances in Neural Information Processing Systems*, 28.

Gorham, J. and Mackey, L. (2017). Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2006). A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems*, 19.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2007). A kernel statistical test of independence. *Advances in Neural Information Processing Systems*, 20.

Grünwald, P., de Heide, R., and Koolen, W. M. (2024). Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology (with discussion)*.

Hinton, G. (2010). A practical guide to training restricted boltzmann machines. *Momentum*, 9(1):926.

Huang, D., Bharti, A., Souza, A., Acerbi, L., and Kaski, S. (2023). Learning robust statistics for simulation-based inference under model misspecification. *Advances in Neural Information Processing Systems*, 36:7289–7310.

Hyvärinen, A. and Dayan, P. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4).

Kanagawa, H., Barp, A., Gretton, A., and Mackey, L. (2022). Controlling moments with kernel stein discrepancies. *arXiv preprint arXiv:2211.05408*.

Key, O., Gretton, A., Briol, F.-X., and Fernandez, T. (2021). Composite goodness-of-fit tests with kernels. *arXiv preprint arXiv:2111.10275*.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F., et al. (2006). A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Lehmann, E. L., Romano, J. P., et al. (1986). *Testing statistical hypotheses*, volume 3. Springer.

Liu, Q., Lee, J., and Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pages 276–284. PMLR.

Liu, S., Kanamori, T., Jitkrittum, W., and Chen, Y. (2019). Fisher efficient inference of intractable models. *Advances in Neural Information Processing Systems*, 32.

Masserano, L., Dorigo, T., Izbicki, R., Kuusela, M., and Lee, A. (2023). Simulator-based inference with waldo: Confidence regions by leveraging prediction algorithms and posterior estimators for inverse problems. In *International Conference on Artificial Intelligence and Statistics*, pages 2960–2974. PMLR.

Matsubara, T., Knoblauch, J., Briol, F.-X., and Oates, C. J. (2022). Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):997–1022.

Podkopaev, A., Blöbaum, P., Kasiviswanathan, S., and Ramdas, A. (2023). Sequential kernelized independence testing. In *International Conference on Machine Learning*, pages 27957–27993. PMLR.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601.

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*.

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. M. (2022). Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 141:83–109.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press.

Schrab, A., Guedj, B., and Gretton, A. (2022). Ksd aggregated goodness-of-fit test. *Advances in Neural Information Processing Systems*, 35.

Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(2):407–431.

Shafer, G., Shen, A., Vereshchagin, N., and Vovk, V. (2011). Test martingales, Bayes factors and p-values. *Statistical Science*.

Shafer, G. and Vovk, V. (2019). *Game-theoretic foundations for probability and finance*, volume 455. John Wiley & Sons.

Shekhar, S. and Ramdas, A. (2023). Nonparametric two-sample testing by betting. *IEEE Transactions on Information Theory*.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, volume 6, pages 583–603. University of California Press.

Ville, J. (1939). *Etude critique de la notion de collectif*. Gauthier-Villars Paris.

Waudby-Smith, I. and Ramdas, A. (2024). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27.

Wynne, G., Kasprzak, M. J., and Duncan, A. B. (2025). A fourier representation of kernel stein discrepancy with application to goodness-of-fit tests for measures on infinite dimensional hilbert spaces. *Bernoulli*, 31(2):868–893.

Xu, W. and Matsuda, T. (2020). A stein goodness-of-fit test for directional distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 320–330. PMLR.

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2015). Graphical models via univariate exponential family distributions. *The Journal of Machine Learning Research*, 16(1):3813–3847.

Yang, J., Liu, Q., Rao, V., and Neville, J. (2018). Goodness-of-fit testing for discrete distributions via stein discrepancy. In *International Conference on Machine Learning*, pages 5561–5570. PMLR.

Yang, J., Rao, V., and Neville, J. (2019). A stein–papangelou goodness-of-fit test for point processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 226–235. PMLR.

Zhou, Z. and Liu, W. (2024). Sequential kernel goodness-of-fit testing. In *Forty-first International Conference on Machine Learning*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Not Applicable

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

A ADDITIONAL EXPERIMENTS

A.1 Logarithmic wealth process of a Gaussian-Bernoulli restricted Boltzmann machine

In Section 6, we displayed the performance of the proposed test for the Gaussian and intractable models. The performance of the proposed test for the remaining model introduced in Section 5, a Gaussian-Bernoulli restricted Boltzmann machine, is now exhibited in Figure 3. In particular, we take $d_h = 10$, and $d = 50$. We sample from it using Gibbs sampling with a burn-in of 1000 iterations. Under the null, we take $b = 0$, $c = 0$, and matrix B such that B_{ij} is one if visible node i is connected to hidden node j, and zero otherwise, with each hidden node connected to five visible nodes. We consider two different alternatives. For one of them, each entry of B is shifted by 0.5. For the other one, $b = 1$.

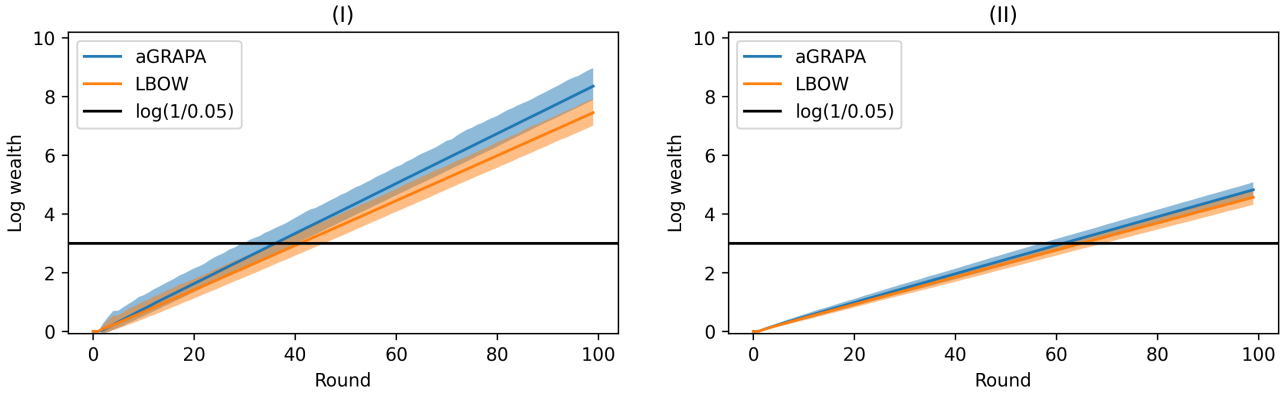

Figure 3: Average logarithmic wealth alongside 95% empirical confidence intervals for 1000 simulations under the alternative for (I) the restricted Boltzmann machine with shifted B, (II) the restricted Boltzmann machine with bias $b = 1$. We highlight once again the exponential growth of the wealth processes.

A.2 Type-I error control

In Section 6, we illustrated the empirical power of the proposed test. We devote this appendix to displaying the empirical type-I error control of the proposed test. Figure 4 exhibits the wealth processes for the examples considered in Section 6 under the null. We highlight that the wealth processes do not cross the threshold $1/0.05$, and hence the nulls are not rejected.

We note that Figure 4 exhibits the anytime validity of the tests until round 100 (the same number of rounds as considered in Section 6). However, we can verify anytime validity without committing to a sample size for those simulations with known normalizing constants. More specifically, let P be the distribution under the null, and Q any other distribution such that dP/dQ is known and we can sample from Q. Let τ denote the random stopping time. Since $\mathbb{P}_P(\tau < \infty) = E_P[1(\tau < \infty)] = E_Q[dP/dQ1(\tau < \infty)]$, instead of sampling from P and calculating $\mathbb{P}_P(\tau < \infty)$ by Monte-Carlo, we can instead sample from Q and calculate $E_Q[dP/dQ1(\tau < \infty)]$ by Monte-Carlo. For $Q \neq P$, the procedure will stop relatively fast (since the wealth grows exponentially fast under the alternative), and we can average these likelihood ratios at the stopping time to estimate the type 1 error. For simulations in which we know the normalizing constants (even though we may not want to use them in order to test our methods), this importance sampling technique for estimating the type-1 error works well. For the Gaussian case considered in Section 6 (i.e., P is a standard Gaussian distribution), taking Q to be a Gaussian distribution with unit variance centered at 0.5 leads to an estimated $\mathbb{P}_P(\tau < \infty)$ of 0.0006 for $\alpha = 0.1$ (showing, again, the empirical anytime validity of the proposed test).

A.3 Comparison to batch setting

We illustrated in Section 6 some of the advantages of the sequential test over the batch test. We emphasize that the proposed test is anytime valid, counting with the subsequent substantial advantages that have been discussed

in the main body of the work. Thus, it is expected to be outperformed in the batch setting by algorithms that are tailored to fixed sample sizes. Nonetheless, it is of interest to explore how the proposed test compares with the more classical kernel Stein discrepancy test. We present in this appendix the empirical performance of both the sequential and fixed sample size algorithms.

Figure 5 exhibits the proportion of rejections for the Gaussian distribution considered in Section 6 with different alternatives. As expected, the classical batch setting test outperforms the proposed test for fixed sample sizes, with the proportions of rejections of the former being always larger than those of the latter. Nonetheless, the figure illustrates the convenience of employing anytime valid tests. If using the classical kernel Stein discrepancy test, 50 observations suffice to reject the null hypothesis for every sample in the right plot, but they only allow to reject for around 80% of the samples in the left plot. The null hypothesis cannot be ever rejected for the samples encompassed in the remaining 20%: based on the interpretation of p-values, it is not possible to keep gathering evidence to rerun the test. In stark contrast, all the null hypotheses are eventually rejected by the proposed test, being always able to collect more data and keep running the experiment with anytime validity.

We would like to emphasize the lack of anytime validity of the classical, batch setting test. Figure 6 displays the proportion of rejections for both the proposed sequential test, and the batch test run sequentially (once every time a new observation is observed). Not surprisingly, the batch test rejection rates go well above the desired $\alpha = 0.05$ type-I error. This example illustrates the lack of anytime validity of classical test.

