A Unified Framework for Statistical Testing of Invariance

Ashkan Soleymani [*1] **Behrooz Tahmasebi** [*2] **Patrick Jaillet** [1] **Stefanie Jegelka** [3 2]

Abstract

While invariances naturally arise in almost any type of real-world data, no efficient and robust test exists for detecting them in observational data under arbitrarily given group actions. We tackle this problem by studying discrepancy-based measures of invariance that can capture even subtle distributional asymmetries. Our first contribution is to show that, while detecting worst-case asymmetries can be *computationally intractable*, a randomized method can estimate closeness measures to invariance within *universal constant factors*. This provides a general framework for statistical testing of invariance under compact group actions. Despite the extensive and well-established literature on group-based testing, our methodology, to the best of our knowledge, is the *first* to provide statistical tests for general group invariances with *finite-sample guarantees on Type II errors* against worst-case alternatives. We instantiate the framework for common probability discrepancies, including total variation, Wasserstein distances, integral probability metrics, energy distance, and maximum mean discrepancy, obtaining explicit sample-complexity guarantees from empirical convergence rates.

1. Introduction

Invariance is a central organizing principle across science. Across empirical disciplines, observed data often contain patterns that persist under transformations of viewpoint, coordinates, time, or representation (Bronstein et al., 2017). Physics provides a canonical example, where fundamental laws are expected to remain stable under changes of coordinates and translations in time, thereby supporting the universality of the principles they describe (Wigner, 1949, 1964; Smidt, 2021). Machine learning has successfully adopted this principle by designing models that respect known symmetries of the data, often improving both statistical efficiency and computational behavior (Bronstein et al., 2017). Yet this design choice rests on an assumption that is too often imposed rather than examined. Before enforcing invariance, whether through model architecture or post-processing, one must first determine whether the observed data distribution is truly invariant under the proposed group action. This verification problem is the central focus of our work.

Group invariance hypothesis testing draws on a broad and well-developed statistical tradition, including permutation tests and randomization tests (Westfall & Young, 1993; Tusher et al., 2001; Anderson & Robinson, 2001; Onghena, 2017; Hemerik & Goeman, 2021; Koning & Hemerik, 2024; Domingo-Enrich et al., 2025). At their core, these nonparametric procedures test whether the data distribution remains unchanged under the action of a prescribed group G of transformations (Lehmann et al., 1986). Algebraically, the group G is closed under composition, contains an identity element, and assigns an inverse to every element $g \in G$. Recent work by Koning & Hemerik (2024); Koning (2024) and Hemerik (2024) emphasizes that this group-invariance viewpoint already appears in classical sign-flipping tests, tracing the idea back to the early contributions of Fisher (1935); Fisher et al. (1966) and Efron (1969). This perspective also suggests that several familiar procedures, including t-tests (Eden & Yates, 1933; Lehmann & Stein, 1949), can be interpreted as tests of group invariance.

Despite this rich history, existing methods are typically built around the structure of a particular transformation class, such as permutations, sign flips, or rotations, or around carefully chosen subgroups and test-specific reference distributions. This specialization has led to powerful procedures in important settings, including subgroup-based and coded-subgroup methods for permutation and sign-flipping tests, as well as rotation-based tests for multivariate and genomic applications (Langsrud, 2005; Perry & Owen, 2010; Solari et al., 2014; Wu et al., 2010). These developments provide strong tools for particular symmetries, and often give exact or valid Type I error control. What remains missing, to the best of our knowledge, is a single group-agnostic route from a probability discrepancy D to finite-sample Type II

[*]Equal contribution [1]MIT EECS and MIT LIDS [2]MIT EECS and MIT CSAIL [3]School of CIT, MCML, and MDSI, Technical University of Munich (TUM). Correspondence to: Ashkan Soleymani <ashkanso@mit.edu>, Behrooz Tahmasebi <bzt@mit.edu>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

guarantees for the original worst-case invariance alternative. Our work fills this gap by treating the group action and the discrepancy measure abstractly, and by showing that worst-case non-invariance can be detected through random group elements with non-asymptotic guarantees.

In this paper, we study hypothesis testing for invariance, given a general compact topological group G, which may be finite or infinite and acts on the domain of datapoints $\mathcal{X} \subset \mathbb{R}^d$. We test whether the input distribution $\mu \in \mathcal{P}(\mathcal{X})$ is invariant with respect to the 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 by the existence of $g \in G$ such that $D(\mu, g\mu) \geq \epsilon$, where D is a (pseudo)metric on the probability space $\mathcal{P}(\mathcal{X})$. The threshold ϵ separates the null from alternatives that are statistically detectable from finite samples. Thus, we formulate the problem as follows.

Input. n independent and identically distributed (i.i.d.) samples from an unknown probability distribution μ, a group G acting on the domain of datapoints $\mathcal{X}$, a (pseudo)metric D over the space of probability measures, and a threshold ϵ.
Output. Either H_0 or H_1, where

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

At first sight, it is surprising that such a general methodology has not already emerged from the near-a-century long literature on group-based testing. The formulation above reveals why the problem is difficult. There are two obstacles, and both become severe in modern settings where G can be large, continuous, or high-dimensional.

