A Robust Kernel Statistical Test of Invariance: Detecting Subtle Asymmetries

Ashkan Soleymani[1*], Behrooz Tahmasebi[1*], Stefanie Jegelka[1,2], Patrick Jaillet[1]

[1] MIT, [2] TUM
{ashkanso,bzt,stefje,jaillet}@mit.edu

While invariances naturally arise in almost any type of real-world data, no efficient and robust test exists for detecting them in observational data for arbitrarily given group actions. We tackle this problem by studying measures of invariance that can capture even negligible underlying patterns. Our first contribution is to show that, while detecting subtle asymmetries is computationally intractable, a randomized method can be used to estimate robust closeness measures to invariance within constant factors. This provides a general framework for robust statistical tests of invariance. In addition, we focus on kernel methods and propose deterministic algorithms for robust testing with respect to both finite and infinite groups, accompanied by a rigorous analysis of their convergence rates and sample complexity. Finally, we revisit the general framework in the specific case of kernel methods, showing that recent closeness measures to invariance, defined via group averaging, are provably robust, leading to powerful randomized algorithms.

1. Introduction

Invariances are ubiquitous. Almost all scientific fields study data that manifest consistent patterns that remain unchanged under various transformations [1]. For example, the laws of physics exhibit invariances under coordinate changes or changes in time, promising the universality of underlying principles [2–4]. Traditionally, machine learning models are designed to be invariant with respect to the symmetries of the data by construction, leading to better computational and statistical properties [1]. However in general, prior to introducing invariances into models, either by design or through post-processing steps, it is essential to first verify whether the observational data is invariant with respect to a given algebraic group or not, which is the main focus of this work.

Group invariance hypothesis testing methods encompass a broad range of statistical approaches, including permutation tests and randomization tests [5–10]. These nonparametric tests examine the null hypothesis that the data distribution is invariant under a group action G of transformations [11]. In algebraic terms, the group action G is closed under composition, contains an identity element, and has an inverse for each element $g \in G$. Koning and Hemerik [10], Koning [12], Hemerik [13] argue that, by considering sign-flipping tests, the class of invariance tests can be traced back to the early works of Fisher [14], Fisher et al. [15], Efron [16]. They further extend their argument by suggesting that even standard methods, such as t-tests [17, 18], can be interpreted as tests for group invariances. Testing other class of invariances, e.g., w.r.t. rotations with broader applications has also been explored in the literature [19–22]. However, our focus is on developing general recipes, rather than emphasizing a specific class of invariances.

In this paper, we study hypothesis testing of invariances, given a general topological compact group G, which may be finite or infinite, acting on the domain of datapoints $\mathcal{X}$. We test whether the input distribution $\mu \in \mathcal{P}(\mathcal{X})$ is invariant with respect to transformations induced by G. We define the null hypothesis H_0 as the assumption that μ is the same as $g\mu$ for all $g \in G$. The alternative hypothesis H_1 is defined as the existence of $g \in G$ such that $D(\mu, g\mu) \geq \epsilon$, where D is a metric on the probability

*These authors contributed equally to this work.

Preprint. Work in progress.

space $\mathcal{P}(\mathcal{X})$. This definition of the alternative hypothesis H_1 is designed to *robustly* demonstrate that μ is not G-invariant. The threshold ϵ is introduced to ensure the distinguishability between hypotheses H_0 and H_1. In terms, we formulate the problem as the following.

Input: n independent and identically distributed (i.i.d.) samples from an unknown probability distribution μ; a group action G, and a metric D over the space of probability measures, a threshold ϵ.
Output: Either H_0 or H_1, where

$$\begin{aligned} &\mathrm{H}_{\mathbf{0}} : \mu \overset{\mathrm{d}}{\equiv} g\mu \text{ for all } g \in G. \\ &\mathrm{H}_{\mathbf{1}} : \sup_{g \in G} D(\mu, g\mu) \geq \epsilon. \end{aligned} \tag{1}$$

The null hypothesis H_0 can be equivalently rewritten as $\mu \overset{\mathrm{d}}{\equiv} \overline{g}\mu$, where $\overline{g}$ is drawn according to the Haar measure (uniform) defined over the group action G. The main challenge in this class of hypothesis tests is that the group G may be infinite, or finite but with prohibitively large size $|G|$. For example, for the group of orthogonal matrices $O(d)$, G is infinite, while for the permutation group P_d, $|G| = d! \sim \sqrt{d}\left(\frac{d}{e}\right)^d$. As another example, for the group of sign-flipping matrices F_d, which are diagonal matrices with elements in $\{\pm 1\}$, we know that $|G| = 2^d$. This computational problem is amplified when searching for a certificate $\widehat{g}$ such that $D(\mu, \widehat{g}\mu) \geq \epsilon$, which serves as evidence for the hypothesis H_1. We note that, for almost all uncountable choices of the group G, e.g., Lie groups, the Optimization Problem 1 is highly non-convex, even if we assume the measure μ is readily accessible.

Additionally, there is a major statistical barrier for Formulation 1. Recall that we do not have access to μ directly; instead, we only have the empirical measure $\widehat{\mu}$ induced by n i.i.d. samples. Therefore, we cannot evaluate the optimization problem $\sup\limits_{g \in G} D(\mu, g\mu)$ directly. Instead, we can only estimate it from the observations. The trivial estimator $\sup\limits_{g \in G} D(\widehat{\mu}, g\widehat{\mu})$ is highly biased, and it is not clear how to derive non-asymptotic consistency guarantees for this estimator for general choices of distributions μ and group actions G.

Our fundamental result solves these obstacles. We show that there is no need to exhaustively search over the space G for such a certificate $\widehat{g}$. We demonstrate that, under minimal assumptions on the metric D, $\sup_{g \in G} D(\mu, g\mu)$ is surprisingly sandwiched by constant factors of $\mathbb{E}_g[D(\mu, g\mu)]$, where the randomness is induced by g drawn from the Haar measure over the compact group G. An informal version of this theorem is provided below, with the formal details deferred to subsequent sections.

Theorem 1 (Informal version of Theorem 3)**.** *Under the minimal assumption that the metric D is shift-invariant with respect to G,*

$$\mathbb{E}_g[D(\mu, g\mu)] \leq \sup_{g \in G} D(\mu, g\mu) \leq 4\mathbb{E}_g[D(\mu, g\mu)],$$

where expectation is w.r.t. left Haar (uniform) measure over the group G.

This result is rather surprising, as at first glance, $\sup_{g \in G} D(\mu, g\mu)$ appears to be computationally intractable. And indeed this is the case, as we show in subsequent sections. Even for a finite group G, the exact computation of $\arg\sup_{g \in G} D(\mu, g\mu)$ is NP-hard, even in the benign setting without randomness, such as when μ is a Dirac delta measure. However, Theorem 1 shows that it can be approximated within a factor of 4 by introducing randomization, which can be efficiently estimated by data observations. This theorem is general and holds for many choices of the metric D (Section 4) and any compact topological group G, including Lie groups. In light of this flexibility, we propose a *general recipe* below.

General recipe. We introduce another alternative hypothesis $\widetilde{\mathrm{H}}_{\mathbf{1}}$, where

$$\widetilde{\mathrm{H}}_{\mathbf{1}} : \mathbb{E}_g[D(\mu, g\mu)] \geq \epsilon', \tag{2}$$

with a threshold parameter ϵ'. By Theorem 1, non-asymptotic bounds on Type I and Type II errors of the newly designed test 2 can be converted to non-asymptotic bounds on Type I and Type II errors of the original hypothesis test 1. Furthermore, in contrast to the optimization problem $\sup_{g\in G} D(\mu, g\mu)$, the term $\mathbb{E}_g[D(\mu, g\mu)]$ can be readily estimated from i.i.d. observations by calculating the empirical mean of $D(\mu, g\mu)$. We recall again that it is not clear how to estimate $\sup_{g\in G} D(\mu, g\mu)$ from observations with non-asymptotic guarantees in general.

Next, while our framework is general, we focus on hypothesis testing described by H_0 versus H_1 for the special case of kernel Maximum Mean Discrepancy (MMD) distances, due to their favorable computational and statistical properties. We propose solutions to achieve consistent hypothesis testing for H_0 and H_1 with finite sample guarantees for Type I and Type II errors in the case of finite groups. Furthermore, we illustrate how similar ideas can extend to infinite groups G, by elaborating on the case of rotation invariances. Finally, we revisit the hypothesis testing based on our general recipe and discuss its implications by analyzing the hypothesis test of H_0 versus $\widetilde{\mathrm{H}}_1$, as opposed to H_1.

The structure of this paper is as follows. We begin with discussion on the related work and defer a detailed review of the preliminaries on invariances, kernels and embeddings of measures to the appendix. Next, we discuss the robust invariance hypothesis testing of H_0 versus H_1 and its computational hardness results. We then present our general framework, Theorem 1, and explain how it allows us to reformulate the problem. We further explore the special case of Maximum Mean Discrepancy (MMD) distance for testing H_0 versus H_1, offering solutions for both finite and infinite group settings. Finally, we revisit the MMD setting in the context of H_0 versus $\widetilde{\mathrm{H}}_1$ and discuss its implications. We provide rigorous analysis, confidence intervals, algorithms and consistency results for each one of these settings. In the end, we complement our theory in Theorem 1 with experiments on rotational symmetries, practically showing that $\sup_{g\in G} D(\mu, g\mu)$ is within constant factor of the term $\mathbb{E}_g[D(\mu, g\mu)]$.

2. Related Work

As discussed in the previous section, testing invariances is a prolonged fundamental problem in machine learning and statistics. Here, we review some of the most recent works on this topic. In a slightly less related topic, Law et al. [23] proposed probability distance measures that inherently encoded invariance to additive symmetric noise within the embeddings, so as to account for measurement and data collection noises. Bellot and van der Schaar [24] presented testing on set-valued data with applications in electronic health records. Dobriban [25] discusses the consistency of randomization tests based on invariances for signal-plus-noise models. Kashlak [26] shows that specific functions of random variables exhibit certain invariances in the limit. Koning and Hemerik [10] suggest statistically selective deterministic group transformation testing as opposed to traditional Monte Carlo group-invariance tests based on a uniformly randomly selected subset of the elements of the group. In a follow-up, Koning [12] introduce a tradeoff between the power of the test and computational complexity by selecting a coded subgroup, a very tiny subgroup which is not necessarily easy-to-find for all types of group actions. Ramdas et al. [27] observed that, in the special case of permutations, sampling from any subset (not necessarily a subgroup) of the permutations according to an arbitrary distribution (not necessarily uniform) suffices for the test. Chiu and Bloem-Reddy [28] proposed measuring the invariance of a distribution by considering its distance to the orbit-averaged distribution. In contrast to these works, we focus on robust hypothesis testing, where a certificate of the Type described in Equation (1) is required. Additionally, we discuss general remedies for invariance testing over arbitrary compact groups.

For further discussion on related work, particularly on invariances in machine learning and kernels, please refer to Appendix A.

3. Background

In this section, we provide a short review of the necessary background for the paper, while deferring the detailed version to Appendix B.

Throughout this paper, we consider a complete metric space $\mathcal{X}$ and we study (Borel) probability measures $\mu \in \mathcal{P}(\mathcal{X})$. Moreover, we consider a compact topological group G, endowed with the (uniform) left Haar measure α, acting on $\mathcal{X}$ via homeomorphisms. Indeed, each group element g corresponds to a continuous bijection on $\mathcal{X}$, and the group operation is composing of functions. For any probability measure $\mu \in \mathcal{X}$, let $g\mu \in \mathcal{P}(\mathcal{X})$ denote the push-forward measure according to the action of g on $\mathcal{X}$. Similarly, one can define μ^G which is the distribution of gx when $x \sim \mu$ and $g \in G$ is chosen according to left Haar (uniform) measure on G, independently.

A probability metric $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ is a metric on the space of (Borel) probability measures $\mathcal{P}(\mathcal{X})$. It is called shift-invariant, if and only if $D(g\mu, g\nu) = D(\mu, \nu)$ for any probability measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$.

4. Main Results

We start this section by asserting that the exact computation of $\arg\sup_{g\in G} D(\mu, g\mu)$, even when the group G is finite and the distribution μ is a single-point Dirac delta distribution—and there is no randomness—is computationally intractable.

Theorem 2 (Computational intractability)**.** *There exists a shift-invariant pseudometric $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$, a finite group G, and a discrete probability measure μ such that solving the optimization problem $\arg\sup\limits_{g\in G} D(\mu, g\mu)$ is NP-complete.*

The proof of Theorem 2 is presented in Appendix E.1. We carefully craft a pseudometric $D(.,.)$, a finite group action G, and a deterministic measure μ such that $\arg\sup_{g\in G} D(\mu, g\mu)$ solves a special variant of Travelling Salesman Problem (TSP), which we prove to be NP-complete. Theorem 2 implies that, even in the simplest settings, the optimization problem of Equation (1) is computationally intractable, let alone the statistical challenges in estimating $\sup_{g\in G} D(\mu, g\mu)$ from observations. Next, we state our main theorem that enables a randomized approximation for $\sup_{g\in G} D(\mu, g\mu)$ instead.

Theorem 3 (Probabilistic approximation (formal version of Theorem 1))**.** *Let $\mathcal{X}$ be a complete metric space and $\mathcal{P}(\mathcal{X})$ denote the space of (Borel) probability measures on $\mathcal{X}$. Let G be a compact topological group acting continuously on $\mathcal{X}$. Consider a shift-invariant probability metric $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$. Then,*

$$\mathbb{E}_g[D(\mu, g\mu)] \leq \sup_{g\in G} D(\mu, g\mu) \leq 4\mathbb{E}_g[D(\mu, g\mu)],$$

where expectation is taken w.r.t. left Haar (uniform) measure over the group G.

The proof of Theorem 3 is presented in Appendix E.2.

Remark 4.1. The shift-invariance of D w.r.t. group G in Theorem 3 is a general assumption, satisfied in many settings: Wasserstein distance and any isometry group G; Sobolev Integral Probability Metric and isometry group G; TV distance and any general group G, and MMD distance with shift-invariant kernels.

5. Kernel Maximum Invariance Criterion (KMaxIC)

In this section, we consider the special case of kernel Maximum Mean Discrepancy (MMD) distances[2] and focus on proposing algorithms for the hypothesis testing described by H_0 versus H_1 in Equation (1).

[2]A detailed review of the theory of kernel mean embeddings is provided in Appendix B.