A.4 Composite null hypotheses

Section 4.2 presented the extension of the proposed test to composite null hypotheses. We devote this appendix to illustrating the validity and power of such an extension. Let the null hypothesis be $H_0 : Q \in \mathcal{P}$ against $H_1 : Q \notin \mathcal{P}$. Throughout, the null hypotheses will be combinations of the Gaussian-Bernoulli restricted Boltzmann machines (RBMs) from Section 5.

In particular (and analogously to Appendix A.1), we take $d_h = 10$, and $d = 50$. We sample from such RBMs using Gibbs sampling with a burn-in of 1000 iterations. Throughout, Q is taken as an RBM with $b = 0$, $c = 0$, and matrix B such that B_{ij} is one if visible node i is connected to hidden node j, and zero otherwise, with each hidden node connected to five visible nodes. For two different experiments, we consider the two null hypotheses $\mathcal{P} = \{Q, Q_{a,1}\}$ and $\mathcal{P} = \{Q_{a,1}, Q_{a,2}\}$, where $Q_{a,1}$ is an RBM with each entry of B is shifted by 0.5 and $Q_{a,2}$ is taken with $b = 1$. Figure 7 displays the logarithmic wealth processes for these two scenarios. Note that the logarithmic wealth process never crosses the threshold $\log(1/0.05)$ in the first case (showing the empirical anytime validity of the test), while the latter case shows comparable power to those experiments presented in Section A.1.

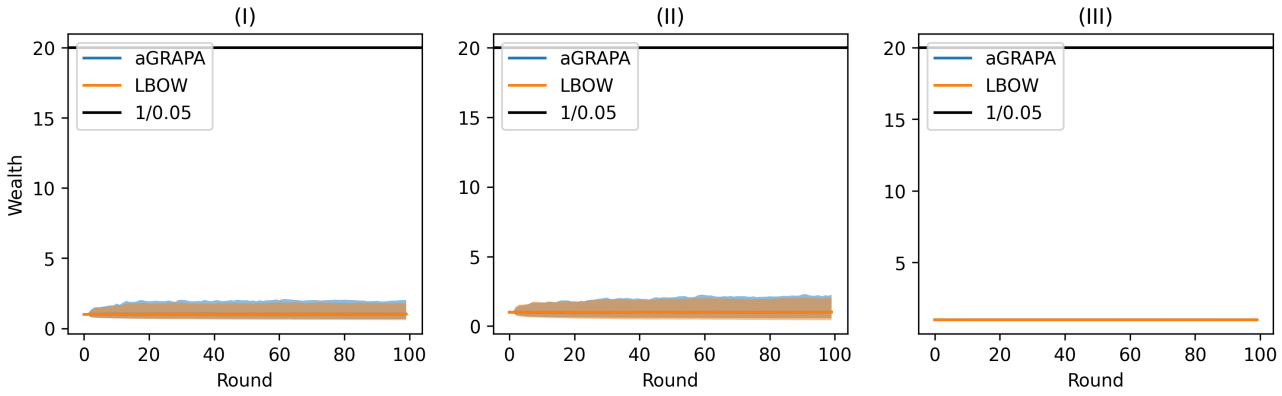

Figure 4: Average wealth alongside 95% empirical confidence intervals for 1000 simulations under the null for (I) the Gaussian distribution, (II) the intractable model, (III) the restricted Boltzmann machine. We emphasize that the wealth processes do not cross the threshold 1/0.05, and hence the nulls are not rejected, showing the empirical type-I error control.

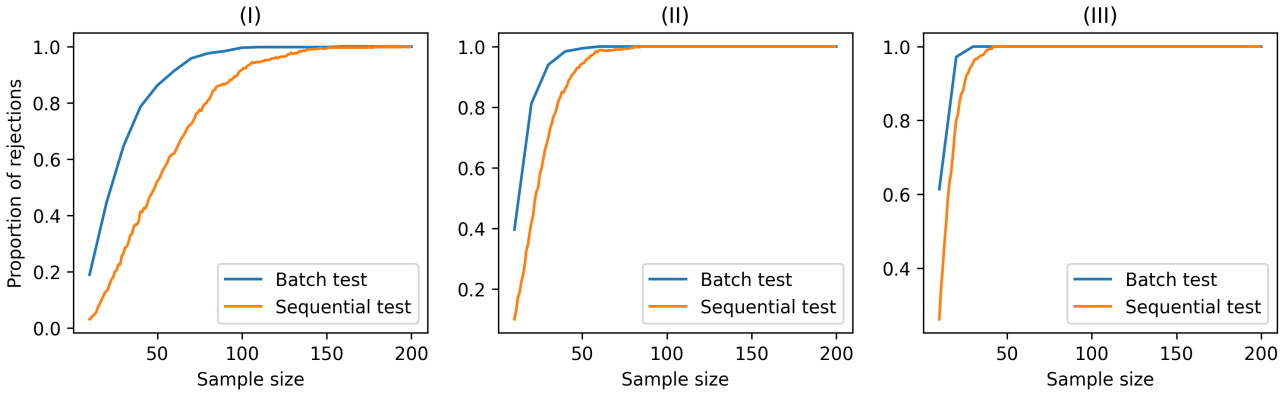

Figure 5: Proportion of rejections for 500 simulations for the Gaussian distribution considered in Section 6 under three different alternatives with (I) $\theta = 0.5$, (II) $\theta = 0.75$, (III) $\theta = 1$. We emphasize that the (proposed) sequential test always ends up rejecting the null hypotheses, while the batch test will not do so if the original sample size is too small.

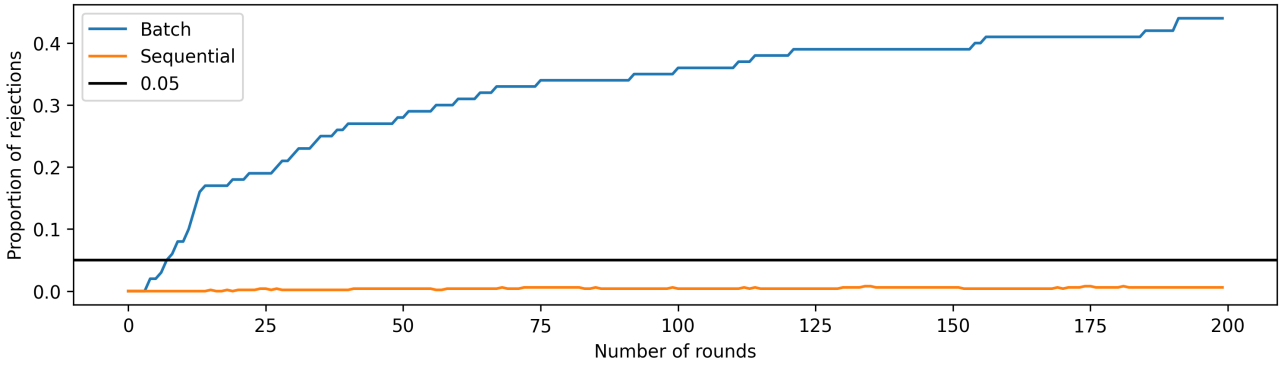

Figure 6: Proportion of rejections for 100 simulations for the Gaussian distribution considered in Section 6 under the alternative for the proposed sequential test, and the batch test run sequentially (once every time a new observation is observed). While the sequential test preserves anytime valid guarantees, the 'sequentialized' batch test lacks type-I error control.

A.5 On the tightness of the bounds and statistical power

In Section 5, we carefully exhibited how to derive sensible bounds M_p for different families of distributions. Here, we explore how the tightness of the bounds affects the statistical power of the test. Figure 8 displays the wealth processes for the Gaussian distribution from Section 5 under three different alternatives, using the original bound obtained in Section 5 alongside looser bounds. We highlight that all logarithmic wealth processes are linear (i.e., the wealth growth is exponential), but looser bounds lead to smaller slopes.

A.6 Empirical Verification of the Lower Bound of the Wealth

In Section 4, we stressed that the stopping time τ of the proposed test (i.e. τ is the smallest t verifying $\mathcal{K}_t \geq 1/\alpha$) is roughly upper bounded by $\log(1/\alpha)/r^*$ (where r^* is defined as in Theorem 2). We devote this section to exhibit the empirical validity of this claim. Note that, while r^* is unknown in practice, it can be easily estimated via a Monte Carlo approach when we have access to the ground truth distribution (as it is the case in our experimental settings), given that it only depends on the first and second moments of the h_p and M_p evaluations.

Figure 9 displays the stopping times τ of the proposed test for the Gaussian setting (presented in Section 5.2 and

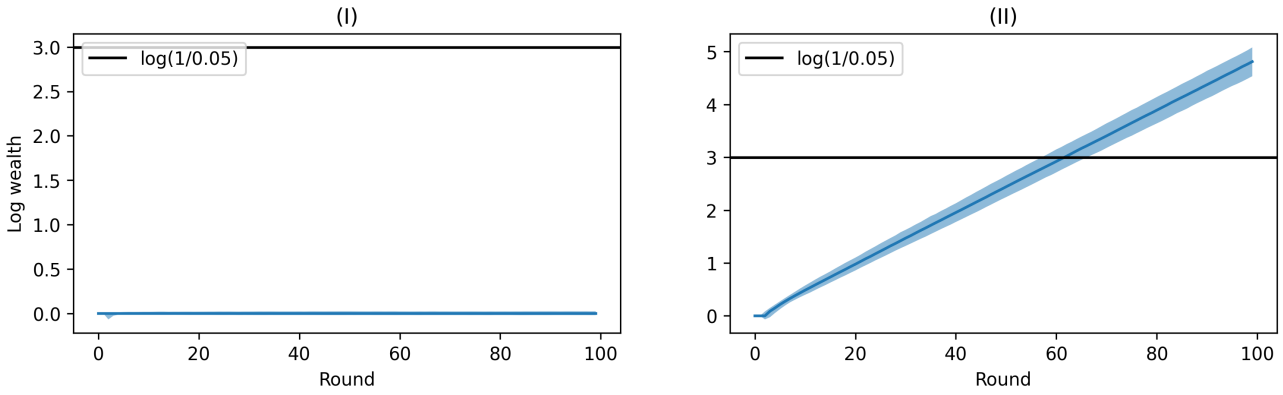

Figure 7: Average logarithmic wealth for 1000 simulations for the composite null hypothesis (I) $\mathcal{P} = \{Q, Q_{a,1}\}$, (II) $\mathcal{P} = \{Q_{a,1}, Q_{a,2}\}$, where Q is the RBM defined in Section 5 with $b = 0$, $c = 0$, and matrix B such that B_{ij} is one if visible node i is connected to hidden node j, and zero otherwise, with each hidden node connected to five visible nodes; $Q_{a,1}$ is the same RBM with each entry of B is shifted by 0.5, and $Q_{a,2}$ is taken with $b = 1$.