Let $\mathcal{H}$ denote the Reproducing Kernel Hilbert Space (RKHS) of a given Positive Definite Symmetric (PDS) kernel K, and let $\mu_{\mathcal{H}} \in \mathcal{H}$ denote the embedding of $\mu \in \mathcal{P}(\mathcal{X})$ into $\mathcal{H}$. Then, consider the probability metric $D(\mu, \nu) = \mathrm{MMD}(\mu, \nu) := \|\mu_{\mathcal{H}} - \nu_{\mathcal{H}}\|_{\mathcal{H}}$.

The Kernel Maximum Invariance Criterion (KMaxIC) measures closeness to invariance by uniformly bounding the MMD distance across all group elements transformations.

Definition 1 (Kernel Maximum Invariance Criterion (KMaxIC))**.** For any probability measure $\mu \in \mathcal{P}(\mathcal{X})$, the Kernel Maximum Invariance Criterion (KMaxIC) is defined as

$$\mathrm{KMaxIC}(\mu) := \sup_{g \in G} \left\|(g\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\right\|_{\mathcal{H}}^2,$$

where $g\mu$ is the shifted version of μ with respect to the group element $g \in G$.

First, we note that KMaxIC successfully distinguishes G-invariant measures from non-invariants:

Theorem 4 (Definiteness of KMaxIC)**.** *For any probability measure $\mu \in \mathcal{P}(\mathcal{X})$, we have* $\mathrm{KMaxIC}(\mu) = 0$ *if and only if μ is G-invariant, assuming the kernel is universal.*

The proof of Theorem 4 is provided in Appendix D.6. This result demonstrates that KMaxIC provides a well-defined notion of distance to G-invariance for probability measures.

In the next section, we propose solutions to achieve consistent hypothesis testing for H_0 and H_1 (Equation (1)) with finite sample guarantees for Type I and Type II errors in the case of finite groups.

6. Testing Invariances via KMaxIC: Finite Groups

In this section, we present a *deterministic* hypothesis testing algorithm for H_0 and H_1 in Equation (1) based on KMaxIC. For simplicity, we first focus on finite groups, and later we generalize to infinite groups.

Note that KMaxIC does not admit a representation as expectations over kernels. To overcome this challenge in designing statistical hypothesis tests using KMaxIC, we leverage group-theoretic properties.

We begin with the following definition:

Definition 1 (Generating sets)**.** A set $S \subseteq G$ is called a generating set for a group G if for every $g \in G$, there exists $k \in \mathbb{N}$ and $s_1, s_2, \ldots, s_k \in S$, such that for each $i \in [k]$, either $s_i \in S$ or $s_i^{-1} \in S$, and $g = s_1 s_2 \ldots s_k$.

Intuitively, generating sets are subsets of a group that can "generate" the entire group when their elements (or their inverses) are multiplied together. For any (not necessarily generating) set $S \subseteq G$, we have the following inequality:

$$\mathrm{KMaxIC}(\mu) \geq \max_{g \in S} \left\|(g\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\right\|_{\mathcal{H}}^2.$$

However, with generating sets $S \subseteq G$, we can establish a converse to the above inequality.

Theorem 5 (Definiteness of KMaxIC via generating sets)**.** *Assuming the underlying kernel used to define KMaxIC is universal, for any arbitrary generating set $S \subseteq G$ and any probability measure $\mu \in \mathcal{P}(\mathcal{X})$, if*

$$\max_{g \in S} \left\|(g\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\right\|_{\mathcal{H}}^2 = 0,$$

then $\mathrm{KMaxIC}(\mu) = 0$*, which implies that μ is G-invariant.*

The proof of Theorem 5 is provided in Appendix D.7.

This result suggests that it is sufficient to test over a generating set rather than the entire group. Generating sets typically have much smaller cardinality compared to G, leading to significant reductions in sample complexity. In fact, one can show that:

Proposition 1 (Size of generating sets)**.** *Any finite group G has a generating set $S \subseteq G$ of size at most $\log_2(|G|)$.*

The proof of Proposition 1 is presented in Appendix D.8. Therefore, to test whether a probability measure is G-invariant, we can estimate $\left\|(g\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\right\|_{\mathcal{H}}^2$ from data for each $g \in S$:

Proposition 2. *For any $g \in G$ and any probability measure $\mu \in \mathcal{P}(\mathcal{X})$, we have*

$$\left\|(g\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\right\|_{\mathcal{H}}^2 = 2\mathbb{E}_{x,x'}[K(x,x')] - 2\mathbb{E}_{x,x'}[K(x,gx')],$$

where $x, x' \sim \mu$ are independent random variables.

The proof of Proposition 2 is provided in Appendix D.9. This identity leads to Algorithm 1.

Algorithm 1 Testing Invariance via KMaxIC

Input: n i.i.d. samples $x_i \sim \mu$, $i \in [n]$, a generating set $S \subseteq G$, and a threshold $c \in (0, \infty)$.

1: For each $g \in S$, compute:

$$\begin{aligned}\widehat{c}_g = &\frac{4}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} K(x_i, x_j) \\ &- \frac{4}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} K(x_i, gx_j).\end{aligned}$$

2: **if** $\max_{g \in S} \widehat{c}_g \leq c$ **then**

3: **return** There is not enough evidence to reject the null hypothesis H_0 that μ is G-invariant.

4: **else**

5: **return** H_1: μ is not G-invariant.

6: **end if**

The total runtime of Algorithm 1 on n samples is $\mathcal{O}(n^2|S|)$, assuming that the kernel function can be computed for each pair of points in constant time. Thanks to Proposition 1, the time complexity is logarithmic in the group size when an appropriate generating set is used, without the need to sample from G. The time complexity can be further reduced to $\mathcal{O}(n|S|)$ by replacing U-statistics with empirical estimates over disjoint pairs of independent samples.

6.1. Confidence Intervals for KMaxIC

In this section, we provide confidence intervals for Algorithm 1. To begin, we introduce the following definition. For any generating set $S \subseteq G$, let $\ell(S)$ denote the maximum length of the minimal representations of group elements $g \in G$ as products of elements (or inverses of elements) from S. This quantity plays a crucial role in the confidence intervals derived for the parameter c in Algorithm 1.

Theorem 6. *Consider Algorithm 1 ran on n samples from a G-invariant probability measure μ. Then, the probability of a Type I error (i.e., incorrectly rejecting the invariance) is bounded as*

$$\begin{aligned}\mathbb{P}\Big(\mathrm{H}_1 | \mathrm{H}_0\Big) &= \mathbb{P}\Big(\max_{g \in S} \widehat{c}_g > c \,\Big|\, \mu \text{ is } G\text{-invariant}\Big) \\ &\leq |S| \exp\left(-\frac{nc^2}{128 c_1^2}\right),\end{aligned}$$

where $c_1 := \sup_{x \in \mathcal{X}} K(x,x)$. Moreover, the Type II error, which is the probability of incorrectly accepting a non-invariant measure using Algorithm 1, approaches zero as the sample size increases. Quantitatively, for *any probability measure $\mu \in \mathcal{P}(\mathcal{X})$ such that* $\mathrm{KMaxIC}(\mu) \geq 2c' > c\ell(S)^2$*, we have*

$$\mathbb{P}\Big(\mathrm{H}_{\mathbf{0}}|\mathrm{H}_{\mathbf{1}}\Big) = \mathbb{P}\Big(\max_{g \in S} \widehat{c}_g \leq c \,\Big|\, \mu \text{ is not } G\text{-invariant}\Big)$$
$$\leq \exp\Big(-\frac{n\big(\frac{2c'}{\ell(S)^2} - c\big)^2}{128c_1^2}\Big).$$

The proof of Theorem 6 is presented in Appendix E.3. The theorem allows us to conclude:

Corollary 6.2. *For any $\epsilon, \delta > 0$ and any finite group G, Algorithm 1 can distinguish G-invariant probability measures from non-invariant measures with $\mathrm{KMaxIC}(\mu) \geq 2\epsilon$, with probability at least $1-\delta$, given*

$$n \geq \frac{128c_1^2\ell(S)^4}{\epsilon^2}\log\left(\frac{|S|}{\delta}\right),$$

i.i.d. samples, via the threshold $c = \dfrac{\epsilon}{\ell(S)^2}$. In other words, the sample complexity of Algorithm 1 is $\mathcal{O}\Big(\dfrac{\ell(S)^4}{\epsilon^2}\log\left(\dfrac{|S|}{\delta}\right)\Big)$.

In the next section, we provide detailed explanations about how to achieve appropriate generating sets for different finite groups to evaluate the results. Note that the runtime of Algorithm 1 depends linearly on $|S|$, which demands small size generating sets, while the sample complexity depends quadratically with $\ell(G)$, and it is also required to be small.

Remark 6.3. Algorithm 1 provides a hypothesis test with confidence level (i.e., Type I error) δ for the null hypothesis that μ is G-invariant with the acceptance threshold

$$c = \sqrt{\frac{-128c_1^2}{\epsilon^2}\log\left(\frac{\delta}{|S|}\right)},$$

where $c_1 := \sup_{x\in\mathcal{X}} K(x,x)$. Moreover, the Type II error (i.e., the probability of incorrectly accepting a non-invariant measure using Algorithm 1) vanishes as the sample size increases, as shown in Theorem 6. Hence, the test in Algorithm 1 is *consistent*, in the statistical sense.

7. Examples and Applications to Finite Groups

In this section, we evaluate the performance of Algorithm 1 across several well-known finite groups from the literature by computing their generating sets and analyzing their sample complexity.

7.1. Permutation Invariance Testing

To apply Algorithm 1 to the permutation group P_d, we need to find generating sets $S \subseteq P_d$ that minimize both $|S|$ and $\ell(S)$. To this end, we define $\sigma_i := (i \quad i+1)$ for each $i \in [d-1]$, meaning that σ_i swaps element i with $i+1$ while leaving the other elements unchanged. We then consider the following generating set:

$$S^\star := \Big\{\sigma_i \in P_d : i \in [d-1]\Big\}. \tag{3}$$

Proposition 3. *The set $S^\star \subseteq P_d$ defined via Equation (3) is a generating set for P_d and satisfies*

$$\ell(S^\star) \leq \frac{d(d-1)}{2}.$$

The proof of Proposition 3 is presented in Appendix D.10. This shows that one can use Algorithm 1 to test permutation invariance with sample complexity

$$n = \mathcal{O}\Big(\frac{d^8}{\epsilon^2}\log\left(\frac{d}{\delta}\right)\Big).$$

7.2. Sign-Flips Invariance Testing

The group of d-dimensional sign-flips F_d consists of 2^d diagonal matrices:

$$F_d \coloneqq \left\{ A = \mathrm{diag}(v) \in \mathbb{R}^{d\times d} : v \in \{\pm 1\}^d \right\}.$$

Although F_d is a large group, it can be generated simply using the following set:

$$S^\star \coloneqq \left\{ A = \mathrm{diag}\left(\mathbf{1}_d - 2e_i\right) \in \mathbb{R}^{d\times d} : i \in [d] \right\},$$

where $e_i \in \mathbb{R}^d$ denotes the unit vector in coordinate $i \in [d]$ and $\mathbf{1}_d \in \mathbb{R}^d$ denotes the all-one vector. Moreover, it is evident that $\ell(S^\star) = d$. Therefore, using Algorithm 1, one can test invariance to sign-flipping with sample complexity:

$$n = \mathcal{O}\left(\frac{d^4}{\epsilon^2} \log\left(\frac{d}{\delta}\right) \right).$$

7.3. Testing Invariances to Cyclic Groups

As a final application of testing invariance via KMaxIC, we study the cyclic group $G = \mathbb{Z}/m\mathbb{Z}$ with size m. Note that cyclic groups are generated by only one element, $1 \in \mathbb{Z}/m\mathbb{Z}$, but this is not an appropriate generating set since it has $\ell(S) = m$. To construct a generating set with smaller $\ell(G)$, consider the following set:

$$S^\star \coloneqq [m] \cap \left\{ 2^k : k = 0, 1, \dots \right\}. \tag{4}$$

Proposition 4. *The set $S^\star \subseteq G$ defined via Equation (4) is a generating set for G and satisfies*

$$\ell(S^\star) \leq \log_2(m).$$

The proof of Proposition 4 is presented in Appendix D.11. Note that this gives a much better bound compared to the one-element generating set. Indeed, using Algorithm 1 with $S^\star$ defined above provides a statistical test of invariance to cyclic groups with sample complexity:

$$n = \mathcal{O}\left(\frac{\log^4(m) + \log\left(\frac{1}{\delta}\right)}{\epsilon^2} \right).$$

8. Testing Invariances via KMaxIC: Infinite Groups

To apply Algorithm 1 to infinite groups, we need to find generating sets with small $\ell(G)$. However, unlike finite groups, infinite groups can only have generating sets S with $\ell(S) < \infty$ when $|S| = \infty$. Therefore, if we naively use a generating set S to apply Algorithm 1 to an infinite group, we would need to test over infinitely many group elements, which is impossible.

To resolve this issue, we fix a generating set $S \subseteq G$ with $\ell(S) < \infty$, and then refine it to a smaller finite set $\widehat{S} \subseteq S$ that provides an appropriate *covering* of the original set S. For simplicity, in this section, we focus on matrix groups consisting of orthogonal matrices $G \subseteq O(d)$ acting on $\mathcal{X} \subseteq \mathbb{R}^d$. The general case follows using a similar approach.

Definition 1 (Covering sets)**.** Given $S \subseteq O(d)$, we say that a finite set $\widehat{S} \subseteq S$ provides a γ-covering of S if and only if

$$\sup_{s\in S} \min_{\widehat{s}\in\widehat{S}} \|s - \widehat{s}\|_{\mathrm{op}} < \gamma,$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm of matrices.

Using the concept of covering sets, we can apply Algorithm 1 over $\widehat{S}$ with provable guarantees on both Type I and Type II errors:

Theorem 7. *Consider a PDS kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is a closed subset, and let $G \subseteq O(d)$ be an orthogonal subgroup acting on $\mathcal{X}$. Assume that $K(x, \cdot) : \mathcal{X} \to \mathbb{R}$ is an r-Lipschitz function with respect to the norm $\|\cdot\|_2$ on $\mathbb{R}^d$, for each $x \in \mathcal{X}$. Let $S \subseteq G$ be a generating set for G with $\ell(G) < \infty$, and let $\widehat{S}$ be a γ-covering of S.*

Then, when applying Algorithm 1 via $\widehat{S}$ to test invariance to G, the probability of a Type I error (i.e., incorrectly rejecting the invariance) is bounded as

$$\begin{aligned} \mathbb{P}\Big(\mathrm{H}_{\mathbf{1}} | \mathrm{H}_{\mathbf{0}}\Big) &= \mathbb{P}\Big(\max_{g \in \widehat{S}} \widehat{c}_g > c \,\Big|\, \mu \text{ is } G\text{-invariant}\Big) \\ &\leq |\widehat{S}| \exp\left(-\frac{nc^2}{128c_1^2}\right), \end{aligned}$$

where $c_1 := \sup_{x \in \mathcal{X}} K(x, x)$. Moreover, the Type II error, which is the probability of incorrectly accepting a non-invariant measure using Algorithm 1, approaches zero as the sample size increases. Specifically, for any probability measure $\mu \in \mathcal{P}(\mathcal{X})$ with $\mathbb{E}_{x\sim\mu}[\|x\|_2] \leq b$ such that $\mathrm{KMaxIC}(\mu) \geq 3c' > c\ell(S)^2 + 2rb\gamma$, we have

$$\begin{aligned} \mathbb{P}\Big(\mathrm{H}_{\mathbf{0}} | \mathrm{H}_{\mathbf{1}}\Big) &= \mathbb{P}\Big(\max_{g \in \widehat{S}} \widehat{c}_g \leq c \,\Big|\, \mu \text{ is not } G\text{-invariant}\Big) \\ &\leq \exp\left(-\frac{n\big(\frac{3c'}{\ell(S)^2} - 2r\gamma b - c\big)^2}{128c_1^2}\right). \end{aligned}$$

The proof of Theorem 7 is presented in Appendix E.4.

Similar to the case with finite groups, the test in Algorithm 1 is *statistically consistent* for infinite groups. Moreover, we conclude the following important result:

Corollary 8.2. *Let $G \subseteq O(d)$ denote an infinite group with a generating set $S \subseteq G$ such that $\ell(S) < \infty$, and let $\widehat{S} \subseteq S$ be a γ-covering of S with $\gamma = \dfrac{\epsilon}{2rb\ell(S)^2}$. Then, for any $\epsilon, \delta > 0$, Algorithm 1 can distinguish G-invariant probability measures from non-invariant measures with $\mathrm{KMaxIC}(\mu) \geq 3\epsilon$, with probability at least $1 - \delta$, given*

$$n \geq \frac{128c_1^2\ell(S)^4}{\epsilon^2} \log\left(\frac{|\widehat{S}|}{\delta}\right),$$

i.i.d. samples, via the threshold $c = \dfrac{\epsilon}{\ell(S)^2}$. In other words, the sample complexity of Algorithm 1 is $\mathcal{O}\left(\dfrac{\ell(S)^4}{\epsilon^2} \log\left(\dfrac{|\widehat{S}|}{\delta}\right)\right)$.

We conclude this section by noting that the method we used here to obtain upper bounds differs from traditional methods that focus on covering the entire group (e.g., group codes [12]). Here, we focused on covering the generating set, which, as we will see, allows for exact constructions for rotational symmetries $SO(d)$ in the next section.

9. Examples and Applications to Infinite Groups

In this section, we apply the theory from the previous section to an important infinite group testing problem: rotational symmetries, denoted by $SO(d)$ on $\mathcal{X} = \mathbb{R}^d$, assuming that $\mathbb{E}_{x\sim\mu}[\|x\|_2] \leq 1$. This group is formally defined as:

$$SO(d) := \left\{A \in \mathbb{R}^{d\times d} : AA^{\mathsf{T}} = I_d,\ \det(A) = 1\right\}.$$

To apply Algorithm 1, we need to find a generating set $S \subseteq SO(d)$ with small $\ell(S)$ and a good γ-covering $\widehat{S} \subseteq S$. Define $R_{ij}(\theta_{ij}) \in \mathbb{R}^{d\times d}$ to be the ordinary rotation matrix rotating in the ij-plane in $\mathbb{R}^d$ by an angle θ_{ij}, while keeping all other coordinates fixed. We use the following generating set:

$$S \coloneqq \left\{ R_{ij}(\theta_{ij}) : \theta_{ij} \in [0, 2\pi),\ i, j \in [d],\ i < j \right\}.$$

It is well-known that this set generates $SO(d)$. Specifically, for any $A \in SO(d)$, there exist angles θ_{ij} for $i, j \in [d]$, $i < j$, such that $A = \prod_{i<j} R_{ij}(\theta_{ij})$. Thus, S is a generating set for $SO(d)$ with $\ell(S) \leq \frac{d(d-1)}{2}$. Moreover, we can construct a finite γ-covering set $\widehat{S} \subseteq S$ as follows. Fix a parameter $k \in \mathbb{N}$, and for each $i < j$, define

$$\widehat{S}_{ij} \coloneqq \left\{ R_{ij}(\theta_{ij}) : \theta_{ij} = \frac{2\pi t}{k},\ t = 0, 1, \ldots, k-1 \right\},$$

and let $\widehat{S} \coloneqq \bigcup_{i<j} \widehat{S}_{ij}$. Note that the set $\widehat{S}$ contains $\frac{kd(d-1)}{2}$ elements. Moreover, there exists a constant c' such that

$$\sup_{\theta} \min_{t} \left\| R_{ij}(\theta) - R_{ij}\left(\frac{2\pi t}{k}\right) \right\|_{\text{op}} < \frac{c'}{k}.$$

Thus, to obtain a γ-covering, we set $k = \frac{c'}{\gamma}$.

To compute the sample complexity of Algorithm 1 using the proposed set $\widehat{S}$, we follow Corollary 8.2 and set $\gamma = \frac{\epsilon}{2r\ell(S)^2}$, which gives $k = \frac{2c'r\ell(S)^2}{\epsilon} = \mathcal{O}\left(\frac{d^4}{\epsilon}\right)$. This implies that $|\widehat{S}| = \frac{kd(d-1)}{2} = \mathcal{O}\left(\frac{d^6}{\epsilon}\right)$. We can now run Algorithm 1 with the threshold $c = \frac{\epsilon}{\ell(S)^2}$ to test invariance to $SO(d)$ with n i.i.d. samples.

Therefore, for any $\epsilon, \delta > 0$, Algorithm 1 can distinguish $SO(d)$-invariant probability measures from non-invariant ones with $\mathrm{KMaxIC}(\mu) \geq 3\epsilon$, with probability at least $1-\delta$, given $n = \mathcal{O}\left(\frac{d^8}{\epsilon^2} \log\left(\frac{d}{\delta}\right)\right)$, i.i.d. samples.

Remark 9.1. The method proposed in this section for exactly constructing coverings for $SO(d)$ also applies to many other matrix groups (such as $O(d)$ or Stiefel manifold), as we have explicit small generating sets for them. Here, we focused on rotational symmetries as an important application of our method, but it can be generalized to other well-known infinite groups as well.

10. Kernel Mean Invariance Criterion (KMIC)

In this section, we revisit the general recipe for testing invariances via the alternative hypothesis $\widetilde{\mathrm{H}}_\mathbf{1}$:

$$\widetilde{\mathrm{H}}_\mathbf{1} : \mathbb{E}_g[D(\mu, g\mu)] \geq \epsilon.$$

In other words, we focus on proposing algorithms for the hypothesis testing described by H_0 versus $\widetilde{\mathrm{H}}_\mathbf{1}$ in Formulation 2. Similar to KMaxIC, here we focus on the special case of kernel Maximum Mean Discrepancy (MMD) distances $D \equiv \mathrm{MMD}$. Observe that according to Proposition 2,

$$\begin{aligned} D^2(\mu, g\mu) &= \left\| (g\mu)_\mathcal{H} - \mu_\mathcal{H} \right\|_\mathcal{H}^2 \\ &= 2\mathbb{E}_{x,x'}[K(x, x')] - 2\mathbb{E}_{x,x'}[K(x, gx')], \end{aligned}$$

where $x, x' \sim \mu$ independently, and $g \in G$ is chosen uniformly at random and independently of x and x'. This means that

$$\mathbb{E}_g[D^2(\mu, g\mu)] = 2\mathbb{E}_{x,x'}[K(x, x')] - 2\mathbb{E}_{g,x,x'}[K(x, gx')].$$

Let μ^G denote the distribution of gx, where $x \sim \mu$ and $g \in G$ in uniformly distributed over the group. Surprisingly, for shift-invariant kernels, we also have the following identity:

$$2\big\|\mu_{\mathcal{H}}^G - \mu_{\mathcal{H}}\big\|_{\mathcal{H}}^2 = 2\mathbb{E}_{x,x'}[K(x,x')] - 2\mathbb{E}_{g,x,x'}[K(x,gx')],$$

See Proposition 6 for a proof. This means that

$$\mathbb{E}_g[D^2(\mu, g\mu)] = 2\big\|\mu_{\mathcal{H}}^G - \mu_{\mathcal{H}}\big\|_{\mathcal{H}}^2.$$

The right hand side of the above identity, termed as the Kernel Mean Invariance Criterion (KMIC) in this paper, is also introduced recently as a measure of closeness to invariance.

Definition 1 ([28])**.** Let $\mu \in \mathcal{P}(\mathcal{X})$. The Kernel Mean Invariance Criterion (KMIC) is defined as

$$\mathrm{KMIC}(\mu) \coloneqq \big\|\mu_{\mathcal{H}}^G - \mu_{\mathcal{H}}\big\|_{\mathcal{H}}^2,$$

where $\mu_{\mathcal{H}}^G, \mu_{\mathcal{H}} \in \mathcal{H}$ are the kernel mean embeddings of μ^G and μ, respectively.

KMIC also quantifies the distance to G-invariance: $\mathrm{KMIC}(\mu) = 0$ if and only if μ is G-invariant, assuming the kernel is universal (Appendix C). More importantly, using our main result (Theorem 3), we have

$$\begin{aligned}\mathrm{KMaxIC}(\mu) = \sup_{g\in G} D^2(\mu, g\mu) &\leq 16\big(\mathbb{E}_g[D(\mu, g\mu)]\big)^2 \\ &\leq 16\mathbb{E}_g[D^2(\mu, g\mu)] = 32\,\mathrm{KMIC}(\mu).\end{aligned}$$

Moreover, we also have

$$\begin{aligned}\mathrm{KMaxIC}(\mu) = \sup_{g\in G} D^2(\mu, g\mu) &\geq \mathbb{E}_g[D^2(\mu, g\mu)] \\ &= 2\,\mathrm{KMIC}(\mu).\end{aligned}$$

Therefore, we conclude that the optimal convergence rates and the Type I and Type II error bounds for both tests according to KMIC and KMaxIC are equivalent to each other, up to constant factors. In other words, while KMIC only provides an averaged measure of being invariance, it also provides an algorithm, *robust* to all group transformations.

We provide a detailed review of testing invariance via KMIC as well as a detailed study of its convergence rate and Type I and Type II errors in Appendix C. The corresponding testing algorithm is also presented in Algorithm 2.

11. KMIC vs. KMaxIC: A Discussion and Comparison

In this paper, we proposed and analyzed two distinct methods for deriving testing algorithms: KMaxIC (Algorithm 1) and KMIC (Algorithm 2). Thanks to Theorem 3, the two measures of distance to invariance are equivalent up to a constant factor. Here, we provide a brief discussion on the differences between their corresponding algorithms..

First, note that testing via KMIC is a *randomized* algorithm, as it involves generating n i.i.d. uniform samples from the group to achieve μ^G. On the other hand, KMaxIC offers a *deterministic* testing algorithm, with no need to sample from G, unlike KMIC. While the KMIC testing algorithm requires n i.i.d. samples from G, KMaxIC evaluates invariance over a *fixed* subset of the group, which remains independent of the number of samples.

Note that to propose a testing algorithm according to KMaxIC formulation, one needs to specifically construct generating sets and coverings which needs problem specific designs. However, KMIC allows to achieve a *universal* testing algorithm, which needs no design other than being able to uniformly sample from the group.

12. Experiments

In this section, we examine Theorem 3 on synthetic data to validate the constant factor approximation. Since the problem is intractable for large groups (Theorem 2), we focus on small-sized groups of rotational symmetries.

We consider two-dimensional data $x \in \mathbb{R}^2$ generated independently according to a zero-mean multivariate Gaussian distribution $\mu = \mathcal{N}(0, \Sigma)$, where $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$. Moreover, we work with a group of rotational symmetries of size $k \in \mathbb{N}$:

$$G = \left\{ R\Big(\frac{2\pi t}{k}\Big) : t = 0, 1, \dots, k-1 \right\},$$

where $R(\theta) \coloneqq \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \in \mathbb{R}^{2\times 2}$. Let $\widehat{\mu}$ denote the empirical measure obtained from the data.

In our experiments, we use $n = 2000$ data points and consider a rotational group of size $k = 100$. We adopt the 1-Wasserstein distance as the metric on probability measures, formulated through the optimal transport problem (i.e., we instantiate Theorem 3 with $D \equiv W$). In Figure 1, we plot the optimal transport distance $W(\widehat{\mu}, g\widehat{\mu})$ for all $g \in G$ and its average over $g \in G$. The parameter $\theta = \frac{2\pi t}{k}$ runs from 0 to 2π, representing all group elements.

As observed in Figure 1, the function $W(\widehat{\mu}, g\widehat{\mu})$ is not concave over $[0, 2\pi]$, aligning with Theorem 2, which states that the overall maximization problem $\sup_{g\in G} W(\widehat{\mu}, g\widehat{\mu})$ is generally intractable. Furthermore, by plotting the ratio between $W(\widehat{\mu}, g\widehat{\mu})$ and $\sup_{g\in G} W(\widehat{\mu}, g\widehat{\mu})$, we observe that it is uniformly bounded above over the group by a constant (approximately 1.85). This is consistent with Theorem 3, which proves a constant factor of four approximation through randomization.

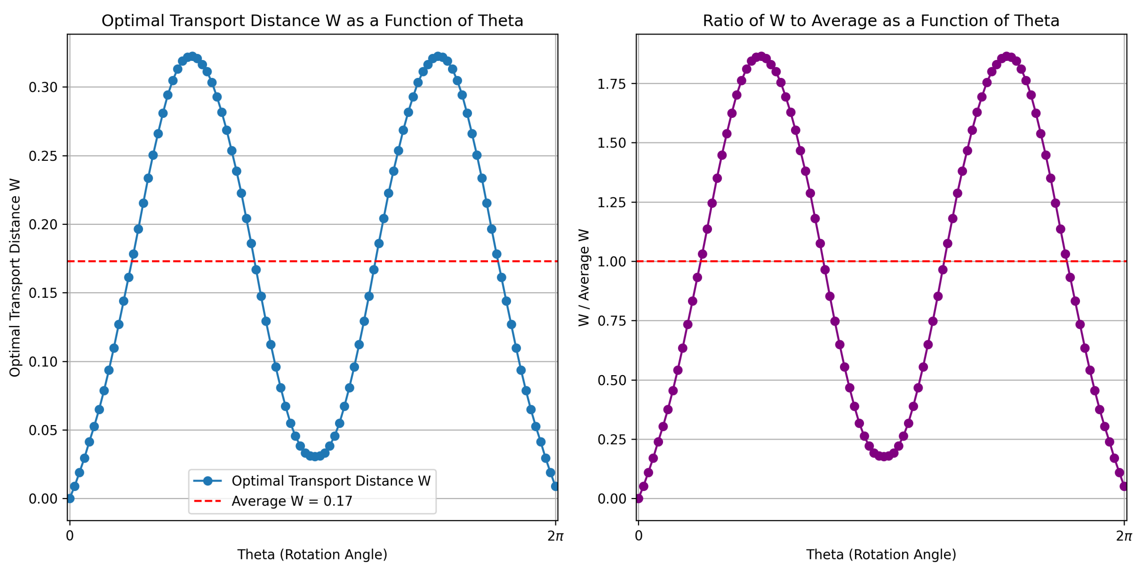

Figure 1: A constant-factor approximation of the worst-case optimal transport distance, $\sup_{g\in G} W(\widehat{\mu}, g\widehat{\mu})$ where G is the group of rotational symmetries in two dimensions, and $\widehat{\mu}$ is the empirical measure obtained from n samples of a non-isotropic multivariate Gaussian distribution.