Section 6) as a function of r^*, both for the LBOW and aGRAPA strategies. We take $\theta_0 = 0$ under the null, and a range of θ_1 under different alternatives. Intuitively, the larger the distance between the means θ_0 and θ_1 is, the larger the theoretical quantities r^* become. Thus, this setting provides a range of r^* for which we can study the empirical stopping times τ. Figure 9 shows that the average stopping time curve, as well as its empirical 95% confidence interval, are empirically dominated by the upper bound $\log(1/\alpha)/r^*$ (and once more, we note that the aGRAPA strategy empirically outperforms the LBOW strategy).

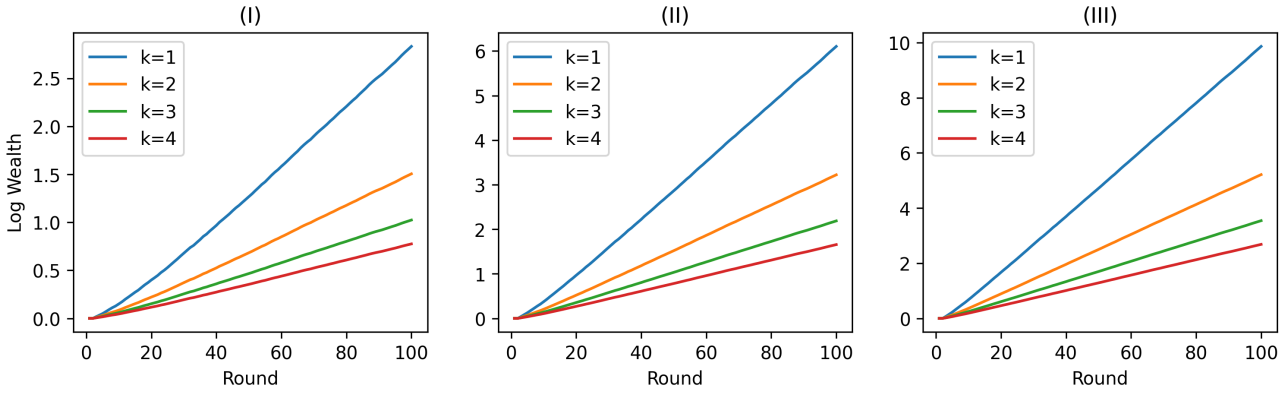

Figure 8: Average logarithmic wealth for 1000 simulations for the Gaussian distribution when using bounds of the form $k \cdot M(X)$, where $M(X)$ is the original bound derived in Section 5 and $k \in \{1, 2, 3, 4\}$, under the alternatives (I) $\theta_1 = 0.5$, (II) $\theta_1 = 0.75$, (III) $\theta_1 = 1$. It can be clearly seen that all the logarithmic wealth processes are linear (i.e., the wealth growth is exponential), but higher multiplicative factors k (i.e., looser bounds) have smaller slopes.

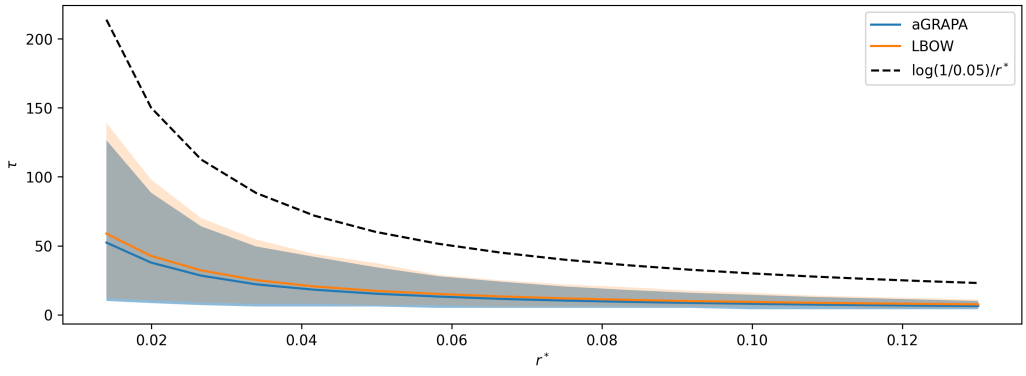

Figure 9: Average empirical stopping values τ, alongside their 95% empirical confidence intervals for 1000 simulations, as a function of r^* (defined in Theorem 2) in the Gaussian setting, with $\alpha = 0.05$, across 500 simulations, for the LBOW and aGRAPA strategies. We emphasize that all the stopping times are dominated by the approximate upper bound $log(1/\alpha)/r^*$.

B Comparison to related work

We present in this appendix a comprehensive comparison of our contribution and that of Zhou and Liu (2024), which also proposed a sequential KSD goodness-of-fit test that is based on a supermartingale construction and Ville's inequality.

The primary distinction between the contributions lies in the uniform boundedness assumption, which is adopted in Zhou and Liu (2024) but deliberately avoided in our work. Specifically, their work assumed that the scores are uniformly bounded by 1 (Zhou and Liu, 2024, Assumption 2.4). We would like to emphasize that this is, in principle, a rather restrictive assumption. For instance, even the simple case of a Gaussian distribution with an IMQ kernel does not satisfy this assumption. Thus, our method is strictly more general and can accommodate more scenarios, e.g., the three experiments presented in Section 6 could not have been addressed using their procedure.

In principle, there could be other alternatives to work around unbounded score functions. For instance, one could work with tilted base kernels that make the Stein kernel bounded even if the score function is not bounded (Barp et al., 2024, Theorem 7). Nonetheless, current results such as Barp et al. (2024, Theorem 7) still require bounds to 'tilt' the base kernel, but do not provide such bounds (the authors' goal is solely to prove that boundedness of the Stein kernel may be assumed without loss of generality in their setting, without further interest in the bound itself). One of the main contributions of our paper is to show how to obtain tight workable bounds, and illustrate that such derivations yield powerful tests. In order to potentially use Barp et al. (2024, Theorem 7) to obtain sequential tests, one would have to carefully study the actual bounds derived from such a theorem. This would eventually lead to an analysis very similar to the one presented throughout this contribution.

The lack of uniform boundedness also conveys a number of deep theoretical challenges. The theoretical foundation of Zhou and Liu (2024) is analogous to the one presented in Podkopaev et al. (2023). See e.g. Zhou and Liu (2024, Theorem 2.5) and Podkopaev et al. (2023, Theorem 2.4), alongside their proofs. However, the lack of uniform boundedness makes such proofs break in (at least) two different places: (i) after applying the Cauchy-Schwarz inequality, see Zhou and Liu (2024, Equation (47)), (ii) the analysis of the Online Newton Step (ONS) strategy for selecting betting fractions. In order to circumvent the above challenges, this contribution replaces the ONS strategy for aGRAPA/LBOW strategies and presents novel theoretical guarantees, which are exhibited in Appendix C.3.[5] Lastly, we have experimentally found that the aGRAPA and LBOW strategies proposed in our contribution yield substantially more powerful tests than the ONS strategy proposed in Zhou and Liu (2024) in all the cases considered. As exhibited in Figure 10, the average log wealth at round 100 for aGRAPA roughly doubles the log wealth achieved by the ONS strategy for all the experiments.

[5]Some of the arguments presented in Appendix C.3 may be of independent interest. To the best of our knowledge, no other contribution in the 'testing by betting' literature exploits outcomes that are lower bounded but not upper bounded.

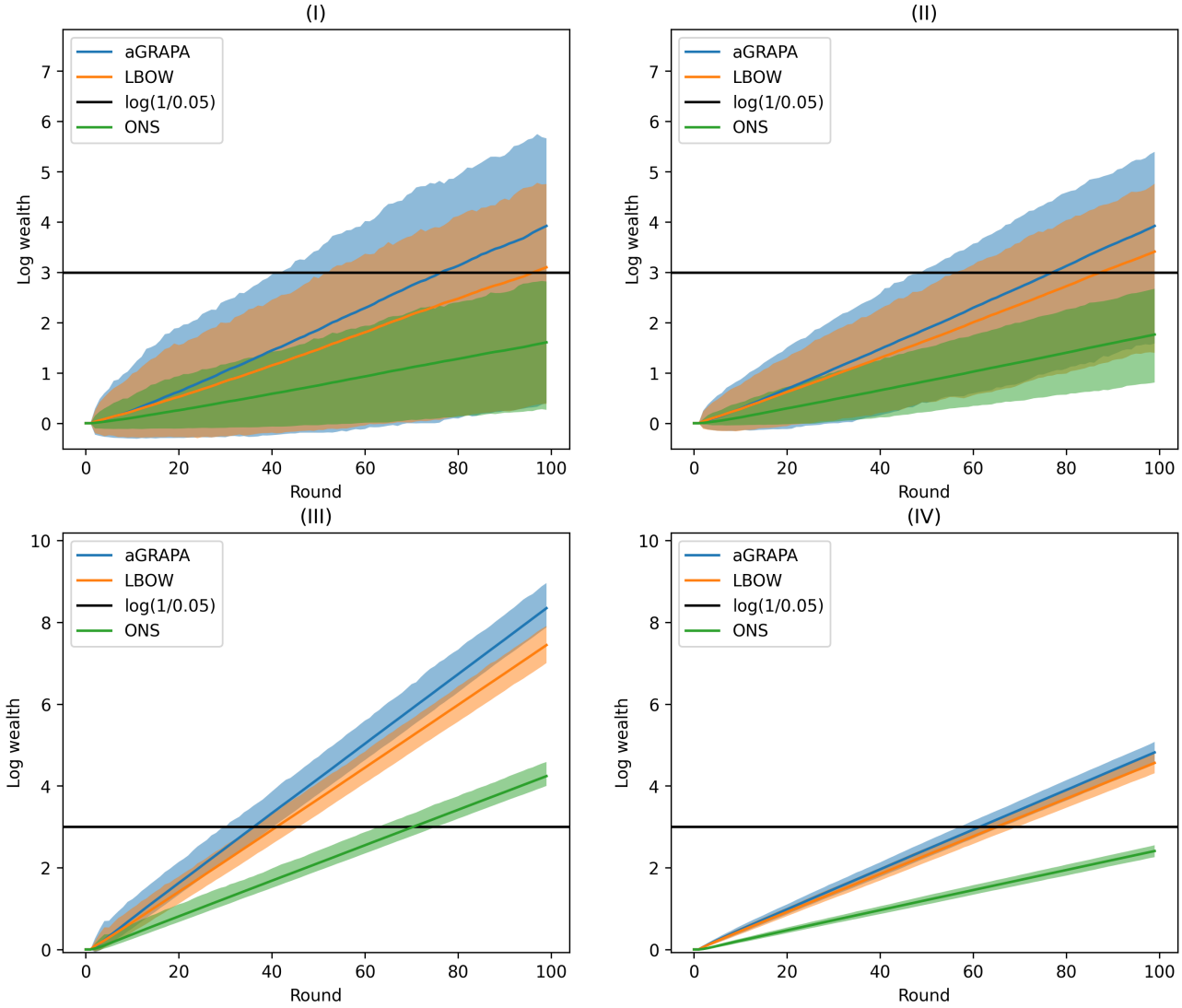

Figure 10: Average logarithmic wealth alongside 95% empirical confidence intervals for 1000 simulations under the alternative for (I) the Gaussian distribution, (II) the intractable model, (III) the restricted Boltzmann machine with shifted B, (IV) the restricted Boltzmann machine with bias $b = 1$. We highlight that the logarithmic wealth achieved by ONS is roughly doubled by that obtained by the aGRAPA strategy.

C PROOFS

C.1 Notation

Throughout, we denote

$$g_i(x) = \frac{1}{\frac{1}{t-1}\sum_{k=1}^{i-1} M_p(X_k)} \left(\frac{1}{i-1} \sum_{k=1}^{i-1} h_p(X_k, x) \right), \quad g^*(x) = \frac{1}{\mathbb{E}[M_p(X)]} \mathbb{E}[h_p(X, x)],$$
$$f_i(x) = \frac{1}{i-1} \sum_{k=1}^{i-1} h_p(X_k, x), \quad f^*(x) = \mathbb{E}[h_p(X, x)],$$
$$M_p^i = \frac{1}{i-1} \sum_{k=1}^{t-1} M_p(X_k), \quad M_p = E[M_p(X)].$$

C.2 Preliminary Results

For completeness, we enunciate the following theorems, which will be exploited in subsequent proofs.

Theorem 4 (SLLN for Banach-valued random variables. Bosq (2000, Theorem 2.4).)**.** *Let B be a separable Banach space with norm $\|\cdot\|_B$. Let $(\chi_t)_{t\geq 1}$ be a sequence of i.i.d. integrable B-valued random variables. Then*

$$\| \frac{1}{t} \sum_{i=1}^{t} \chi_i - \mathbb{E}[\chi] \|_B \overset{a.s.}{\to} 0.$$

Theorem 5 (Ville's inequality.)**.** *Let $(S_t)_{t\geq 0}$ be a nonnegative supermartingale process adapted to a filtration $\{\mathcal{F}_t \geq 0\}$. It holds that*

$$\mathbb{P}(\exists t \geq 1 : S_t \geq x) \leq \frac{\mathbb{E}[S_0]}{x}$$

for any $x > 0$.

C.3 Auxiliary Results

We present here two results that Theorem 2 builds on.

Proposition 1. *If $\mathbb{E}[\sqrt{h_p(X,X)}] < \infty$ and $M_p < \infty$, then*

$$\frac{1}{t} \sum_{i=1}^{t} |g_i(X_i) - g^*(X_i)| \overset{a.s.}{\to} 0, \quad \frac{1}{t} \sum_{i=1}^{t} g_i(X_i) \overset{a.s.}{\to} \mathbb{E}[g^*(X)].$$

Proof. Without loss of generality, we can assume that $M_p > 0$. We first highlight that, based on $0 < \mathbb{E}[\sqrt{h_p(X,X)}] < \infty$,

- $\mathbb{E}[|f^*(X)|]$ is finite, as $\mathbb{E}[|f^*(X)|] = \mathbb{E}_{X,X'}[|h_p(X', X)|] \leq \mathbb{E}_{X,X'} \left[\|\xi_p(\cdot, X)\|_{\mathcal{H}^d} \|\xi_p(\cdot, X')\|_{\mathcal{H}^d} \right] = \mathbb{E}_X \left[\|\xi_p(\cdot, X)\|_{\mathcal{H}^d} \right]^2 = \mathbb{E}[\sqrt{h_p(X,X)}]^2 < \infty$,
- $\mathbb{E}\left[\|\xi_p(\cdot, X)\|_{\mathcal{H}^d} \right] = \mathbb{E}[\sqrt{h_p(X,X)}]$ is finite.
- $\mathbb{E}[|g^*(X)|]$ is finite, as $g^*(x) = f^*(x)/M_p$, $\mathbb{E}[|f^*(X)|]$ is also finite and $M_p > 0$.

We now note that $\frac{1}{t}\sum_{i=1}^{t} |g_i(X_i) - g^*(X_i)| \overset{a.s.}{\to} 0$ implies $\frac{1}{t}\sum_{i=1}^{t} g_i(X_i) \overset{a.s.}{\to} \mathbb{E}[g^*(X)]$, i.e., $\frac{1}{t}\sum_{i=1}^{t} g_i(X_i) - \mathbb{E}[g^*(X)] \overset{a.s.}{\to} 0$. To see this, we decompose

$$\frac{1}{t} \sum_{i=1}^{t} g_i(X_i) - \mathbb{E}[g^*(X)] = \left(\frac{1}{t} \sum_{i=1}^{t} g_i(X_i) - \frac{1}{t} \sum_{i=1}^{t} g^*(X_i) \right) + \left(\frac{1}{t} \sum_{i=1}^{t} g^*(X_i) - \mathbb{E}[g^*(X)] \right),$$

and we highlight that the second term converges to zero almost surely in view of the strong law of large numbers (SLLN) and $\mathbb{E}[|g^*(X)|] < \infty$. Further, $\frac{1}{t}\sum_{i=1}^{t}|g_i(X_i) - g^*(X_i)| \overset{a.s.}{\to} 0$ implies $\frac{1}{t}\sum_{i=1}^{t} g_i(X_i) - \frac{1}{t}\sum_{i=1}^{t} g^*(X_i) \overset{a.s.}{\to} 0$.

Hence, it remains to prove that

$$\frac{1}{t}\sum_{i=1}^{t}|g_i(X_i) - g^*(X_i)| \overset{a.s.}{\to} 0.$$

That is, for any given $\epsilon > 0$ and $\delta > 0$, there exists $N \in \mathbb{N}$ such that

$$\mathbb{P}\left(\sup_{t\geq N}\frac{1}{t}\sum_{i=1}^{t}|g_i(X_i) - g^*(X_i)| > \epsilon\right) \leq \delta.$$

In order to prove the result, we are going to decompose $\frac{1}{t}\sum_{i=1}^{t}|g_i(X_i) - g^*(X_i)|$, and then combine the scalar-valued SLLN and the Banach space-valued SLLN. We will finish the proof by taking a union bound over the different terms.

Introducing the SLLN terms and the probabilistic bounds: By the scalar valued SLLN and the finiteness of $M_p > 0$, $\mathbb{E}[|f^*(X)|]$, and $\mathbb{E}\left[\|\xi_p(\cdot, X)\|_{\mathcal{H}^d}\right]$, we have that

- $M_p^i \overset{a.s.}{\to} M_p$, and so $\frac{1}{M_p^i} \overset{a.s.}{\to} \frac{1}{M_p}$ (given that $M_p \neq 0$),
- $\frac{1}{t}\sum_{i=1}^{t}|f^*(X_i)| \overset{a.s.}{\to} \mathbb{E}[|f^*(X)|]$,
- $\frac{1}{t}\sum_{i=1}^{t}\|\xi_p(\cdot, X_i)\|_{\mathcal{H}^d} \overset{a.s.}{\to} \mathbb{E}\left[\|\xi_p(\cdot, X)\|_{\mathcal{H}^d}\right]$.

From the Banach space-valued SLLN (Theorem 4) and the finiteness of $\mathbb{E}\left[\|\xi_p(\cdot, X)\|_{\mathcal{H}^d}\right]$, it also follows that

- $\left\|\frac{1}{i-1}\sum_{k=1}^{i-1}\xi_p(\cdot, X_k) - \mathbb{E}[\xi_p(\cdot, X)]\right\|_{\mathcal{H}^d} \overset{a.s.}{\to} 0$.

Hence there exist $B > 0$ and $N_1 \in \mathbb{N}$ such that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\geq N_1}\left|\frac{1}{M_p^i}\right| > B\right) \leq \frac{\delta}{10},\\
&\mathbb{P}\left(\sup_{t\geq N_1}\frac{1}{t}\sum_{i=1}^{t}|f^*(X_i)| > B\right) \leq \frac{\delta}{10},\\
&\mathbb{P}\left(\sup_{t\geq N_1}\frac{1}{t}\sum_{i=1}^{t}\|\xi_p(\cdot, X_i)\|_{\mathcal{H}^d} > B\right) \leq \frac{\delta}{10},\\
&\mathbb{P}\left(\sup_{t\geq N_1}\left|\frac{1}{M_p^i} - \frac{1}{M_p}\right| > \frac{\epsilon}{4B}\right) \leq \frac{\delta}{10},\\
&\mathbb{P}\left(\sup_{i\geq N_1}\left\|\frac{1}{i-1}\sum_{k=1}^{i-1}\xi_p(\cdot, X_k) - \mathbb{E}[\xi_p(\cdot, X)]\right\|_{\mathcal{H}^d} > \frac{\epsilon}{4B^2}\right) \leq \frac{\delta}{10}.
\end{aligned}$$

Note that $\mathbb{P}\left(\sup_{t\geq N_1}\frac{1}{t}\sum_{i=1}^{t}\|\xi_p(\cdot, X_i)\|_{\mathcal{H}^d} > B\right) \leq \frac{\delta}{10}$ and $\mathbb{P}\left(\sup_{t\geq N_1}\frac{1}{t}\sum_{i=1}^{t}|f^*(X_i)| > B\right) \leq \frac{\delta}{10}$ imply that $\mathbb{P}\left(\sup_{t\geq 2N_1}\frac{1}{t-N_1}\sum_{i=N_1}^{t}\|\xi_p(\cdot, X_i)\|_{\mathcal{H}^d} > B\right) \leq \frac{\delta}{10}$ and $\mathbb{P}\left(\sup_{t\geq 2N_1}\frac{1}{t-N_1}\sum_{i=N_1}^{t}|f^*(X_i)| > B\right) \leq \frac{\delta}{10}$, given that the data are iid.