13. Conclusion

In this paper, we study the robust formulation of testing invariance to a set of group transformations. We prove that the robust distance to invariance, defined via probability metrics, while intractable to compute exactly, can be approximated within constant factors through randomization. As a result, we propose a general framework for robust testing of invariances using a new hypothesis testing formulation. Next, we focus on distances defined via kernel methods, specifically Maximum Mean Discrepancies (MMDs), and propose deterministic algorithms for robust testing with respect to both finite and infinite groups, based on generating sets of groups and coverings. Finally, we prove that another studied metric for measuring closeness to invariance, defined via group averaging and kernels, is equivalent to the robust metric up to multiplicative constants. This leads to the result that the group-averaged metric is robust, and we propose randomized testing algorithms with plausible performance.

References

[1] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4): 18–42, 2017. 1

[2] Eugene P Wigner. Invariance in physical theory. *Proceedings of the American Philosophical Society*, 93(7):521–526, 1949. 1

[3] Eugene P Wigner. Events, laws of nature, and invariance principles. *Science*, 145(3636):995–999, 1964.

[4] Tess E Smidt. Euclidean symmetry and equivariance in machine learning. *Trends in Chemistry*, 3(2):82–85, 2021. 1

[5] Peter H Westfall and S Stanley Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons, 1993. 1

[6] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9): 5116–5121, 2001.

[7] Marti J Anderson and John Robinson. Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, 43(1):75–88, 2001.

[8] Patrick Onghena. Randomization tests or permutation tests? a historical and terminological clarification. In *Randomization, masking, and allocation concealment*, pages 209–228. Chapman and Hall/CRC, 2017.

[9] Jesse Hemerik and Jelle J Goeman. Another look at the lady tasting tea and differences between permutation tests and randomisation tests. *International Statistical Review*, 89(2):367–381, 2021.

[10] Nick W Koning and Jesse Hemerik. More efficient exact group invariance testing: using a representative subgroup. *Biometrika*, 111(2):441–458, 2024. 1, 3

[11] Erich Leo Lehmann, Joseph P Romano, and George Casella. *Testing statistical hypotheses*, volume 3. Springer, 1986. 1

[12] Nick W Koning. More power by using fewer permutations. *Biometrika*, page asae031, 2024. 1, 3, 9

[13] Jesse Hemerik. On the term “randomization test”. *The American Statistician*, pages 1–8, 2024. 1

[14] RA Fisher. The design of experiments. 1935. 1

[15] Ronald Aylmer Fisher, Ronald Aylmer Fisher, Statistiker Genetiker, Ronald Aylmer Fisher, Statistician Genetician, Great Britain, Ronald Aylmer Fisher, and Statisticien Généticien. *The design of experiments*, volume 21. Springer, 1966. 1

[16] Bradley Efron. Student's t-test under symmetry conditions. *Journal of the American Statistical Association*, 64(328):1278–1302, 1969. 1

[17] T Eden and F Yates. On the validity of fisher's z test when applied to an actual example of non-normal data.(with five text-figures.). *The Journal of Agricultural Science*, 23(1):6–17, 1933. 1

[18] Eric L Lehmann and Charles Stein. On the theory of some non-parametric hypotheses. *The Annals of Mathematical Statistics*, 20(1):28–45, 1949. 1

[19] Øyvind Langsrud. Rotation tests. *Statistics and computing*, 15:53–60, 2005. 1

[20] Patrick O Perry and Art B Owen. A rotation test to verify latent structure. *Journal of Machine Learning Research*, 11(2), 2010.

[21] Aldo Solari, Livio Finos, and Jelle J Goeman. Rotation-based multiple testing in the multivariate linear model. *Biometrics*, 70(4):954–961, 2014.

[22] Di Wu, Elgene Lim, François Vaillant, Marie-Liesse Asselin-Labat, Jane E Visvader, and Gordon K Smyth. Roast: rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17):2176–2182, 2010. 1

[23] Ho Chung Law, Christopher Yau, and Dino Sejdinovic. Testing and learning on distributions with symmetric noise invariance. *Advances in Neural Information Processing Systems*, 30, 2017. 3

[24] Alexis Bellot and Mihaela van der Schaar. Application of kernel hypothesis testing on set-valued data. In *Uncertainty in Artificial Intelligence*, pages 194–204. PMLR, 2021. 3

[25] Edgar Dobriban. Consistency of invariance-based randomization tests. *The Annals of Statistics*, 50(4):2443–2466, 2022. 3

[26] Adam B Kashlak. Asymptotic symmetry and group invariance for randomization. *arXiv preprint arXiv:2211.00144*, 2022. 3

[27] Aaditya Ramdas, Rina Foygel Barber, Emmanuel J Candès, and Ryan J Tibshirani. Permutation tests using arbitrary permutation distributions. *Sankhya A*, 85(2):1156–1177, 2023. 3

[28] Kenny Chiu and Benjamin Bloem-Reddy. Hypothesis tests for distributional group symmetry with applications to particle physics. In *NeurIPS 2023 AI for Science Workshop*, 2023. 3, 11, 19

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 17

[30] Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012. 17

[31] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017. 17

[32] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 17

[33] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008. 17

[34] Allan Zhou, Tom Knowles, and Chelsea Finn. Meta-learning symmetries by reparameterization. In *International Conference on Learning Representations*, 2021. 17

[35] Nima Dehmamy, Robin Walters, Yanchen Liu, Dashun Wang, and Rose Yu. Automatic symmetry discovery with lie algebra convolutional network. *Advances in Neural Information Processing Systems*, 34:2503–2515, 2021.

[36] Artem Moskalev, Anna Sepliarskaia, Ivan Sosnovik, and Arnold Smeulders. Liegg: Studying learned lie group generators. *Advances in Neural Information Processing Systems*, 35:25212–25223, 2022.

[37] Jianke Yang, Robin Walters, Nima Dehmamy, and Rose Yu. Generative adversarial symmetry discovery. In *International Conference on Machine Learning*, pages 39488–39508. PMLR, 2023.

[38] Jianke Yang, Nima Dehmamy, Robin Walters, and Rose Yu. Latent space symmetry discovery. In *International Conference on Machine Learning*, 2023. 17

[39] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21*, pages 44–51. Springer, 2011. 17

[40] Hong-Xing Yu, Jiajun Wu, and Li Yi. Rotationally equivariant 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1456–1464, 2022. 17

[41] Hugo Caselles-Dupré, Michael Garcia Ortiz, and David Filliat. Symmetry-based disentangled representation learning requires interaction with environments. *Advances in Neural Information Processing Systems*, 32, 2019. 17

[42] Robin Quessard, Thomas Barrett, and William Clements. Learning disentangled representations and group structure of dynamical environments. *Advances in Neural Information Processing Systems*, 33:19727–19737, 2020.

[43] Giovanni Luca Marchetti, Gustaf Tegnér, Anastasiia Varava, and Danica Kragic. Equivariant representation learning via class-pose decomposition. In *International Conference on Artificial Intelligence and Statistics*, pages 4745–4756. PMLR, 2023. 17

[44] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997. 17

[45] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006. 17

[46] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012. 17, 18, 22, 24

[47] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Int. conference on Algorithmic Learning Theory (ALT)*, 2005. 17

[48] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007. 17

[49] Youssef Mroueh, Tom Sercu, Mattia Rigotti, Inkit Padhi, and Cicero Nogueira dos Santos. Sobolev independence criterion. *Advances in Neural Information Processing Systems*, 32, 2019. 17

[50] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011. 17

[51] Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *UAI*, pages 132–141, 2014. 17

[52] Ilya Tolstikhin, Bharath K Sriperumbudur, Krikamol Mu, et al. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18(86):1–47, 2017. 17

[53] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[54] Ingmar Schuster, Mattes Mollenhauer, Stefan Klus, and Krikamol Muandet. Kernel conditional density operators. In *International Conference on Artificial Intelligence and Statistics*, pages 993–1004. PMLR, 2020.

[55] Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. *Advances in neural information processing systems*, 33:21247–21259, 2020.

[56] Krikamol Muandet, Motonobu Kanagawa, Sorawit Saengkyongam, and Sanparith Marukatat. Counterfactual mean embeddings. *Journal of Machine Learning Research*, 22(162):1–71, 2021.

[57] Boris Muzellec, Francis Bach, and Alessandro Rudi. A note on optimizing distributions using kernel mean embeddings. *arXiv preprint arXiv:2106.09994*, 2021.

[58] Cristopher Salvi, Maud Lemercier, Chong Liu, Blanka Horvath, Theodoros Damoulas, and Terry Lyons. Higher order kernel mean embeddings to capture filtrations of stochastic processes. *Advances in Neural Information Processing Systems*, 34:16635–16647, 2021.

[59] Jonas M Kübler, Wittawat Jitkrittum, Bernhard Schölkopf, and Krikamol Muandet. A witness two-sample test. In *International Conference on Artificial Intelligence and Statistics*, pages 1403–1419. PMLR, 2022.

[60] Jonas M Kübler, Vincent Stimper, Simon Buchholz, Krikamol Muandet, and Bernhard Schölkopf. Automl two-sample test. *Advances in Neural Information Processing Systems*, 35:15929–15941, 2022.

[61] Antoine Chatalic, Nicolas Schreuder, Lorenzo Rosasco, and Alessandro Rudi. Nyström kernel mean embeddings. In *International Conference on Machine Learning*, pages 3006–3024. PMLR, 2022. 17

[62] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019. 23, 30

A. Additional Related Work

Learning and Symmetries. Designing invariant machine learning models by construction has a rich and long-standing history. To name a few, Convolutional Neural Networks (CNNs) were introduced to exploit local shift-invariance structures in images [29, 30]. Deep Sets were developed to handle set-structured data [31], and PointNets were proposed for point cloud data that are invariant to permutations [32]. Graph Neural Networks (GNNs) [33] were designed for graph-structured data.

Recent efforts have explored alternative approaches, such as automatically discovering the underlying symmetries in data [34–38]. Another line of work focuses on learning equivariant representations given known symmetries [39, 40], particularly targeting symmetric disentangled representations [41–43]. Despite this extensive body of work, the problem of testing invariances—central to our work—remains relatively underexplored in the machine learning literature.

Kernels and Embedding of Distributions. The relationship between kernels and distributions has been extensively studied over the past decades. Müller [44] introduced the notion of the Integral Probability Metric (IPM) over a function class. Gretton et al. [45, 46] coined the term Maximum Mean Discrepancy (MMD) when the function class is restricted to a Reproducing Kernel Hilbert Space (RKHS). They showed that under the universality assumption of the RKHS, the MMD distance is definite, meaning $\mathrm{MMD}(p, q) = 0$ if and only if $p = q$. This led to the development of two-sample testing using empirical MMD estimates.

Gretton et al. [47, 48] introduced the Hilbert-Schmidt Independence Criterion (HSIC) as a measure of independence between random variables, defined as the Hilbert-Schmidt norm of the cross-covariance operator. They demonstrated that independence could be tested using observations in the form of universal kernels. Mroueh et al. [49] extended these ideas to gradient-regularized IPM and explored its applications in feature selection.

Sriperumbudur et al. [50] characterized the relationship between characteristic and universal kernels, providing necessary and sufficient conditions for the bijectivity of the kernel mean embedding of distributions. Doran et al. [51] reduced kernel-based conditional independence testing to kernel two-sample tests through permutations.

This area of research has seen continuous development. We conclude by highlighting a subset of recent works contributing to this line of inquiry [52–61].

B. Background

In this section, we provide the necessary background group actions and kernels used in the paper.

B.1. Group Actions and Invariant Measures

The continuous action of a compact topological group G on a complete metric space $\mathcal{X}$ is defined by a continuous function $\theta : G \times \mathcal{X} \to \mathcal{X}$, such that for each $g \in G$, the mapping $\theta(g, \cdot)$ is a homeomorphism on $\mathcal{X}$. Additionally, it satisfies the property $\theta(g_2, \theta(g_1, x)) = \theta(g_2 g_1, x)$ for any $g_1, g_2 \in G$ and any $x \in \mathcal{X}$. For brevity, we denote the action of $g \in G$ on $x \in \mathcal{X}$ as $gx := \theta(g, x)$. We endow the group G with its associated unique (left) Haar probability measure, which provides the uniform distribution over the group elements.

Examples of groups acting on spaces include the permutation group P_d, which acts on $\mathbb{R}^d$ via permutation matrices, and the orthogonal group $O(d)$, which acts on $\mathbb{R}^d$ via orthogonal matrices.

Let $\mathcal{P}(\mathcal{X})$ denote the space of all Borel probability measures on $\mathcal{X}$. For each $\mu \in \mathcal{P}(\mathcal{X})$, let $g\mu \in \mathcal{P}(\mathcal{X})$ be a Borel probability measure defined by $(g\mu)(A) = \mu(g^{-1}A)$ for any Borel-measurable set $A \subseteq \mathcal{X}$ and any group element $g \in G$, where $gA := \{ga : a \in A\}$. We say that $\mu \in \mathcal{P}(\mathcal{X})$ is *G-invariant* if and only if $\mu = g\mu$ for all $g \in G$.

In particular, a probability measure is invariant with respect to the action of a group G if and only if it assigns the same probabilities to each event and its "shifted version" according to the group action. For example, Gaussian random variables define G-invariant probability measures on $\mathbb{R}^d$ with respect to the group of orthogonal matrices $G = O(d)$.

B.2. Positive Definite Symmetric Kernels

Let $\mathcal{X}$ be a complete metric space. A Positive Definite Symmetric (PDS) kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a continuous symmetric function with the following property: for any positive integer n and any points $x_1, x_2, \ldots, x_n \in \mathcal{X}$, the Gram matrix $[K(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{n\times n}$ is positive semi-definite.