Combining the probabilistic bounds: We start by noting that

$$\frac{1}{t}\sum_{i=1}^{t}|g_i(X_i) - g^*(X_i)| = \frac{1}{t}\sum_{i=1}^{N_1-1}|g_i(X_i) - g^*(X_i)| + \frac{1}{t}\sum_{i=N_1}^{t}|g_i(X_i) - g^*(X_i)|.$$

Consider the sequence of random variables $Y_t = \frac{1}{t}\left(\sum_{i=1}^{N_1-1} |g_i(X_i) - g^*(X_i)|\right)$. Clearly, $Y_t \overset{a.s.}{\to} 0$. Thus there exists N_2 such that $\sup_{t\geq N_2} Y_t \leq \frac{\epsilon}{2}$ with probability at least $1-\delta/2$.

Moreover,

$$\begin{aligned}
\frac{1}{t}\sum_{i=N_1}^{t} |g_i(X_i) - g^*(X_i)| &= \frac{1}{t}\sum_{i=N_1}^{t} \left|\frac{1}{M_p^i} f_i(X_i) - \frac{1}{M_p} f^*(X_i)\right| \\
&= \frac{1}{t}\sum_{i=N_1}^{t} \left|\frac{1}{M_p^i} f_i(X_i) - \frac{1}{M_p^i} f^*(X_i) + \frac{1}{M_p^i} f^*(X_i) - \frac{1}{M_p} f^*(X_i)\right| \\
&\leq \underbrace{\frac{1}{t}\sum_{i=N_1}^{t} \left|\frac{1}{M_p^i} f_i(X_i) - \frac{1}{M_p^i} f^*(X_i)\right|}_{(I)} + \\
&+ \underbrace{\frac{1}{t}\sum_{i=N_1}^{t} \left|\frac{1}{M_p^i} f^*(X_i) - \frac{1}{M_p} f^*(X_i)\right|}_{(II)}.
\end{aligned}$$

We now handle terms (I) and (II) separately. For term (II), we observe that

$$\begin{aligned}
\sup_{t\geq 2N_1} \frac{1}{t}\sum_{i=N_1}^{t} \left|\frac{1}{M_p^i} f^*(X_i) - \frac{1}{M_p} f^*(X_i)\right| &= \sup_{t\geq 2N_1} \frac{1}{t}\sum_{i=N_1}^{t} \left|\frac{1}{M_p^i} - \frac{1}{M_p}\right| |f^*(X_i)| \\
&\leq \left(\sup_{t\geq 2N_1} \left|\frac{1}{M_p^i} - \frac{1}{M_p}\right|\right)\left(\sup_{t\geq 2N_1} \frac{1}{t}\sum_{i=N_1}^{t} |f^*(X_i)|\right) \\
&\leq \left(\sup_{t\geq N_1} \left|\frac{1}{M_p^i} - \frac{1}{M_p}\right|\right)\left(\sup_{t\geq 2N_1} \frac{1}{t-N_1}\sum_{i=N_1}^{t} |f^*(X_i)|\right).
\end{aligned}$$

Note that this is upper bounded by $\frac{\epsilon}{4B} B = \frac{\epsilon}{4}$ with probability $1 - \frac{2\delta}{10}$ in view of the union bound.

For term (I), we have that

$$\sup_{t\geq 2N_1} \frac{1}{t}\sum_{i=N_1}^{t} \left|\frac{1}{M_p^i} f_i(X_i) - \frac{1}{M_p^i} f^*(X_i)\right| \leq \left(\sup_{t\geq N_1} \left|\frac{1}{M_p^i}\right|\right)\left(\sup_{t\geq 2N_1} \frac{1}{t}\sum_{i=N_1}^{t} |f_i(X_i) - f^*(X_i)|\right).$$

Now we observe that

$$\begin{aligned}
\sup_{t\geq 2N_1} \frac{1}{t}\sum_{i=N_1}^{t} |f_i(X_i) - f^*(X_i)| &= \sup_{t\geq 2N_1} \frac{1}{t}\sum_{i=N_1}^{t} \left\langle \frac{1}{i-1}\sum_{k=1}^{i-1} \xi_p(\cdot, X_k) - \mathbb{E}[\xi_p(\cdot, X)], \xi_p(\cdot, X_i)\right\rangle_{\mathcal{H}^d} \\
&\leq \sup_{t\geq 2N_1} \frac{1}{t}\sum_{i=N_1}^{t} \left\| \frac{1}{i-1}\sum_{k=1}^{i-1} \xi_p(\cdot, X_k) - \mathbb{E}[\xi_p(\cdot, X)]\right\|_{\mathcal{H}^d} \|\xi_p(\cdot, X_i)\|_{\mathcal{H}^d} \\
&\leq \sup_{i\geq 2N_1} \left\{ \left\| \frac{1}{i-1}\sum_{k=1}^{i-1} \xi_p(\cdot, X_k) - \mathbb{E}[\xi_p(\cdot, X)]\right\|_{\mathcal{H}^d}\right\} \sup_{t\geq 2N_1} \left\{ \frac{1}{t}\sum_{i=N_1}^{t} \|\xi_p(\cdot, X_i)\|_{\mathcal{H}^d}\right\} \\
&\leq \sup_{i\geq N_1} \left\{ \left\| \frac{1}{i-1}\sum_{k=1}^{i-1} \xi_p(\cdot, X_k) - \mathbb{E}[\xi_p(\cdot, X)]\right\|_{\mathcal{H}^d}\right\} \\
&\quad \times \sup_{t\geq 2N_1} \left\{ \frac{1}{t-N_1}\sum_{i=N_1}^{t} \|\xi_p(\cdot, X_i)\|_{\mathcal{H}^d}\right\}
\end{aligned}$$

Note that this is upper bounded by $\frac{\epsilon}{4B^2}B = \frac{\epsilon}{4B}$ with probability $\frac{2\delta}{10}$ in view of the union bound. Thus term (I) is upper bounded by $\frac{\epsilon}{4B}B = \frac{\epsilon}{4}$ with probability $1 - \frac{3\delta}{10}$ in view of the union bound.

Concluding the step by considering the union bound over all the terms: Taking $M := \max(2N_1, N_2)$, we conclude that

$$\sup_{t \geq M} \left| \frac{1}{t} \sum_{i=1}^{t} g_i(X_i) - \frac{1}{t} \sum_{i=1}^{t} g^*(X_i) \right| \leq \frac{\epsilon}{2} + \frac{\epsilon}{4} + \frac{\epsilon}{4} = \epsilon$$

with probability $1 - (\frac{\delta}{2} + \frac{2\delta}{10} + \frac{3\delta}{10}) = 1 - \delta$ in view of the union bound.

□

Proposition 2. *If $\mathbb{E}[h_p(X,X)] < \infty$ and $M_p < \infty$, then*

$$\frac{1}{t} \sum_{i=1}^{t} g_i^2(X_i) \overset{a.s.}{\to} \mathbb{E}[(g^*(X))^2].$$

Proof. Without loss of generality, we can further assume that $M_p > 0$. We first highlight that, based on $0 < \mathbb{E}[h_p(X,X)] < \infty$,

- both $\mathbb{E}[|f^*(X)|]$ and $\mathbb{E}[(f^*(X))^2]$ are finite, as $\mathbb{E}[(f^*(X))^2] = \mathbb{E}_X[\mathbb{E}_{X'}[h_p^2(X',X)]] = \mathbb{E}_{X,X'}[h_p^2(X',X)] \leq \mathbb{E}_{X,X'}\left[\|\xi_p(\cdot,X)\|_{\mathcal{H}^d}^2 \|\xi_p(\cdot,X')\|_{\mathcal{H}^d}^2\right] = \mathbb{E}_X\left[\|\xi_p(\cdot,X)\|_{\mathcal{H}^d}^2\right]^2 = \mathbb{E}[h_p(X,X)]^2 < \infty$ and $\mathbb{E}^2[f^*(X)] \leq \mathbb{E}[(f^*(X))^2]$,
- both $\mathbb{E}\left[\|\xi_p(\cdot,X)\|_{\mathcal{H}^d}\right]$ and $\mathbb{E}\left[\|\xi_p(\cdot,X)\|_{\mathcal{H}^d}^2\right]$ is finite, as $\mathbb{E}_X\left[\|\xi_p(\cdot,X)\|_{\mathcal{H}^d}^2\right] = \mathbb{E}[h_p(X,X)] < \infty$ and $\mathbb{E}^2\left[\|\xi_p(\cdot,X)\|_{\mathcal{H}^d}\right] \leq \mathbb{E}\left[\|\xi_p(\cdot,X)\|_{\mathcal{H}^d}^2\right]$.
- both $\mathbb{E}[|g^*(X)|]$ and $\mathbb{E}[(g^*(X))^2]$ are finite, as $g^*(x) = f^*(x)/M_p$, $\mathbb{E}[|f^*(X)|]$ and $\mathbb{E}[(f^*(X))^2]$ are also finite, and $M_p > 0$.

We want to show that $\frac{1}{t}\sum_{i=1}^{t} g_i^2(X_i) \overset{a.s.}{\to} \mathbb{E}[(g^*(X))^2]$, i.e. $\frac{1}{t}\sum_{i=1}^{t} g_i^2(X_i) - \mathbb{E}[(g^*(X))^2] \overset{a.s.}{\to} 0$. We decompose

$$\begin{aligned}
\frac{1}{t} \sum_{i=1}^{t} g_i^2(X_i) - \mathbb{E}[(g^*(X))^2] &= \frac{1}{t} \sum_{i=1}^{t} \left(g_i(X_i) - g^*(X_i) + g^*(X_i)\right)^2 - \mathbb{E}[(g^*(X))^2] \\
&= \frac{1}{t} \sum_{i=1}^{t} \left(g_i(X_i) - g^*(X_i)\right)^2 - \frac{2}{t} \sum_{i=1}^{t} \left(g_i(X_i) - g^*(X_i)\right) g^*(X_i) + \\
&+ \frac{1}{t} \sum_{i=1}^{t} \left(g^*(X_i)\right)^2 - \mathbb{E}[(g^*(X))^2].
\end{aligned}$$

The third term converges to zero almost surely in view of $\mathbb{E}[(g^*(X))^2] < \infty$ and the SLLN. For the second term, we apply Cauchy-Schwarz inequality to obtain

$$\begin{aligned}
\left| \frac{2}{t} \sum_{i=1}^{t} \left(g_i(X_i) - g^*(X_i)\right) g^*(X_i) \right| &\leq \frac{2}{t} \sum_{i=1}^{t} \left| \left(g_i(X_i) - g^*(X_i)\right) g^*(X_i) \right| \\
&\leq 2 \left[\frac{1}{t} \sum_{i=1}^{t} \left(g_i(X_i) - g^*(X_i)\right)^2 \right]^{\frac{1}{2}} \left[\frac{1}{t} \sum_{i=1}^{t} \left(g^*(X_i)\right)^2 \right]^{\frac{1}{2}}
\end{aligned}$$

Given that $\frac{1}{t}\sum_{i=1}^{t} \left(g^*(X_i)\right)^2 \overset{a.s.}{\to} \mathbb{E}[(g^*(X))^2]$, we derive

$$\left[\frac{1}{t} \sum_{i=1}^{t} \left(g^*(X_i)\right)^2 \right]^{\frac{1}{2}} \overset{a.s.}{\to} E^{\frac{1}{2}}[(g^*(X))^2].$$

Note that if $\frac{1}{t}\sum_{i=1}^{t}\left(g_i(X_i)-g^*(X_i)\right)^2 \overset{a.s.}{\to} 0$, it also follows that $\left[\frac{1}{t}\sum_{i=1}^{t}\left(g_i(X_i)-g^*(X_i)\right)^2\right]^{\frac{1}{2}} \overset{a.s.}{\to} 0$, implying that the second term converges to zero almost surely. Thus it suffices to show that the first term converges to zero almost surely, i.e. $\frac{1}{t}\sum_{i=1}^{t}\left(g_i(X_i)-g^*(X_i)\right)^2 \overset{a.s.}{\to} 0$, to conclude the result. We prove so similarly to Proposition 1.

Introducing the SLLN terms and the probabilistic bounds: By the SLLN and the finiteness of M_p, $\mathbb{E}[(f^*(X))^2]$, and $\mathbb{E}\left[\|\xi_p(\cdot,X)\|^2_{\mathcal{H}^d}\right]$, we have that

- $M_p^i \overset{a.s.}{\to} M_p$, and so $\frac{1}{M_p^i} \overset{a.s.}{\to} \frac{1}{M_p}$ (given that $M_p \neq 0$),
- $\frac{1}{t}\sum_{i=M_1}^{t}\left(f^*(X_i)\right)^2 \overset{a.s.}{\to} \mathbb{E}[(f^*(X))^2]$,
- $\frac{1}{t}\sum_{i=1}^{t}\|\xi_p(\cdot,X_i)\|^2_{\mathcal{H}^d} \overset{a.s.}{\to} \mathbb{E}\left[\|\xi_p(\cdot,X)\|^2_{\mathcal{H}^d}\right]$.

From the Banach space-valued SLLN (Theorem 4) and the finiteness of $\mathbb{E}\left[\|\xi_p(\cdot,X)\|_{\mathcal{H}^d}\right]$, it also follows that

- $\left\|\frac{1}{i-1}\sum_{k=1}^{i-1}\xi_p(\cdot,X_k)-\mathbb{E}[\xi_p(\cdot,X)]\right\|_{\mathcal{H}^d} \overset{a.s.}{\to} 0$.

Hence there exist $B>0$ and $N_1 \in \mathbb{N}$ such that

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{t\geq N_1}\left(\frac{1}{M_p^i}\right)^2 > B\right) \leq \frac{\delta}{10},\\
&\mathbb{P}\left(\sup_{t\geq N_1}\frac{1}{t}\sum_{i=1}^{t}\left(f^*(X_i)\right)^2 > B\right) \leq \frac{\delta}{10},\\
&\mathbb{P}\left(\sup_{t\geq N_1}\frac{1}{t}\sum_{i=1}^{t}\|\xi_p(\cdot,X_i)\|^2_{\mathcal{H}^d} > B\right) \leq \frac{\delta}{10},\\
&\mathbb{P}\left(\sup_{t\geq N_1}\left(\frac{1}{M_p^i}-\frac{1}{M_p}\right)^2 > \frac{\epsilon}{8B}\right) \leq \frac{\delta}{10}.\\
&\mathbb{P}\left(\sup_{i\geq N_1}\left\|\frac{1}{i-1}\sum_{k=1}^{i-1}\xi_p(\cdot,X_k)-\mathbb{E}[\xi_p(\cdot,X)]\right\|^2_{\mathcal{H}^d} > \frac{\epsilon}{8B^2}\right) \leq \frac{\delta}{10}.
\end{aligned}
$$

Note that $\mathbb{P}\left(\sup_{t\geq N_1}\frac{1}{t}\sum_{i=1}^{t}\|\xi_p(\cdot,X_i)\|^2_{\mathcal{H}^d} > B\right) \leq \frac{\delta}{10}$ and $\mathbb{P}\left(\sup_{t\geq N_1}\frac{1}{t}\sum_{i=1}^{t}\left(f^*(X_i)\right)^2 > B\right) \leq \frac{\delta}{10}$ imply that $\mathbb{P}\left(\sup_{t\geq 2N_1}\frac{1}{t-N_1}\sum_{i=N_1}^{t}\|\xi_p(\cdot,X_i)\|^2_{\mathcal{H}^d} > B\right) \leq \frac{\delta}{5}$ and $\mathbb{P}\left(\sup_{t\geq 2N_1}\frac{1}{t-N_1}\sum_{i=N_1}^{t}\left(f^*(X_i)\right)^2 > B\right) \leq \frac{\delta}{10}$, given that the data are iid.

Combining the probabilistic bounds: We start by noting that

$$
\frac{1}{t}\sum_{i=1}^{t}\left(g_i(X_i)-g^*(X_i)\right)^2 = \frac{1}{t}\sum_{i=1}^{N_1-1}\left(g_i(X_i)-g^*(X_i)\right)^2 + \frac{1}{t}\sum_{i=N_1}^{t}\left(g_i(X_i)-g^*(X_i)\right)^2.
$$

Consider the sequence of random variables $Y_t = \frac{1}{t}\left(\sum_{i=1}^{N_1-1}\left(g_i(X_i)-g^*(X_i)\right)^2\right)$. Clearly, $Y_t \overset{a.s.}{\to} 0$. Thus there exists N_2 such that $\sup_{t\geq N_2} Y_t \leq \frac{\epsilon}{2}$ with probability $1-\delta/2$.

Moreover,

$$\begin{aligned}\frac{1}{t}\sum_{i=N_1}^{t}(g_i(X_i)-g^*(X_i))^2 &= \frac{1}{t}\sum_{i=N_1}^{t}\left(\frac{1}{M_p^i}f_i(X_i)-\frac{1}{M_p^i}f^*(X_i)+\frac{1}{M_p^i}f^*(X_i)-\frac{1}{M_p}f^*(X_i)\right)^2 \\ &\overset{(i)}{\leq} \underbrace{\frac{2}{t}\sum_{i=N_1}^{t}\left(\frac{1}{M_p^i}f_i(X_i)-\frac{1}{M_p^i}f^*(X_i)\right)^2}_{(I)} + \\ &+ \underbrace{\frac{2}{t}\sum_{i=N_1}^{t}\left(\frac{1}{M_p^i}f^*(X_i)-\frac{1}{M_p}f^*(X_i)\right)^2}_{(II)},\end{aligned}$$

where (i) is obtained in view of $(a+b)^2 \leq 2a^2+2b^2$. It now remains to prove that terms (I) and (II) converge almost surely to zero. For term (II), we observe that

$$\begin{aligned}2\sup_{t\geq 2N_1}\frac{1}{t}\sum_{i=N_1}^{t}\left(\frac{1}{M_p^i}f^*(X_i)-\frac{1}{M_p}f^*(X_i)\right)^2 &= 2\sup_{t\geq 2N_1}\frac{1}{t}\sum_{i=N_1}^{t}\left(\frac{1}{M_p^i}-\frac{1}{M_p}\right)^2(f^*(X_i))^2 \\ &\leq 2\left(\sup_{t\geq 2N_1}\left(\frac{1}{M_p^i}-\frac{1}{M_p}\right)^2\right)\left(\sup_{t\geq 2N_1}\frac{1}{t}\sum_{i=N_1}^{t}(f^*(X_i))^2\right) \\ &\leq 2\left(\sup_{t\geq N_1}\left(\frac{1}{M_p^i}-\frac{1}{M_p}\right)^2\right)\left(\sup_{t\geq 2N_1}\frac{1}{t-N_1}\sum_{i=N_1}^{t}(f^*(X_i))^2\right).\end{aligned}$$

Note that this is upper bounded by $2\frac{\epsilon}{8B}B=\frac{\epsilon}{4}$ with probability $1-\frac{2\delta}{10}$ in view of the union bound.