Kernels serve as measures of similarity. Notable examples of PDS kernels include the Gaussian kernel, defined as $K(x_1, x_2) = \exp\left(-\frac{1}{2\sigma^2}\|x_1 - x_2\|_2^2\right)$, where the kernel is defined over the space $\mathcal{X} = \mathbb{R}^d$.

Let $L^2(\mathcal{X})$ denote the space of square-integrable real-valued functions on $\mathcal{X}$. For each PSD kernel K, there is an associated Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H} \subseteq L^2(\mathcal{X})$ with an inner product denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, which satisfies the following properties:

- For each point $x \in \mathcal{X}$, the *feature* function $\Phi(x) = K(\cdot, x)$ belongs to the RKHS $\mathcal{H}$.
- For any $f \in \mathcal{H}$ and any $x \in \mathcal{X}$, we have the reproducing property: $f(x) = \langle f, K(\cdot, x)\rangle_{\mathcal{H}}$.

Combining these two properties, we find that $K(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2)\rangle_{\mathcal{H}}$ for all $x_1, x_2 \in \mathcal{X}$.

Note. For technical reasons, we consider uniformly bounded kernels: $\sup_{x\in\mathcal{X}} K(x, x) < \infty$.

B.3. Shift-Invariant Kernels

As mentioned earlier, kernels introduce similarity measures on metric spaces. The concept of a *shift-invariant* kernel refers to those kernels that measure similarity regardless of how the pair of points is shifted according to a given group action.

Definition 1. Given a compact topological group G acting continuously on a complete metric space $\mathcal{X}$, and a PSD kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we say that K is *shift-invariant* if and only if

$$K(x_1, x_2) = K(gx_1, gx_2),$$

for any $g \in G$ and any $x_1, x_2 \in \mathcal{X}$.

For example, the Gaussian kernel is shift-invariant with respect to $G = O(d)$.

B.4. Kernel Mean Embeddings of Measures

For any probability measure $\mu \in \mathcal{P}(\mathcal{X})$ and any PSD kernel K, the kernel mean embedding of μ, denoted as $\mu_{\mathcal{H}}$, is a unique element of the RKHS $\mathcal{H}$ that satisfies the following identity:

$$\langle f, \mu_{\mathcal{H}}\rangle_{\mathcal{H}} = \mathbb{E}_{x\sim\mu}[f(x)] = \mathbb{E}_{x\sim\mu}[\langle f, \phi(x)\rangle],$$

for each $f \in \mathcal{H}$. The existence and uniqueness of such a $\mu_{\mathcal{H}} \in \mathcal{H}$ are guaranteed by the Riesz representation theorem for Hilbert spaces (see, for instance, [46]), therefore it can be inferred that $\mu_{\mathcal{H}} = \mathbb{E}_{x\sim\mu}[\phi(x)]$.

It is well-known that one can also uniquely recover the original probability measure μ from its kernel mean embedding, provided that the PSD kernel K is *universal*. A PSD kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with an RKHS $\mathcal{H}$ is said to be universal if, for any continuous function $f : \mathcal{X} \to \mathbb{R}$ and any positive ϵ, there exists a function $\hat{f} \in \mathcal{H}$ such that $\sup_{x\in\mathcal{X}} |f(x) - \hat{f}(x)| < \epsilon$. The ability to uniquely recover probability measures from their kernel mean embeddings leads to the following definition of Maximum Mean Discrepancy (MMD) as a metric for comparing probability measures:

$$\mathrm{MMD}(\mu, \nu) := \|\mu_{\mathcal{H}} - \nu_{\mathcal{H}}\|_{\mathcal{H}},$$

for any $\mu, \nu \in \mathcal{P}(\mathcal{X})$.

C. Kernel Mean Invariance Criterion (KMIC)

In this section, we provide a detailed review of the properties of KMIC [28].

The idea of KMIC is to construct a canonical G-invariant probability measure via *group averaging*, and then compare it to the original measure using the Maximum Mean Discrepancy (MMD) metric to quantify how far the measure is from being G-invariant.

Proposition 5 (Invariant Measure)**.** *Let $\mu \in \mathcal{P}(\mathcal{X})$ be a probability measure defined on a complete metric space $\mathcal{X}$, and let G be a compact topological group acting continuously on $\mathcal{X}$. For each measurable set $A \subseteq \mathcal{X}$, define*

$$\mu^G(A) := \mathbb{E}_g[(g\mu)(A)] = \mathbb{E}_g[\mu(gA)],$$

where the expectation is over uniformly sampled $g \in G$, according to its unique (left) Haar probability measure. Then, $\mu^G \in \mathcal{P}(\mathcal{X})$ defines a G-invariant (Borel) probability measure on $\mathcal{X}$.

The proof of Proposition 5 is presented in Appendix D.1.

This proposition motivates the following definition of the Kernel Mean Invariance Criterion (KMIC).

Definition 1 (Kernel Mean Invariance Criterion (KMIC))**.** Let $\mathcal{X}$ be a complete metric space and let G be a compact topological group acting continuously on $\mathcal{X}$. For any probability measure $\mu \in \mathcal{P}(\mathcal{X})$, the Kernel Mean Invariance Criterion (KMIC) is defined as

$$\mathrm{KMIC}(\mu) := \left\|\mu_{\mathcal{H}}^G - \mu_{\mathcal{H}}\right\|_{\mathcal{H}}^2,$$

where $\mu_{\mathcal{H}}^G, \mu_{\mathcal{H}} \in \mathcal{H}$ denote the kernel mean embeddings of the probability measures μ^G and μ, respectively.

Note that $\mathrm{KMIC}(\mu) \geq 0$ for all $\mu \in \mathcal{P}(\mathcal{X})$. Moreover, KMIC provides a notion of distance to being G-invariant: $\mathrm{KMIC}(\mu) = 0$ if and only if μ is G-invariant.

Theorem 8 (Definiteness of KMIC)**.** *Let K be a universal PDS kernel defined on a complete metric space $\mathcal{X}$. Let $\mu \in \mathcal{P}(\mathcal{X})$ be a probability measure. Then, $\mathrm{KMIC}(\mu) = 0$ if and only if μ is G-invariant.*

The proof of Theorem 8 is presented in Appendix D.2.

The above theorem demonstrates how KMIC is capable of distinguishing probability measures from their canonical G-invariant probability measures. However, to propose statistical tests using KMIC, an efficient representation is necessary to compute it using i.i.d. samples. The following proposition facilitates this representation.

Proposition 6. *Consider a shift-invariant PDS kernel K defined on the complete metric space $\mathcal{X}$. Then, KMIC can be alternatively represented as*

$$\mathrm{KMIC}(\mu) = \mathbb{E}_{x,x'}[K(x, x')] - \mathbb{E}_{g,x,x'}[K(x, gx')],$$

where $x, x' \sim \mu$ independently, and $g \in G$ is chosen uniformly at random and independently of x and x'.

The proof of Proposition 6 is presented in Appendix D.3. While we focused on shift-invariant kernels in the above proposition, a general formula for arbitrary kernels is derived in the proof.

C.1. Testing Invariances via KMIC

Given n i.i.d. samples $x_i \sim \mu$, $i \in [n]$, how can one provide estimates of $\mathrm{KMIC}(\mu)$? Proposition 6 allows us to provide empirical estimates from data:

$$\widehat{\mathrm{KMIC}}(\mu) = \frac{2}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} K(x_i, x_j) - \frac{2}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} K(x_i, g_j x_j).$$

Here, we utilize n i.i.d. samples g_j, $j \in [n]$, each uniformly distributed over the group G. Note that $\widehat{\mathrm{KMIC}}(\mu)$, as a sum of two U-statistics, provides an unbiased estimator for $\mathrm{KMIC}(\mu)$.

Algorithm 2 Testing Invariances via KMIC

Input: n i.i.d. samples $x_i \sim \mu$, $i \in [n]$, and a threshold $c \in (0, \infty)$.

1: Generate n i.i.d. samples $g_j \in G$, $j \in [n]$, each uniformly distributed over G.

2: Compute the following:

$$\widehat{\mathrm{KMIC}}(\mu) = \frac{2}{n(n-1)} \sum_{\substack{i,j=1 \\ i<j}}^{n} K(x_i, x_j) - \frac{2}{n(n-1)} \sum_{\substack{i,j=1 \\ i<j}}^{n} K(x_i, g_j x_j).$$

3: **if** $\widehat{\mathrm{KMIC}}(\mu) \leq c$ **then**

4: **return** There is not enough evidence to reject the null hypothesis H_0 that μ is G-invariant.

5: **else**

6: **return** $\widetilde{\mathrm{H}}_{\mathbf{1}}$: μ is not G-invariant.

7: **end if**

The above estimator gives rise to Algorithm 2, a hypothesis testing algorithm with a threshold $c \in (0, \infty)$.

It is worth mentioning that the total runtime of Algorithm 2 is $\mathcal{O}(n^2)$, provided that we can sample from G and compute the kernel function for each pair of points in constant time. Moreover, the time complexity can be further improved to $\mathcal{O}(n)$ by modifying the algorithm and replacing the U-statistics with empirical estimates over disjoint pairs of independent samples.

C.2. Convergence Rates and Confidence Intervals for KMIC

In this section, we analyze Algorithm 2. First, we derive the convergence rate of the empirical estimator for $\mathrm{KMIC}(\mu)$, and then we focus on obtaining confidence intervals to design the parameter $c \in (0, \infty)$ appropriately.

Theorem 9 (Convergence rate for $\widehat{\mathrm{KMIC}}(\mu)$)**.** *For the estimator $\widehat{\mathrm{KMIC}}(\mu)$ defined in Algorithm 2, we have*

$$\mathbb{E}\left[\left|\widehat{\mathrm{KMIC}}(\mu) - \mathrm{KMIC}(\mu)\right|^2\right] \lesssim \frac{c_1^2}{n}, \tag{5}$$

where $c_1 \coloneqq \sup_{x \in \mathcal{X}} K(x, x)$.

The proof of Theorem 9 is presented in Appendix D.4.

The above result shows that the estimator provided in Algorithm 2 converges in the *mean*. However, to design statistical hypothesis tests, it is desirable to obtain confidence intervals based on the threshold $c \in (0, \infty)$. The following theorem provides such bounds.

Theorem 10. *For the estimator $\widehat{\mathrm{KMIC}}(\mu)$ defined in Algorithm 2, we have*

$$\mathbb{P}\left(\left|\widehat{\mathrm{KMIC}}(\mu) - \mathrm{KMIC}(\mu)\right| \geq t\right) \leq 4 \exp\left(-\frac{nt^2}{32 c_1^2}\right),$$

where $c_1 \coloneqq \sup_{x \in \mathcal{X}} K(x, x)$.

The proof of Theorem 10 is presented in Appendix D.5. We note that the result above provides confidence intervals for estimating $\mathrm{KMIC}(\mu)$ from data. Specifically, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have

$$\widehat{\mathrm{KMIC}}(\mu) \in \left[\mathrm{KMIC}(\mu) - \sqrt{\frac{32 c_1^2}{n} \log\left(\frac{4}{\delta}\right)},\ \mathrm{KMIC}(\mu) + \sqrt{\frac{32 c_1^2}{n} \log\left(\frac{4}{\delta}\right)}\right].$$

In other words, we have

$$\mathbb{P}\left(\widehat{\mathrm{KMIC}}(\mu) > c \,\middle|\, \mu \text{ is } G\text{-invariant}\right) \leq \delta,$$

whenever $n \geq \dfrac{32c_1^2 \log\left(\frac{4}{\delta}\right)}{c^2}$. This result shows that with an appropriate choice of the threshold c, the Type I error of the proposed statistical test (i.e., the probability of failing to detect invariances in data generated according to a G-invariant probability measure) is at most δ.

Corollary C.2. *Algorithm 2 provides a hypothesis test with confidence level δ for the null hypothesis that μ is G-invariant, with the acceptance threshold given by $c = \sqrt{\frac{32c_1^2}{n}\log\left(\frac{4}{\delta}\right)}$, where $c_1 := \sup_{x\in\mathcal{X}} K(x,x)$.*

Moreover, the Type II error, which is the probability of incorrectly accepting a non-invariant measure using Algorithm 2, approaches zero as the sample size increases (Theorem 10). This demonstrates that the test in Algorithm 2 is *consistent* in the statistical sense. Quantitatively, for any probability measure $\mu \in \mathcal{P}(\mathcal{X})$ such that $\mathrm{KMIC}(\mu) \geq 2c$, we have

$$\mathbb{P}\left(\widehat{\mathrm{KMIC}}(\mu) \leq c \,\middle|\, \mu \text{ is not } G\text{-invariant}\right) \leq \delta,$$

whenever $n \geq \dfrac{32c_1^2 \log\left(\frac{4}{\delta}\right)}{c^2}$. This shows that Algorithm 2 with threshold c can distinguish G-invariant probability measures from non-invariant ones with $\mathrm{KMIC}(\mu) \geq 2c$, with sample complexity $n = \dfrac{32c_1^2 \log\left(\frac{4}{\delta}\right)}{c^2}$, with probability at least $1-\delta$.

D. Proofs

D.1. Proof of Proposition 5

Proposition 5 (Invariant Measure). *Let $\mu \in \mathcal{P}(\mathcal{X})$ be a probability measure defined on a complete metric space $\mathcal{X}$, and let G be a compact topological group acting continuously on $\mathcal{X}$. For each measurable set $A \subseteq \mathcal{X}$, define*

$$\mu^G(A) := \mathbb{E}_g[(g\mu)(A)] = \mathbb{E}_g[\mu(gA)],$$

where the expectation is over uniformly sampled $g \in G$, according to its unique (left) Haar probability measure. Then, $\mu^G \in \mathcal{P}(\mathcal{X})$ defines a G-invariant (Borel) probability measure on $\mathcal{X}$.

Proof. We start the proof by showing that μ^G is a valid (Borel) probability measure, in order to do so, we show that it satisfies the following conditions, and hence it is a valid Borel measure.