For term (I), we have that

$$2\sup_{t\geq 2N_1}\frac{1}{t}\sum_{i=N_1}^{t}\left(\frac{1}{M_p^i}f_i(X_i)-\frac{1}{M_p^i}f^*(X_i)\right)^2 \leq 2\left(\sup_{t\geq N_1}\left(\frac{1}{M_p^i}\right)^2\right)\left(\sup_{t\geq 2N_1}\frac{1}{t}\sum_{i=N_1}^{t}(f_i(X_i)-f^*(X_i))^2\right).$$

Now we observe that

$$\begin{aligned}\sup_{t\geq 2N_1}\frac{1}{t}\sum_{i=N_1}^{t}(f_i(X_i)-f^*(X_i))^2 &= \sup_{t\geq 2N_1}\frac{1}{t}\sum_{i=N_1}^{t}\left\langle \mathbb{E}[\xi_p(\cdot,X)]-\frac{1}{i-1}\sum_{k=1}^{i-1}\xi_p(\cdot,X_k),\xi_p(\cdot,X_i)\right\rangle_{\mathcal{H}^d}^2 \\ &\leq \sup_{t\geq 2N_1}\frac{1}{t}\sum_{i=N_1}^{t}\left\|\mathbb{E}[\xi_p(\cdot,X)]-\frac{1}{i-1}\sum_{k=1}^{i-1}\xi_p(\cdot,X_k)\right\|_{\mathcal{H}^d}^2\|\xi_p(\cdot,X_i)\|_{\mathcal{H}^d}^2 \\ &\leq \left\|\mathbb{E}[\xi_p(\cdot,X)]-\frac{1}{i-1}\sum_{k=1}^{i-1}\xi_p(\cdot,X_k)\right\|_{\mathcal{H}^d}^2\sup_{t\geq 2N_1}\frac{1}{t}\sum_{i=N_1}^{t}\|\xi_p(\cdot,X_i)\|_{\mathcal{H}^d}^2 \\ &\leq \left\|\mathbb{E}[\xi_p(\cdot,X)]-\frac{1}{i-1}\sum_{k=1}^{i-1}\xi_p(\cdot,X_k)\right\|_{\mathcal{H}^d}^2\sup_{t\geq 2N_1}\frac{1}{t-N_1}\sum_{i=N_1}^{t}\|\xi_p(\cdot,X_i)\|_{\mathcal{H}^d}^2\end{aligned}$$

Note that this is upper bounded by $\frac{\epsilon}{8B^2}B=\frac{\epsilon}{8}B$ with probability $1-\frac{2\delta}{10}$ in view of the union bound, and hence term (I) is upper bounded by $2\frac{\epsilon}{8B}B=\frac{\epsilon}{4}$ with probability $1-\frac{3\delta}{10}$ in view of the union bound.

Concluding the step by considering the union bound over all the terms:

Taking $N := \max(2N_1, N_2)$, we conclude that

$$\sup_{t\geq N}\left|\frac{1}{t}\sum_{i=1}^{t}g_i(X_i)-\frac{1}{t}\sum_{i=1}^{t}g^*(X_i)\right| \leq \frac{\epsilon}{2}+\frac{\epsilon}{4}+\frac{\epsilon}{4}=\epsilon$$

with probability $1-(\frac{\delta}{2}+\frac{2\delta}{10}+\frac{3\delta}{10})=1-\delta$ in view of the union bound.

□

C.4 Proofs of Main Theorems

Proof of Theorem 1. We first note that

$$\begin{aligned}
\mathbb{E}_{H_0}\left[g_t(X_t)\middle|\mathcal{F}_{t-1}\right] &= \mathbb{E}_{H_0}\left[\frac{1}{\frac{1}{t-1}\sum_{i=1}^{t-1}M_p(X_i)}\left(\frac{1}{t-1}\sum_{i=1}^{t-1}h_p(X_i,X_t)\right)\middle|\mathcal{F}_{t-1}\right] \\
&= \frac{1}{\frac{1}{t-1}\sum_{i=1}^{t-1}M_p(X_i)}\mathbb{E}_{H_0}\left[\frac{1}{t-1}\sum_{i=1}^{t-1}h_p(X_i,X_t)\middle|\mathcal{F}_{t-1}\right] \\
&\overset{(i)}{=} \frac{1}{\frac{1}{t-1}\sum_{i=1}^{t-1}M_p(X_i)}\times 0 \\
&= 0,
\end{aligned}$$

where (i) is obtained in view of (4).

Hence

$$\begin{aligned}
\mathbb{E}_{H_0}\left[\mathcal{K}_t\middle|\mathcal{F}_{t-1}\right] &= \mathbb{E}_{H_0}\left[\mathcal{K}_{t-1}\times(1+\lambda_t g_t(X_t))\middle|\mathcal{F}_{t-1}\right] \\
&= \mathcal{K}_{t-1}\times\left(1+\lambda_t\mathbb{E}_{H_0}\left[g_t(X_t)\middle|\mathcal{F}_{t-1}\right]\right) \\
&= \mathcal{K}_{t-1}\times(1+0) \\
&= \mathcal{K}_{t-1},
\end{aligned}$$

which implies that $\mathcal{K}_{t-1}$ is a martingale. Furthermore,

$$\begin{aligned}
g_t(x) &= \frac{1}{\frac{1}{t-1}\sum_{i=1}^{t-1}M_p(X_i)}\left(\frac{1}{t-1}\sum_{i=1}^{t-1}h_p(X_i,x)\right) \\
&\geq \frac{1}{\frac{1}{t-1}\sum_{i=1}^{t-1}M_p(X_i)}\left(\frac{1}{t-1}\sum_{i=1}^{t-1}-M_p(X_i)\right) \\
&= -1.
\end{aligned}$$

Given that $\lambda_t\in[0,1]$, $\lambda_t g_t$ is also lower bounded by -1, and so $\mathcal{K}_t$ is non-negative. Thus, $\mathcal{K}_t$ is a test martingale. In view of $\mathbb{E}_{H_0}[\mathcal{K}_0]=1$ and Ville's inequality, we conclude that τ is a level-α sequential test.

□

Proof of Theorem 2. We want to prove that $\liminf_{t\to\infty}\frac{\log\mathcal{K}_t}{t}\geq\frac{\left(\mathbb{E}_{H_1}[g^*(X)]\right)^2/2}{\mathbb{E}_{H_1}[g^*(X)]+\mathbb{E}_{H_1}[(g^*(X))^2]}=:L$.

Following the LBOW strategy, we take

$$\lambda_t=\max\left(0,\frac{\frac{1}{t-1}\sum_{i=1}^{t-1}g_i(X_i)}{\frac{1}{t-1}\sum_{i=1}^{t-1}g_i(X_i)+\frac{1}{t-1}\sum_{i=1}^{t-1}g_i^2(X_i)}\right).$$

Given that $\mathbb{E}[h_p(X,X)]<\infty$, it also holds that $\mathbb{E}[\sqrt{h_p(X,X)}]<\infty$. Consequently, Proposition 1 and Proposition 2 yield $\frac{1}{t-1}\sum_{i=1}^{t-1}g_i(X_i)\overset{a.s.}{\to}\mathbb{E}[g^*(X)]$, as well as $\frac{1}{t-1}\sum_{i=1}^{t-1}g_i^2(X_i)\overset{a.s.}{\to}\mathbb{E}[(g^*(X))^2]$. Given that $\mathbb{E}[g^*(X)]>0$, it follows that

$$\lambda_t\overset{a.s.}{\to}\frac{\mathbb{E}[g^*(X)]}{\mathbb{E}[g^*(X)]+\mathbb{E}[(g^*(X))^2]}=:\lambda^*\in(0,1).$$

For $y \geq -1$ and $\lambda \in [0,1)$, it holds that

$$\log(1+\lambda y) \geq \lambda y + y^2 \left(\log(1-\lambda) + \lambda\right).$$

Further, for $\lambda \in [0,1)$, it holds that

$$\log(1-\lambda) + \lambda \geq -\frac{\lambda^2}{2(1-\lambda)}.$$

Thus, for $y \geq -1$ and $\lambda \in [0,1)$,

$$\log(1+\lambda y) \geq \lambda y - y^2 \frac{\lambda^2}{2(1-\lambda)}.$$

Consequently, we derive that

$$\begin{aligned}
\frac{\log \mathcal{K}_t}{t} &= \frac{1}{t}\sum_{i=1}^{t} \log\left(1+\lambda_i g_i(X_i)\right) \\
&\geq \frac{1}{t}\sum_{i=1}^{t} \lambda_i g_i(X_i) - g_i^2(X_i)\frac{\lambda_i^2}{2(1-\lambda_i)}.
\end{aligned}$$

It thus suffices to prove that $\frac{1}{t}\sum_{i=1}^{t} \lambda_i g_i(X_i) - g_i^2(X_i)\frac{\lambda_i^2}{2(1-\lambda_i)} \overset{a.s.}{\to} L$ to conclude the result. To see this, we denote $\kappa(\lambda) = \frac{\lambda^2}{2(1-\lambda)}$ and highlight that

$$\begin{aligned}
\frac{1}{t}\sum_{i=1}^{t} \lambda_i g_i(X_i) - g_i^2(X_i)\frac{\lambda_i^2}{2(1-\lambda_i)} &= \frac{1}{t}\sum_{i=1}^{t} \lambda_i g_i(X_i) - g_i^2(X_i)\kappa(\lambda_i) \\
&= \frac{1}{t}\sum_{i=1}^{t} (\lambda_i - \lambda^* + \lambda^*) g_i(X_i) - \left(\kappa(\lambda_i) - \kappa(\lambda^*) + \kappa(\lambda^*)\right) g_i^2(X_i) \\
&= \underbrace{\frac{1}{t}\sum_{i=1}^{t} \lambda^* g_i(X_i) - \kappa(\lambda^*) g_i^2(X_i)}_{(I)} + \\
&\quad + \underbrace{\frac{1}{t}\sum_{i=1}^{t} (\lambda_i - \lambda^*) g_i(X_i)}_{(II)} - \underbrace{\frac{1}{t}\sum_{i=1}^{t} \left(\kappa(\lambda_i) - \kappa(\lambda^*)\right) g_i^2(X_i)}_{(III)}.
\end{aligned}$$

Now we note that term (I) converges almost surely to $\lambda^* \mathbb{E}[g^*(X)] - \kappa(\lambda^*)\mathbb{E}[(g^*(X))^2] = L$, given that $\frac{1}{t}\sum_{i=1}^{t} g_i(X_i) \overset{a.s.}{\to} \mathbb{E}[g^*(X)]$ and $\frac{1}{t}\sum_{i=1}^{t} g_i^2(X_i) \overset{a.s.}{\to} \mathbb{E}[(g^*(X))^2]$. Thus, it suffices to prove that (II) and (III) converge almost surely to zero to derive that

$$\frac{1}{t}\sum_{i=1}^{t} \lambda_i g_i(X_i) - g_i^2(X_i)\frac{\lambda_i^2}{2(1-\lambda_i)} \overset{a.s.}{\to} L,$$

hence concluding that

$$\liminf_{t\to\infty} \frac{\log \mathcal{K}_t}{t} \geq \liminf_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} \lambda_i g_i(X_i) - g_i^2(X_i)\frac{\lambda_i^2}{2(1-\lambda_i)} = L.$$

Let us now prove that (II) and (III) converge almost surely to 0. Let $\epsilon > 0$ and $\delta > 0$ be arbitrary. We ought to show that there exists $N \in \mathbb{N}$ such that

$$\mathbb{P}\left(\sup_{t\geq N}\left|\frac{1}{t}\sum_{i=1}^{t}(\lambda_i - \lambda^*) g_i(X_i)\right| > \epsilon\right) \leq \delta, \quad \mathbb{P}\left(\sup_{t\geq N}\left|\frac{1}{t}\sum_{i=1}^{t}\left(\kappa(\lambda_i) - \kappa(\lambda^*)\right) g_i^2(X_i)\right| > \epsilon\right) \leq \delta.$$

Based on $\frac{1}{t}\sum_{i=1}^{t}|g_i(X_i)-g^*(X_i)| \overset{a.s.}{\to} 0$, we derive that that

$$\begin{aligned}\frac{1}{t}\sum_{i=1}^{t}|g_i(X_i)| &= \frac{1}{t}\sum_{i=1}^{t}|g_i(X_i)-g^*(X_i)+g^*(X_i)| \\ &\leq \underbrace{\frac{1}{t}\sum_{i=1}^{t}|g_i(X_i)-g^*(X_i)|}_{\overset{a.s.}{\to}0} + \underbrace{\frac{1}{t}\sum_{i=1}^{t}|g^*(X_i)|}_{\overset{a.s.}{\to}\mathbb{E}[|g^*(X)|]},\end{aligned}$$

and so $\limsup_{t\to\infty}\frac{1}{t}\sum_{i=1}^{t}|g_i(X_i)| \leq \mathbb{E}[|g^*(X)|]$. Furthermore, we know that $\frac{1}{t}\sum_{i=1}^{t}g_i^2(X_i) \overset{a.s.}{\to} \mathbb{E}[(g^*(X))^2]$. Given that $\lambda \overset{a.s.}{\to} \lambda^*$ and κ is continuous, it follows that $\kappa(\lambda) \overset{a.s.}{\to} \kappa(\lambda^*)$. Thus, by the SLLN, there exist $B>0$ and $N_1 \in \mathbb{N}$ such that

$$\begin{aligned}&\mathbb{P}\left(\sup_{t\geq N_1}\frac{1}{t}\sum_{i=1}^{t}|g_i(X_i)|>B\right)\leq\frac{\delta}{3}, \quad \mathbb{P}\left(\sup_{t\geq N_1}\frac{1}{t}\sum_{i=1}^{t}g_i^2(X_i)>B\right)\leq\frac{\delta}{3}, \\ &\mathbb{P}\left(\sup_{t\geq N_1}|\lambda_i-\lambda^*|>\frac{\epsilon}{2B}\right)\leq\frac{\delta}{3}, \quad \mathbb{P}\left(\sup_{t\geq N_1}|\kappa(\lambda_i)-\kappa(\lambda^*)|>\frac{\epsilon}{2B}\right)\leq\frac{\delta}{3}.\end{aligned}$$

Note that $\mathbb{P}\left(\sup_{t\geq N_1}\frac{1}{t}\sum_{i=1}^{t}|g_i(X_i)|>B\right) \leq \frac{\delta}{3}$ and $\mathbb{P}\left(\sup_{t\geq N_1}\frac{1}{t}\sum_{i=1}^{t}g_i^2(X_i)>B\right) \leq \frac{\delta}{3}$ imply that $\mathbb{P}\left(\sup_{t\geq 2N_1}\frac{1}{t-N_1}\sum_{i=N_1}^{t}|g_i(X_i)|>B\right)\leq\frac{\delta}{3}$ and $\mathbb{P}\left(\sup_{t\geq 2N_1}\frac{1}{t-N_1}\sum_{i=N_1}^{t}g_i^2(X_i)>B\right)\leq\frac{\delta}{3}$, given that the data are iid.

Thus,

$$\begin{aligned}\sup_{t\geq 2N_1}\left|\frac{1}{t}\sum_{i=N_1}^{t}(\lambda_i-\lambda^*)g_i(X_i)\right| &\leq \sup_{t\geq 2N_1}\frac{1}{t}\sum_{i=N_1}^{t}|(\lambda_i-\lambda^*)g_i(X_i)| \\ &\leq \sup_{t\geq N_1}|\lambda_i-\lambda^*|\sup_{t\geq 2N_1}\frac{1}{t}\sum_{i=N_1}^{t}|g_i(X_i)| \\ &\leq \sup_{t\geq N_1}|\lambda_i-\lambda^*|\sup_{t\geq 2N_1}\frac{1}{t-N_1}\sum_{i=N_1}^{t}|g_i(X_i)| \\ &\leq \frac{\epsilon}{2B}B \\ &= \frac{\epsilon}{2}\end{aligned}$$

with probability $1-\frac{2}{3}\delta$, as well as

$$\begin{aligned}\sup_{t\geq 2N_1}\left|\frac{1}{t}\sum_{i=N_1}^{t}\left(\kappa(\lambda_i)-\kappa(\lambda^*)\right)g_i^2(X_i)\right| &\leq \sup_{t\geq 2N_1}\frac{1}{t}\sum_{i=N_1}^{t}\left|\left(\kappa(\lambda_i)-\kappa(\lambda^*)\right)g_i^2(X_i)\right| \\ &\leq \sup_{t\geq 2N_1}|\kappa(\lambda_i)-\kappa(\lambda^*)|\sup_{t\geq 2N_1}\frac{1}{t}\sum_{i=N_1}^{t}g_i^2(X_i) \\ &\leq \sup_{t\geq N_1}|\kappa(\lambda_i)-\kappa(\lambda^*)|\sup_{t\geq 2N_1}\frac{1}{t-N_1}\sum_{i=N_1}^{t}g_i^2(X_i) \\ &\leq \frac{\epsilon}{2B}B \\ &= \frac{\epsilon}{2}\end{aligned}$$

with probability $1-\frac{2}{3}\delta$.

Consider now the sequence of random variables

$$Y_t = \frac{1}{t}\sum_{i=1}^{N_1-1}(\lambda_i - \lambda^*)g_i(X_i), \quad \tilde{Y}_t = \frac{1}{t}\sum_{i=1}^{N_1-1}\left(\kappa(\lambda_i) - \kappa(\lambda^*)\right)g_i^2(X_i).$$

Clearly, $Y_t \overset{a.s.}{\to} 0$ and $\tilde{Y}_t \overset{a.s.}{\to} 0$. Thus there exists N_2 such that $Y_t \leq \frac{\epsilon}{2}$ and $\tilde{Y}_t \leq \frac{\epsilon}{3}$, both with probability $1 - \frac{\delta}{3}$. Taking $N = \max(2N_1, N_2)$ and in view of the union bound, we derive that

$$\begin{aligned}
\sup_{t\geq N}\left|\frac{1}{t}\sum_{i=1}^{t}(\lambda_i - \lambda^*)g_i(X_i)\right| &\leq \sup_{t\geq N}\left|\frac{1}{t}\sum_{i=1}^{N_1-1}(\lambda_i - \lambda^*)g_i(X_i) + \frac{1}{t}\sum_{i=N_1}^{t}(\lambda_i - \lambda^*)g_i(X_i)\right| \\
&\leq \sup_{t\geq N}\left|\frac{1}{t}\sum_{i=1}^{N_1-1}(\lambda_i - \lambda^*)g_i(X_i)\right| + \sup_{t\geq N}\left|\frac{1}{t}\sum_{i=N_1}^{t}(\lambda_i - \lambda^*)g_i(X_i)\right| \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&= \epsilon
\end{aligned}$$

with probability $1 - (\frac{2}{3}\delta + \frac{1}{3}\delta) = 1 - \delta$, as well as

$$\begin{aligned}
\sup_{t\geq N}\left|\frac{1}{t}\sum_{i=1}^{t}\left(\kappa(\lambda_i) - \kappa(\lambda^*)\right)g_i^2(X_i)\right| &\leq \sup_{t\geq N}\left|\frac{1}{t}\sum_{i=1}^{N_1-1}\left(\kappa(\lambda_i) - \kappa(\lambda^*)\right)g_i^2(X_i) + \frac{1}{t}\sum_{i=N_1}^{t}\left(\kappa(\lambda_i) - \kappa(\lambda^*)\right)g_i^2(X_i)\right| \\
&\leq \sup_{t\geq N}\left|\frac{1}{t}\sum_{i=1}^{N_1-1}\left(\kappa(\lambda_i) - \kappa(\lambda^*)\right)g_i^2(X_i)\right| + \sup_{t\geq N}\left|\frac{1}{t}\sum_{i=N_1}^{t}\left(\kappa(\lambda_i) - \kappa(\lambda^*)\right)g_i^2(X_i)\right| \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&= \epsilon
\end{aligned}$$

with probability $1 - (\frac{2}{3}\delta + \frac{1}{3}\delta) = 1 - \delta$. □

Proof of Theorem 3. Let $\theta_0 \in \Theta$ be such that $P_{\theta_0} = P$. By definition of $\mathcal{K}_t^C$, we have that $\mathcal{K}_t^C \leq \mathcal{K}_t^{\theta_0}$, with $\mathcal{K}_t^{\theta_0}$ being a test martingale. The validity of the test is concluded by noting that

$$\mathbb{P}\left(\mathcal{K}_t^C \geq \frac{1}{\alpha}\right) \leq \mathbb{P}\left(\mathcal{K}_t^{\theta_0} \geq \frac{1}{\alpha}\right) \leq \alpha,$$

where the second inequality is obtained in view of Ville's inequality.

□