- $\mu^G(\emptyset) = \mathbb{E}_g[\mu(g\emptyset)] = \mathbb{E}_g[\mu(\emptyset)] = 0$.
- Countable additivity: if $\{A_i\}_i^\infty$ is a sequence of disjoint sets belonging to Borel σ-field, then

$$\begin{aligned}
\mu^G(\bigcup_{i=1}^\infty A_i) &= \mathbb{E}_g[\mu(g\bigcup_{i=1}^\infty A_i)] \\
&= \mathbb{E}_g[\mu(\bigcup_{i=1}^\infty gA_i)] && (6)\\
&= \mathbb{E}_g[\sum_{i=1}^\infty \mu(gA_i)] && (7)\\
&= \sum_{i=1}^\infty \mathbb{E}_g[\mu(gA_i)] && (8)\\
&= \sum_{i=1}^\infty \mu^G(A_i),
\end{aligned}$$

where Equation (6) follows because of exchangeability of group actions and union, Equation (7) is due to σ-additivity of μ, and Equation (8) is derived by Fubini's theorem. We also note that $\mu^G(\mathcal{X}) = \mathbb{E}_g[\mu(g\mathcal{X})] = \mathbb{E}_g[\mu(\mathcal{X})] = 1$ since g is bijective, thereby μ^G is a probability measure. We conclude the proof by showing that μ^G is G-invariant,

$$\forall g_1 \in G \qquad \mu^G(g_1 A) = \mathbb{E}_g[\mu(g^{-1}g_1 A)] = \mathbb{E}_{g'}[\mu(g' A)] = \mu^G(A),$$

where we used the fact that the Haar measure on the group G is invariant with respect to any left action by $g_1 \in G$, and thus $g' = g^{-1}g_1$ is again distributed according to the Haar measure on the group G.

□

D.2. Proof of Theorem 8

Theorem 8 (Definiteness of KMIC)**.** *Let K be a universal PDS kernel defined on a complete metric space $\mathcal{X}$. Let $\mu \in \mathcal{P}(\mathcal{X})$ be a probability measure. Then, $\mathrm{KMIC}(\mu) = 0$ if and only if μ is G-invariant.*

Proof. We notice that by definition $\mathrm{KMIC}(\mu) = \|\mu_{\mathcal{H}}^G - \mu_{\mathcal{H}}\|_{\mathcal{H}} = \mathrm{MMD}(\mu, \mu^G)$. Hence, by Theorem 5 of Gretton et al. [46], $\mathrm{KMIC}(\mu) = 0$ if and only if $\mu = \mu^G$, thus μ is G-invariant. □

D.3. Proof of Proposition 6

Proposition 6. *Consider a shift-invariant PDS kernel K defined on the complete metric space $\mathcal{X}$. Then, KMIC can be alternatively represented as*

$$\mathrm{KMIC}(\mu) = \mathbb{E}_{x,x'}[K(x,x')] - \mathbb{E}_{g,x,x'}[K(x,gx')],$$

where $x, x' \sim \mu$ independently, and $g \in G$ is chosen uniformly at random and independently of x and x'.

Proof. First, we indicate that by definition of μ^G, $\mu_{\mathcal{H}}^G = \mathbb{E}_{x\sim\mu^G}[\phi(x)] = \mathbb{E}_g[\mu_{\mathcal{H}}(gx)] = \mathbb{E}_{x\sim\mu,g}[\phi(gx)]$. In turn,

$$\begin{aligned}
\mathrm{KMIC}(\mu) &= \|\mu_{\mathcal{H}} - \mu_{\mathcal{H}}^G\|_{\mathcal{H}}^2 \\
&= \langle \mu_{\mathcal{H}}, \mu_{\mathcal{H}}\rangle_{\mathcal{H}} + \langle \mu_{\mathcal{H}}^G, \mu_{\mathcal{H}}^G\rangle_{\mathcal{H}} - 2\langle \mu_{\mathcal{H}}^G, \mu_{\mathcal{H}}\rangle_{\mathcal{H}} \\
&= \mathbb{E}_{x\sim\mu}[\mu_{\mathcal{H}}(x)] + \mathbb{E}_{x\sim\mu^G}[\mu_{\mathcal{H}^G}(x)] - 2\mathbb{E}_{x\sim\mu}[\mu_{\mathcal{H}^G}(x)] \\
&= \langle \mu_{\mathcal{H}}(x), \mathbb{E}_{x\sim\mu}[\phi(x)]\rangle + \langle \mu_{\mathcal{H}}^G(x), \mathbb{E}_{x\sim\mu^G}[\phi(x)]\rangle - 2\langle \mu_{\mathcal{H}}^G(x), \mathbb{E}_{x\sim\mu}[\phi(x)]\rangle \\
&= \langle \mu_{\mathcal{H}}(x), \mathbb{E}_{x\sim\mu}[\phi(x)]\rangle + \langle \mathbb{E}_g[\mu_{\mathcal{H}}(gx)], \mathbb{E}_{x\sim\mu^G}[\phi(x)]\rangle - 2\langle \mathbb{E}_g[\mu_{\mathcal{H}}(gx)], \mathbb{E}_{x\sim\mu}[\phi(x)]\rangle \\
&= \langle \mathbb{E}_{x\sim\mu}[\phi(x)], \mathbb{E}_{x\sim\mu}[\phi(x)]\rangle + \langle \mathbb{E}_{x\sim\mu,g}[\phi(gx)], \mathbb{E}_{x\sim\mu,g}[\phi(gx)]\rangle - 2\langle \mathbb{E}_{x\sim\mu,g}[\phi(gx)], \mathbb{E}_{x\sim\mu}[\phi(x)]\rangle \\
&= \mathbb{E}_{x,x'\sim\mu}[K(x,x')] + \mathbb{E}_{x,x'\sim\mu,g,g'}[K(gx,g'x')] - 2\mathbb{E}_{x,x'\sim\mu,g}[K(gx,x')] \\
&= \mathbb{E}_{x,x'\sim\mu}[K(x,x')] + \mathbb{E}_{x,x'\sim\mu,g,g'}[K(g^{-1}gx,g^{-1}g'x')] - 2\mathbb{E}_{x,x'\sim\mu,g}[K(gx,x')] \qquad (9) \\
&= \mathbb{E}_{x,x'\sim\mu}[K(x,x')] + \mathbb{E}_{x,x'\sim\mu,g''}[K(x,g''x')] - 2\mathbb{E}_{x,x'\sim\mu,g}[K(gx,x')] \qquad (10) \\
&= \mathbb{E}_{x,x'\sim\mu}[K(x,x')] - \mathbb{E}_{x,x'\sim\mu,g}[K(x,gx')],
\end{aligned}$$

where Equation (9) follows by shift-invariance property of the kernel and Equation (10) by properties of Haar measures. □

Remark D.1. In the proof of Proposition 6, we leveraged the shift invariance property in Equation (9). However, for general kernels, it is immediate to show that similarly,

$$\mathrm{KMIC}(\mu) = \mathbb{E}_{x,x'}[K(x,x')] + \mathbb{E}_{x,x',g,g'}[K(gx,g'x')] - 2\mathbb{E}_{x,x',g}[K(x,gx')].$$

D.4. Proof of Theorem 9

Theorem 9 (Convergence rate for $\widehat{\mathrm{KMIC}}(\mu)$)**.** *For the estimator* $\widehat{\mathrm{KMIC}}(\mu)$ *defined in Algorithm 2, we have*

$$\mathbb{E}\left[\left|\widehat{\mathrm{KMIC}}(\mu) - \mathrm{KMIC}(\mu)\right|^2\right] \lesssim \frac{c_1^2}{n}, \tag{5}$$

where $c_1 := \sup_{x \in \mathcal{X}} K(x, x)$.

Proof. Define $T_1 = \frac{2}{n(n-1)} \sum_{\substack{i,j=1 \\ i<j}}^n K(x_i, x_j)$ and $T_2 = \frac{2}{n(n-1)} \sum_{\substack{i,j=1 \\ i<j}}^n K(x_i, g_j x_j)$. Note that

$$\mathbb{E}\left[\left|\widehat{\mathrm{KMIC}}(\mu) - \mathrm{KMIC}(\mu)\right|^2\right] \leq 2\mathbb{E}\big[|T_1 - \mathbb{E}[T_1]|^2\big] + 2\mathbb{E}\big[|T_2 - \mathbb{E}[T_2]|^2\big].$$

Let us focus on the first term. Define $a_{ij} = K(x_i, x_j) = \mathbb{E}[K(x_i, x_j)]$. Note that

$$\mathbb{E}\big[|T_1 - \mathbb{E}[T_1]|^2\big] = \mathbb{E}\left[\left|\frac{2}{n(n-1)} \sum_{\substack{i,j=1 \\ i<j}}^n a_{ij}\right|^2\right] = \frac{4}{n^2(n-1)^2} \sum_{\substack{i,j=1 \\ i<j}}^n \sum_{\substack{k,\ell=1 \\ k<\ell}}^n \mathbb{E}[a_{ij} a_{k\ell}].$$

However, note that if $i \neq k$ and $j \neq \ell$, then $\mathbb{E}[a_{ij} a_{k\ell}] = 0$. Therefore, there exists at most $O(n^3)$ non-zero elements in the above sum, and each is at most $\mathbb{E}[a_{ij} a_{k\ell}] \leq c_1^2$. Therefore, $\mathbb{E}\big[|T_1 - \mathbb{E}[T_1]|^2\big] \lesssim n^3 \frac{c_1^2}{n^4} = \frac{c_1^2}{n}$. Similarly, one can conclude that $\mathbb{E}\big[|T_2 - \mathbb{E}[T_2]|^2\big] \lesssim \frac{c_1^2}{n}$, and this completes the proof.

∎

D.5. Proof of Theorem 10

Theorem 10. *For the estimator* $\widehat{\mathrm{KMIC}}(\mu)$ *defined in Algorithm 2, we have*

$$\mathbb{P}\left(\left|\widehat{\mathrm{KMIC}}(\mu) - \mathrm{KMIC}(\mu)\right| \geq t\right) \leq 4 \exp\left(-\frac{nt^2}{32 c_1^2}\right),$$

where $c_1 := \sup_{x \in \mathcal{X}} K(x, x)$.

Proof. Similar to the proof of Theorem 9, let us define $T_1 = \frac{2}{n(n-1)} \sum_{\substack{i,j=1 \\ i<j}}^n K(x_i, x_j)$ and $T_2 = \frac{2}{n(n-1)} \sum_{\substack{i,j=1 \\ i<j}}^n K(x_i, g_j x_j)$. Note that

$$\begin{aligned}
\mathbb{P}\Big(\widehat{\mathrm{KMIC}}(\mu) - \mathrm{KMIC}(\mu) \geq t\Big) &= \mathbb{P}(T_1 + T_2 \geq t) \\
&= \mathbb{P}\big(T_1 + T_2 \geq t \big| T_1 > t/2\big)\mathbb{P}(T_1 > t/2) \\
&\qquad + \mathbb{P}\big(T_1 + T_2 \geq t \big| T_1 \leq t/2\big)\mathbb{P}(T_1 \leq t/2) \\
&\leq \mathbb{P}(T_1 > t/2) + \mathbb{P}\big(T_1 + T_2 \geq t \big| T_1 \leq t/2\big) \\
&\leq \mathbb{P}(T_1 > t/2) + \mathbb{P}(T_2 > t/2).
\end{aligned}$$

Therefore, we conclude that

$$\mathbb{P}\left(\left|\widehat{\mathrm{KMIC}}(\mu) - \mathrm{KMIC}(\mu)\right| \geq t\right) \leq \mathbb{P}(|T_1| > t/2) + \mathbb{P}(|T_2| > t/2).$$

Using standard tail bounds on U-statistics [62, Example 2.23], we know that $\mathbb{P}(|T_1| > t/2) \leq 2 \exp\left(-\frac{nt^2}{4c_1^2}\right)$. A similar upper bound also holds for T_2. The proof is thus complete. ∎

D.6. Proof of Theorem 4

Theorem 4 (Definiteness of KMaxIC)**.** *For any probability measure $\mu \in \mathcal{P}(\mathcal{X})$, we have* $\mathrm{KMaxIC}(\mu) = 0$ *if and only if μ is G-invariant, assuming the kernel is universal.*

Proof. If the measure μ is G-invariant, then for all $g \in G$, $g\mu = \mu$ almost everywhere and hence, $\|(g\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\|_{\mathcal{H}}^2 = 0$ and thereby, $\mathrm{KMaxIC}(\mu) = 0$. Next, assume that $\mathrm{KMaxIC}(\mu) = 0$, then $\|(g\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\|_{\mathcal{H}}^2 = 0$ for all $g \in G$. Thus, by Theorem 5 of Gretton et al. [46], $g\mu = \mu$ and μ is G-invariant. ☐

D.7. Proof of Theorem 5

Theorem 5 (Definiteness of KMaxIC via generating sets)**.** *Assuming the underlying kernel used to define KMaxIC is universal, for any arbitrary generating set $S \subseteq G$ and any probability measure $\mu \in \mathcal{P}(\mathcal{X})$, if*

$$\max_{g \in S} \left\|(g\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\right\|_{\mathcal{H}}^2 = 0,$$

then $\mathrm{KMaxIC}(\mu) = 0$*, which implies that μ is G-invariant.*

Proof. Let the generating set $S = \{g_1, g_2, \dots, g_{|S|}\}$ and let g' be a maxima element that attains KMaxIC

$$g' = \arg\max_{g \in G} \|(g\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\|_{\mathcal{H}}^2.$$

By definition of generating set, there exist a finite set $\{g_i'\}_{i=1}^{\ell}$, where its potential repetitive member $g_i' \in S$ and $g' = \prod_{i=1}^{\ell} g_i'$. Thus,

$$\begin{aligned}
\sqrt{\mathrm{KMaxIC}(\mu)} &= \|(g'\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\|_{\mathcal{H}} \\
&= \|(\prod_{i=1}^{\ell} g_i'\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\|_{\mathcal{H}} \\
&\leq \|(\prod_{i=1}^{\ell} g_i'\mu)_{\mathcal{H}} - (\prod_{i=1}^{\ell-1} g_i'\mu)_{\mathcal{H}}\|_{\mathcal{H}} + \|(\prod_{i=1}^{\ell-1} g_i'\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\|_{\mathcal{H}},
\end{aligned}$$

where we used triangle equality over the $\|.\|_{\mathcal{H}}$ norm. By iterative applications of triangle equality,

$$\sqrt{\mathrm{KMaxIC}(\mu)} \leq \sum_{l=1}^{\ell} \|(\prod_{i=0}^{l} g_i'\mu)_{\mathcal{H}} - (\prod_{i=0}^{l-1} g_i'\mu)_{\mathcal{H}}\|_{\mathcal{H}},$$

where we overload the notation by setting $g_0' = e$ the indentity member in the group G. Now, by induction we prove that for all $l \in [\ell]$, the term $\|(\prod_{i=0}^{l} g_i'\mu)_{\mathcal{H}} - (\prod_{i=0}^{l-1} g_i'\mu)_{\mathcal{H}}\|_{\mathcal{H}} = 0$. We know that $\|(g_1'\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\|_{\mathcal{H}} = 0$, thus $\mu = e\mu = g_1'\mu$ almost surely. Now, assume that by induction assumption for l,

$$\|(\prod_{i=0}^{l} g_i'\mu)_{\mathcal{H}} - (\prod_{i=0}^{l-1} g_i'\mu)_{\mathcal{H}}\|_{\mathcal{H}} = 0.$$

Thus, $\prod_{i=0}^{l} g_i'\mu = \prod_{i=0}^{l-1} g_i'\mu = \cdots = g_1'\mu = \mu$ almost everywhere. Hence,

$$\|(\prod_{i=0}^{l+1} g_i'\mu)_{\mathcal{H}} - (\prod_{i=0}^{l} g_i'\mu)_{\mathcal{H}}\|_{\mathcal{H}} = \|(g_{l+1}'\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\|_{\mathcal{H}} = 0,$$

and the induction hypothesis is proved. Putting all pieces together,

$$\sqrt{\mathrm{KMaxIC}(\mu)} \leq \sum_{l=1}^{\ell} \|(\prod_{i=0}^{l} g_i'\mu)_{\mathcal{H}} - (\prod_{i=0}^{l-1} g_i'\mu)_{\mathcal{H}}\|_{\mathcal{H}} = 0.$$

Therefore, $\mathrm{KMaxIC}(\mu) = 0$ and the proof is concluded. ☐

D.8. Proof of Proposition 1

Proposition 1 (Size of generating sets). *Any finite group G has a generating set $S \subseteq G$ of size at most* $\log_2(|G|)$.

Proof. Let $S = \{g_1, g_2, \ldots, g_{|S|}\}$ be a minimal generating set for the finite group G. For each $k \in \{1, 2, \ldots, |S|\}$, define the subgroup $G_k = \langle g_1, g_2, \ldots, g_k \rangle$, which is generated by the first k elements of S.

For each $k \in \{1, 2, \ldots, |S|\}$, the element g_{k+1} must lie outside G_k. If this were not the case, then the group G could be generated by the set $\{g_1, g_2, \ldots, g_k, g_{k+2}, \ldots, g_{|S|}\}$, which contradicts the assumption that S is minimal.

As a result, the coset $g_{n+1}G_n$ is disjoint from G_n. By construction, we know that $g_{n+1}G_n \cup G_n \subseteq G_{n+1}$, which implies $|G_{n+1}| \geq |g_{n+1}G_n| + |G_n| = 2|G_n|$. By applying, it follows that $|G| = |G_{|S|}| \geq 2^{|S|}|G_1|$, and since $|G_1| = 1$, we conclude that $|G| \geq 2^{|S|}$, thus proving the result.

□

D.9. Proof of Proposition 2

Proposition 2. *For any $g \in G$ and any probability measure $\mu \in \mathcal{P}(\mathcal{X})$, we have*

$$\left\|(g\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\right\|_{\mathcal{H}}^2 = 2\mathbb{E}_{x,x'}[K(x, x')] - 2\mathbb{E}_{x,x'}[K(x, gx')],$$

where $x, x' \sim \mu$ are independent random variables.

Proof. Note that $(g\mu)_{\mathcal{H}} = \mathbb{E}_{x\sim g\mu}[\phi(x)] = \mathbb{E}_{x\sim\mu}[\phi(gx)]$ where $\phi(x) = K(., x)$ for each $x \in \mathcal{X}$. Therefore,

$$\begin{aligned}
\left\|(g\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\right\|_{\mathcal{H}}^2 &= \langle \mu_{\mathcal{H}}, \mu_{\mathcal{H}} \rangle_{\mathcal{H}} + \langle (g\mu)_{\mathcal{H}}, (g\mu)_{\mathcal{H}} \rangle_{\mathcal{H}} - 2\langle \mu_{\mathcal{H}}, (g\mu)_{\mathcal{H}} \rangle_{\mathcal{H}} \\
&= \mathbb{E}_{x\sim\mu}[\mu_{\mathcal{H}}(x)] + \mathbb{E}_{x\sim g\mu}[(g\mu)_{\mathcal{H}}] - 2\mathbb{E}_{x\sim\mu}[(g\mu)_{\mathcal{H}}] \\
&= \mathbb{E}_{x,x'\sim\mu}[K(x, x')] + \mathbb{E}_{x,x'\sim\mu}[K(gx, gx')] - 2\mathbb{E}_{x,x'\sim\mu}[K(x, gx')] \\
&= \mathbb{E}_{x,x'\sim\mu}[K(x, x')] + \mathbb{E}_{x,x'\sim\mu}[K(x, x')] - 2\mathbb{E}_{x,x'\sim\mu}[K(x, gx')] \qquad (11) \\
&= 2\mathbb{E}_{x,x'\sim\mu}[K(x, x')] - 2\mathbb{E}_{x,x'\sim\mu}[K(x, gx')],
\end{aligned}$$

where $x, x' \sim \mu$ are independent and in Equation (11), we used the shift-invariance property of the kernel. The proof is thus complete.

□

D.10. Proof of Proposition 3

Proposition 3. *The set $S^\star \subseteq P_d$ defined via Equation (3) is a generating set for P_d and satisfies*

$$\ell(S^\star) \leq \frac{d(d-1)}{2}.$$

Proof. Let $\sigma \in P_d$ be an arbitrary permutation. We need to show that there exists a sequence $i_1, i_2, \ldots, i_k \in [d]$ of length $k \leq \frac{d(d-1)}{2}$ such that $\sigma = \sigma_{i_1} \circ \sigma_{i_2} \circ \ldots \circ \sigma_{i_k}$. We prove this by induction on d.

First note that the case $d = 2$ is trivial. Fix $d > 2$ and let $\sigma \in P_d$ be an arbitrary permutation. Assume that $\sigma(d) = \ell$, for some $\ell \in [d]$. Consider the following permutation: $\tilde{\sigma} = \sigma_\ell \circ \sigma_{\ell+1} \circ \ldots \sigma_{d-1} \in P_d$. Note that $\tilde{\sigma}(d) = \ell$. Let $\sigma' = \widetilde{\sigma}^{-1} \circ \sigma \in P_d$. Note that $\sigma'(d) = \widetilde{\sigma}^{-1}(\ell) = d$. This means that $\sigma' \in P_d$ can be considered as a permutation of $[d-1]$. Using induction hypothesis, one can represent σ' as composition of at most $\frac{(d-1)(d-2)}{2}$ transpositions. Moreover, since $\sigma = \widetilde{\sigma} \circ \sigma'$, one can represent σ as compositions of at most

$$\frac{(d-1)(d-2)}{2} + (d-1) = \frac{d(d-1)}{2}$$

transpositions, and this completes the proof.

□

D.11. Proof of Proposition 4

Proposition 4. *The set $S^\star \subseteq G$ defined via Equation (4) is a generating set for G and satisfies*

$$\ell(S^\star) \leq \log_2(m).$$

Proof. Let $t \in G = \mathbb{Z}/m\mathbb{Z}$ be an arbitrary group element. Our goal is to find $t_i \in S^\star$, for $i \in [k]$ with $k \leq \log_2(m)$, such that $t = \sum_{i=1}^k t_i$. Note that the elements of $S^\star$ are of the form 2^ℓ for some ℓ. Now, consider the binary representation of t: $t = \sum_{i=1}^k a_i 2^i$, where $a_i \in \{0,1\}$ and $k \leq \log_2(m)$ since $t \in [m]$. This representation provides the necessary decomposition of $t \in [m]$, thus completing the proof.

□

E. Proof of Main results

E.1. Proof of Theorem 2

Theorem 2 (Computational intractability)**.** *There exists a shift-invariant pseudometric $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$, a finite group G, and a discrete probability measure μ such that solving the optimization problem $\arg\sup_{g \in G} D(\mu, g\mu)$ is NP-complete.*

Proof. We demonstrate the computational hardness result by reducing it to a special variant of the metric traveling salesperson problem (Metric TSP), which we refer to as the reward maximization metric traveling salesperson problem (Reward Metric TSP). In this variant, instead of finding the minimum tour that starts and ends at the same node, the objective is to find the maximum (most profitable) tour. In Reward Metric TSP, the edges between nodes are characterized by a reward function, rather than distances, that satisfies the metric property. In Proposition E.1, we show that this special variant is also NP-complete.

Given a complete graph $\mathcal{G}$ with d nodes denoted by the set V, for all nodes $u, v \in V$, we denote the positive reward function between them by $\rho(u, v)$. By the definition, $\rho(.,.)$ is a metric and it satisfies triangle inequality. We want to find the maximum tour C_{max}, i.e., $C_{\text{max}} = \arg\max_{C \text{ is a tour}} \sum_{(u,v) \in C} \rho(u, v)$. We scale the reward function ρ to design the new reward function $d(.,.)$ by $d(u, v) \coloneqq \rho(u, v)/M + 1$, where M is an upper bound on the reward function ρ, i.e, $M = \sup_{u,v \in V} \rho(u, v)$. By definition, $1 \leq d(u, v) \leq 2$ and clearly $d(.,.)$ is also a metric. Additionally, $C_{\text{max}} = \arg\max_{C \text{ is a tour}} \sum_{(u,v) \in C} d(c, v)$.

Number the nodes of the graph $\mathcal{G}$ arbitrarily from 1 to d and call this numbering $e : [d] \to [d]$, $e(i) = i$ for all $i \in [d]$. For any other numbering $g : [d] \to [d]$ of the nodes of $\mathcal{G}$, let $\mathcal{G}_g$ denote the resulting renumbered copy of $\mathcal{G}$. The set of all such numberings corresponds to the permutation group P_d. By definition, for the identity element $e \in P_d$, we have $\mathcal{G}_e = \mathcal{G}$. Next, we choose the group action set $G = P_d$ and define the set $\mathcal{X} \coloneqq \{\mathcal{G}_g \mid g \in P_d\}$, so that $\mathcal{G} \in \mathcal{X}$. Let μ be the Dirac delta measure on the element $\mathcal{G}$, i.e., $\mu = \delta_{\mathcal{G}}$.

In sequel, we define the pseudometric D for any $g \in G$,

$$
\begin{aligned}
D(\mu, g\mu) &\coloneqq \sum_{i=1}^{d} d\Big(e^{-1}g(i), e^{-1}g(i+1 \bmod d)\Big) + \sum_{i=1}^{d} d\Big(g^{-1}e(i), g^{-1}e(i+1 \bmod d)\Big) \\
&= \sum_{i=1}^{d} d\Big(g(i), g(i+1 \bmod d)\Big) + \sum_{i=1}^{d} d\Big(g^{-1}(i), g^{-1}(i+1 \bmod d)\Big) \\
&= 2\sum_{i=1}^{d} d\Big(g(i), g(i+1 \bmod d)\Big), \qquad (12)
\end{aligned}
$$

where Equation (12) follows by the double counting argument on the direction of calculating the value of the resulting tour and the symmetry property of reward function $d(.,.)$. Intuitively, $\frac{1}{2}D(\mu, g\mu)$ is calculating reward of the tour resulted by traversing the graph $\mathcal{G}$ according to the numbering g (or equivalently permutation g of the nodes).

Similarly, for any $g, g' \in G$ we define

$$
D(g'\mu, g\mu) \coloneqq \sum_{i=1}^{d} d\Big({g'}^{-1}g(i), {g'}^{-1}g(i+1 \bmod d)\Big) + \sum_{i=1}^{d} d\Big(g^{-1}g'(i), g^{-1}g'(i+1 \bmod d)\Big),
$$

where we are overloading g, g' and e by using them as the elements of the permutation group and also the mapping induced by the corresponding permutations. Now, we need to show that $D(.,.)$ is indeed a shift invariant pseudometric. We start by showing that $D(.,.)$ is symmetric.

$$
\begin{aligned}
D(g'\mu, g\mu) &= \sum_{i=1}^{d} d\Big({g'}^{-1}g(i), {g'}^{-1}g(i+1 \bmod d)\Big) + \sum_{i=1}^{d} d\Big(g^{-1}g'(i), g^{-1}g'(i+1 \bmod d)\Big) \\
&= \sum_{i=1}^{d} d\Big(g^{-1}g'(i), g^{-1}g'(i+1 \bmod d)\Big) + \sum_{i=1}^{d} d\Big({g'}^{-1}g(i), {g'}^{-1}g(i+1 \bmod d)\Big) \\
&= D(g\mu, g'\mu).
\end{aligned}
$$

Next, we show that $D(.,.)$ is shift-invariant,

$$
\begin{aligned}
D(g''g'\mu, g''g\mu) &= \sum_{i=1}^{d} d\Big((g''g')^{-1}g''g(i), (g''g')^{-1}g''g(i+1 \bmod d)\Big) \\
&\qquad + \sum_{i=1}^{d} d\Big((g''g)^{-1}g''g'(i), (g''g)^{-1}g''g'(i+1 \bmod d)\Big) \\
&= \sum_{i=1}^{d} d\Big({g'}^{-1}{g''}^{-1}g''g(i), {g'}^{-1}{g''}^{-1}g''g(i+1 \bmod d)\Big) \\
&\qquad + \sum_{i=1}^{d} d\Big(g^{-1}{g''}^{-1}g''g'(i), g^{-1}{g''}^{-1}g''g'(i+1 \bmod d)\Big) \\
&= \sum_{i=1}^{d} d\Big({g'}^{-1}g(i), {g'}^{-1}g(i+1 \bmod d)\Big) + \sum_{i=1}^{d} d\Big(g^{-1}g'(i), g^{-1}g'(i+1 \bmod d)\Big)\Big) \\
&= D(g'\mu, g\mu).
\end{aligned}
$$

In the end, we prove the Triangle inequality for $D(.,.)$. In order to do so, we recall that $\frac{1}{2}D(g''\mu, g\mu)$ is the length of a tour C in the graph $\mathcal{G}$ endowed with metric d. Additionally, we designed the metric $d(.,.)$ such that $1 \leq d(u,v) \leq 2$.. Hence, $\sum_{(u,v)\in C_{\max}} d(u,v) \leq 2|V|$ and $|V| \leq \sum_{(u,v)\in C_{\max}} d(u,v)$

Therefore, by terminology and multiple usage of this fact,

$$\begin{aligned}D(g''\mu, g\mu) &= 2 \sum_{(u,v)\in C} d(u,v) \\ &\leq 2 \sum_{(u,v)\in C_{\max}} d(u,v) \\ &\leq 4|V| \\ &\leq 4 \sum_{(u,v)\in C_{\min}} d(u,v) \\ &\leq D(g''\mu, g'\mu) + D(g'\mu, g\mu),\end{aligned}$$

where in the last line, we again exploited the fact that $\frac{1}{2}D(g''\mu, g'\mu)$ and $\frac{1}{2}D(g''\mu, g'\mu)$ are the length of arbitrary tours in $\mathcal{G}$, therefore their length is more than $\sum_{(u,v)\in C_{\min}} d(u,v)$. Putting all of these pieces together, we showed that $D(.,.)$ is a shift invariant pseudometric.

To conclude the proof, given an instance of the Reward Metric TSP $\mathcal{G}$ equipped with a metric $\rho(.,.)$, we form the scaled metric $d(.,.)$, the shift invariant pseudometric $D(.,.)$, finite group G, the set $\mathcal{X}$, and the distribution $\mu = \delta_{\mathcal{G}}$ as above. By our construction, solution to the optimization problem,

$$\sup_{g\in G} D(\mu, g\mu)$$

is the maximum tour for the problem of Reward Metric TSP. Therefore, this optimization problem for a specific choice of parameters is NP-complete. □

Proposition E.1 (Hardness Result for Reward Metric TSP)**.** *Given a complete graph $\mathcal{G}$, equipped with a metric d, finding the maximum tour of the graph is NP-complete.*

Proof. We prove this result by reduction to the problem of finding a Hamiltonian cycle problem. Formally speaking, given a complete weighted graph $\mathcal{G} = (V, E)$, the question is whether this graph has a Hamiltonian cycle or not. In order to build the reduction, we create a complete weighted graph $\mathcal{G}' = (V, E')$, with the exact set of nodes as $\mathcal{G}$ but we assign weights $d(.,.)$ as follows.

- If an edge $(u, v) \in E$ exists in the original graph $\mathcal{G}$, we assign the weight $d(u, v) = 2$.
- If an edge $(u, v) \notin E$ doesn't exist in the original graph $\mathcal{G}$, we assign the weight $d(u, v) = 1$.

All the edges are positive and they satisfy the triangle inequality trivially. This $\mathcal{G}'$ is a metric graph. If the original graph $\mathcal{G}$ has a Hamiltonian cycle, then the Maximum Tour of the metric graph $\mathcal{G}'$ has the size $2|V|$, otherwise the Maximum Tour of the metric graph $\mathcal{G}'$ has a size strictly lower than $2|V|$. Therefore, the reduction is complete and since the problem of checking existence of a Hamiltonian cycle is NP-complete, the Reward Metric TSP is also NP-complete. □

E.2. Proof of Theorem 3

Theorem 3 (Probabilistic approximation (formal version of Theorem 1))**.** *Let $\mathcal{X}$ be a complete metric space and $\mathcal{P}(\mathcal{X})$ denote the space of (Borel) probability measures on $\mathcal{X}$. Let G be a compact topological group acting continuously on $\mathcal{X}$. Consider a shift-invariant probability metric $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$. Then,*

$$\mathbb{E}_g[D(\mu, g\mu)] \leq \sup_{g\in G} D(\mu, g\mu) \leq 4\mathbb{E}_g[D(\mu, g\mu)],$$

where expectation is taken w.r.t. left Haar (uniform) measure over the group G.

Proof. First, note that

$$\mathbb{E}_g[D(\mu, g\mu)] \leq \sup_{g\in G} D(\mu, g\mu),$$

for each $g \in G$. Therefore, we focus of the proof of the other inequality. Fix a probability measure $\mu \in \mathcal{P}(\mathcal{X})$. Let

$$g^\star := \arg\max_{g \in G} D(\mu, g\mu).$$

Note that such $g^\star \in G$ exists according to the compactness of G. Define the following function

$$\Delta(g) := D(\mu, g\mu), \quad \forall g \in G.$$

Note that for any $g_1, g_2 \in G$, we have

$$\begin{aligned}
\Delta(g_1 g_2) &= D(\mu, (g_1 g_2)\mu) \\
&\leq D(\mu, g_1\mu) + D(g_1\mu, (g_1 g_2)\mu) \\
&= \Delta(g_1) + D(g_1\mu, (g_1 g_2)\mu),
\end{aligned}$$

using the triangle inequality for the metric D. Now, note that D is shift-invariant, meaning that we have

$$D(g_1\mu, (g_1 g_2)\mu) = D(\mu, g_2\mu) = \Delta(g_2).$$

Therefore, we conclude that

$$\Delta(g_1 g_2) \leq \Delta(g_1) + \Delta(g_2), \quad \forall g_1, g_2 \in G.$$

In other words, the function Δ is sub-linear. Now specify the above inequality to $g_1 = g^{-1}$ and $g_2 = gg^\star$ for an arbitrary $g \in G$ to get

$$\begin{aligned}
\Delta(g^\star) &\leq \Delta(g^{-1}) + \Delta(gg^\star) \\
&= \Delta(g) + \Delta(gg^\star), \quad \forall g \in G.
\end{aligned} \tag{13}$$

In above, we used $\Delta(g) = \Delta(g^{-1})$ which holds from the shift-invariance of D. Now define the following set:

$$A := \left\{ g \in G : \Delta(g) \geq \frac{1}{2}\Delta(g^\star) \right\} \subseteq G.$$

Define

$$g^\star A := \left\{ g^\star g \in G : g \in A \right\}.$$

Note that according to Equation (13), for each $g \in G$, either $g \in A$ or $g \in g^\star A$. In other words, $G = A \cup g^\star A$. Let α denote the left Haar measure on G. Then, we conclude that

$$\alpha(A) + \alpha(g^\star A) \geq \alpha(A \cup g^\star A) = \alpha(G) = 1.$$

However, $\alpha(A) = \alpha(g^\star A)$ since α is the Haar measure. This means that $\alpha(A) \geq \dfrac{1}{2}$. Therefore, we conclude

$$\begin{aligned}
\mathbb{E}_{g \sim \alpha}[D(\mu, g\mu)] &= \mathbb{E}_{g \sim \alpha}[\Delta(g)] \\
&\geq \alpha(A) \frac{\Delta(g^\star)}{2} \\
&\geq \frac{\Delta(g^\star)}{4} \\
&= \frac{1}{4} \sup_{g \in G} D(\mu, g\mu),
\end{aligned}$$

which completes the proof.

Remark E.2. The proof we presented here works for pseudometric, i.e., even if D is not *definite*, as we only used the triangle inequality and the symmetry of D.

$\square$

E.3. Proof of Theorem 6

Theorem 6. *Consider Algorithm 1 ran on n samples from a G-invariant probability measure μ. Then, the probability of a Type I error (i.e., incorrectly rejecting the invariance) is bounded as*

$$\mathbb{P}\Big(\mathrm{H}_{\mathbf{1}}|\mathrm{H}_{\mathbf{0}}\Big) = \mathbb{P}\Big(\max_{g\in S}\widehat{c}_g > c \;\Big|\; \mu \textit{ is } G\text{-invariant}\Big) \\ \leq |S|\exp\left(-\frac{nc^2}{128c_1^2}\right),$$

where $c_1 := \sup_{x\in\mathcal{X}} K(x,x)$. Moreover, the Type II error, which is the probability of incorrectly accepting a non-invariant measure using Algorithm 1, approaches zero as the sample size increases. Quantitatively, for any probability measure $\mu \in \mathcal{P}(\mathcal{X})$ such that $\mathrm{KMaxIC}(\mu) \geq 2c' > c\ell(S)^2$, we have

$$\mathbb{P}\Big(\mathrm{H}_{\mathbf{0}}|\mathrm{H}_{\mathbf{1}}\Big) = \mathbb{P}\Big(\max_{g\in S}\widehat{c}_g \leq c \;\Big|\; \mu \text{ is not } G\text{-invariant}\Big) \\ \leq \exp\Big(-\frac{n\big(\frac{2c'}{\ell(S)^2}-c\big)^2}{128c_1^2}\Big).$$

Proof. First, we focus on the first inequality. Note that using the union bound

$$\mathbb{P}\Big(\mathrm{H}_{\mathbf{0}}|\mathrm{H}_{\mathbf{1}}\Big) \leq |S|\max_{g\in S}\mathbb{P}\Big(\widehat{c}_g > c \;\Big|\; \mu \text{ is } G\text{-invariant}\Big).$$

Fix a group element $g \in G$. Let $a_{ij} = 2K(x_i,x_j) - 2K(x_i,gx_j)$ for each $i,j \in [n]$. Note that $\widehat{c}_g = \frac{2}{n(n-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{n} a_{ij}$. Assuming that μ is G-invariant, one has $\mathbb{E}[a_{ij}] = 0$ for any $i \neq j$. Therefore

$$\mathbb{P}\Big(\widehat{c}_g > c \;\Big|\; \mu \text{ is } G\text{-invariant}\Big) = \mathbb{P}\Big(\frac{2}{n(n-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{n} a_{ij} > c\Big).$$

Using standard tail bounds on U-statistics [62, Example 2.23], we know that the right-hand side of the above is upper bounded by $\exp\Big(-\frac{nc^2}{128c_1^2}\Big)$, since $|a_{ij}| \leq 4c_1$. This completes the proof of the first inequality.

Now, we prove the second inequality. Assume that $\mathrm{KMaxIC}(\mu) \geq 2c'$. Define $c_g = \mathbb{E}[\hat{c}_g]$ for each $g \in G$. Let $g^\star \in \arg\max_{g\in G}\|(g\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\|_{\mathcal{H}}^2$. According to the assumption, there exists a sequence $g_i \in S$, $i \in [k]$, with $k \leq \ell(S)$, such that $g^\star = g_1g_2\dots g_k$. Then, we have

$$\begin{aligned}\sqrt{\mathrm{KMaxIC}(\mu)} &= \|(g^\star\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\|_{\mathcal{H}} \\ &= \|(g_1g_2\dots g_k\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\|_{\mathcal{H}} \\ &\leq \|(g_1\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\|_{\mathcal{H}} + \|(g_1g_2\dots g_k\mu)_{\mathcal{H}} - (g_1\mu)_{\mathcal{H}}\|_{\mathcal{H}} \\ &= \sqrt{c_{g_1}} + \|(g_2\dots g_k\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\|_{\mathcal{H}},\end{aligned}$$

where the last step follows from the shift-invariance of the chosen kernel. Therefore, by induction, we conclude that

$$\mathrm{KMaxIC}(\mu) = \|(g^\star\mu)_{\mathcal{H}} - \mu_{\mathcal{H}}\|_{\mathcal{H}}^2 \leq \ell(S)^2\ \max_{g\in S} c_g.$$

By assumption, $\mathrm{KMaxIC}(\mu) \geq 2c'$, which means that there exists $\widehat{g} \in S$ such that $c_{\hat{g}} \geq 2c'/\ell(S)^2$. Thus, by specifying to $\hat{g} \in S$ we have

$$\begin{aligned}\mathbb{P}\Big(\mathrm{H}_{\mathbf{0}}|\mathrm{H}_{\mathbf{1}}\Big) &= \mathbb{P}\Big(\max_{g\in S}\widehat{c}_g \leq c \;\Big|\; \mu \text{ is not } G\text{-invariant}\Big) \\ &\leq \mathbb{P}\Big(\widehat{c}_{\hat{g}} \leq c \;\Big|\; \mu \text{ is not } G\text{-invariant}\Big) \\ &= \mathbb{P}\Big(\frac{2}{n(n-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{n} a_{ij} \leq c\Big),\end{aligned}$$

where $\mathbb{E}[a_{ij}] = c_{\hat{g}} \geq 2c'/\ell(S)^2$. Thus, similar to the proof of the previous part and using standard tail bound on U-statistics, we conclude the desired result.

$\square$

E.4. Proof of Theorem 7

Theorem 7. *Consider a PDS kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is a closed subset, and let $G \subseteq O(d)$ be an orthogonal subgroup acting on $\mathcal{X}$. Assume that $K(x, \cdot) : \mathcal{X} \to \mathbb{R}$ is an r-Lipschitz function with respect to the norm $\|\cdot\|_2$ on $\mathbb{R}^d$, for each $x \in \mathcal{X}$. Let $S \subseteq G$ be a generating set for G with $\ell(G) < \infty$, and let $\widehat{S}$ be a γ-covering of S.*

Then, when applying Algorithm 1 via $\widehat{S}$ to test invariance to G, the probability of a Type I error (i.e., incorrectly rejecting the invariance) is bounded as

$$\begin{aligned}\mathbb{P}\Big(\mathrm{H}_{\mathbf{1}}|\mathrm{H}_{\mathbf{0}}\Big) &= \mathbb{P}\Big(\max_{g\in\widehat{S}} \widehat{c}_g > c \,\Big|\, \mu \textit{ is } G\text{-invariant}\Big) \\ &\leq |\widehat{S}| \exp\left(-\frac{nc^2}{128c_1^2}\right),\end{aligned}$$

where $c_1 := \sup_{x\in\mathcal{X}} K(x,x)$. Moreover, the Type II error, which is the probability of incorrectly accepting a non-invariant measure using Algorithm 1, approaches zero as the sample size increases. Specifically, for any probability measure $\mu \in \mathcal{P}(\mathcal{X})$ with $\mathbb{E}_{x\sim\mu}[\|x\|_2] \leq b$ such that $\mathrm{KMaxIC}(\mu) \geq 3c' > c\ell(S)^2 + 2rb\gamma$, we have

$$\begin{aligned}\mathbb{P}\Big(\mathrm{H}_{\mathbf{0}}|\mathrm{H}_{\mathbf{1}}\Big) &= \mathbb{P}\Big(\max_{g\in\widehat{S}} \widehat{c}_g \leq c \,\Big|\, \mu \text{ is not } G\text{-invariant}\Big) \\ &\leq \exp\left(-\frac{n\big(\frac{3c'}{\ell(S)^2} - 2r\gamma b - c\big)^2}{128c_1^2}\right).\end{aligned}$$

Proof. We note that the first inequality follows similarly to the proof of Theorem 6. Thus, we focus on the proof of the second inequality.

We follow the same notation and arguments as in the proof of Theorem 6 to conclude that there exists $\widehat{g} \in S$ such that $c_{\hat{g}} \geq 3c'/\ell(S)^2$. Now, note that we have

$$\begin{aligned}|c_g - c_{\hat{g}}| &= 2|\mathbb{E}[(K(x', gx) - K(x', \widehat{g}x))]| \\ &\leq 2r\mathbb{E}[\|(g - \widehat{g}x)\|_2] \\ &\leq 2r\|g - \widehat{g}\|_{\mathrm{op}}\mathbb{E}[\|x\|_2] \\ &= 2rb\|g - \widehat{g}\|_{\mathrm{op}}.\end{aligned}$$

Therefore, using that $\widehat{g} \in S$ and $\widehat{S}$ is a γ-covering of S, we conclude that there exists $g' \in \widehat{S}$ such that $c_{g'} \geq 3c'/\ell(S)^2 - 2r\gamma b$. The rest of the proof follows similarly to the proof of Theorem 6. We are done.

$\square$