The first obstacle is computational. The group G may be infinite, as in the orthogonal group $O(d)$, or finite but prohibitively large. For the permutation group P_d, for example, $|G| = d! \sim \sqrt{2\pi d}\left(\frac{d}{e}\right)^d$. For the sign-flipping group F_d, consisting of diagonal matrices with entries in $\{\pm 1\}$, we have $|G| = 2^d$. Certifying the alternative in (1) appears to require finding a witness $\widehat{g}$ such that $D(\mu, \widehat{g}\mu) \geq \epsilon$. For continuous groups, this is typically a highly nonconvex optimization problem. For large finite groups, exhaustive search is infeasible. We show that this difficulty is intrinsic. Even when the distributional estimation problem is removed, and even for a finite group with a discrete measure, computing an exact worst-case witness can be NP-hard.

The second obstacle is statistical. We do not observe μ directly; we only observe the empirical measure $\widehat{\mu}$ induced by n i.i.d. samples. A direct plug-in statistic,

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

is tempting but dangerous. The maximization is performed after seeing the data, so the selected transformation can amplify empirical fluctuations rather than genuine distributional asymmetry. Pointwise accuracy for a fixed g is therefore not enough. One would need uniform finite-sample control over an enormous or continuous orbit of transformations, whose complexity may be combinatorial for finite groups and geometric for continuous ones. Thus the challenge is not only to search over G, but to search in a way that does not overfit finite-sample noise.

Our key result resolves these obstacles by showing that the worst-case search over G is not necessary. Under minimal assumptions on the discrepancy D, the largest displacement of μ along its group orbit is controlled, up to a universal constant, by the average displacement under a Haar-random group element. An informal version of this theorem is provided below, with the formal details deferred to subsequent sections.

Theorem 1.1 (Informal version of Theorem 3.2)**.** *Under the minimal assumption that the (pseudo)metric D is shift-invariant with respect to G,*

$$\mathbb{E}_g[D(\mu, g\mu)] \leq \sup_{g \in G} D(\mu, g\mu) \leq 2\mathbb{E}_g[D(\mu, g\mu)],$$

where the expectation is taken with respect to the left Haar measure over the compact group G.

This theorem is the conceptual hinge of the paper. The quantity $\sup_{g \in G} D(\mu, g\mu)$ is the natural object for certifying non-invariance, but it looks computationally and statistically hostile. Theorem 1.1 shows that, up to constants, it is enough to estimate an average over random group elements. The formal result gives sharper constants for genuine pseudometrics and explicit constants for powered or relaxed discrepancies. It applies to many common choices of D, including Wasserstein distances under isometry groups, total variation distance, integral probability metrics, energy distance, and MMD-type discrepancies with invariant kernels.

The same idea gives a particularly simple testing recipe. Instead of hunting for the most adversarial transformation in G, draw a small number of Haar-random transformations and ask whether any of them visibly moves the empirical distribution. For shift-invariant pseudometrics, the proof gives an even sharper random-witness statement. If

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

then a Haar-random group element satisfies

$$D(\mu, g\mu) \geq \epsilon/2$$

with probability at least $1/2$ (Theorem 3.5). Hence $m \geq \lceil \log_2(2/\delta) \rceil$ random group elements $g_1, \ldots, g_m$ are enough to find such a witness with probability at least $1 - \delta/2$.

General recipe. Draw $g_1, \ldots, g_m$ independently from the normalized Haar measure on G, compute

$$D(\widehat{\mu}, g_j \widehat{\mu}) \qquad \text{for } j = 1, \ldots, m,$$

and reject invariance when

$$\max_{1 \leq j \leq m} D(\widehat{\mu}, g_j \widehat{\mu}) > \epsilon/4.$$

The stability of this plug-in rule follows directly from shift-invariance. For every $ginG$,

$$|D(\widehat{\mu}, g\widehat{\mu}) - D(\mu, g\mu)| \leq 2D(\widehat{\mu}, \mu).$$

Therefore, once the empirical measure estimates μ in the chosen discrepancy to accuracy $O(\epsilon)$, the randomized group search yields finite-sample control of both errors. In particular, if $\mathbb{P}\left(D(\widehat{\mu}, \mu) > \epsilon/16\right) \leq \delta/2$, then the above test has Type I error at most δ and Type II error at most δ against the original alternative in (1). Therefore, the key factor controlling the sample complexity of our unified testing framework is the rate of convergence of $\widehat{\mu}$ to μ in the discrepancy measure (or metric) D as a function of n and d; that is, the rate at which the term $D(\widehat{\mu}, \mu)$ converges to zero as the number of samples n increases.

This reduction is the main contribution of the paper. It separates the intrinsic statistical difficulty of estimating the discrepancy D from the extrinsic difficulty of searching over G. Classical and modern group-randomization methods provide powerful tests for specific symmetries, often with exact Type I validity. By contrast, our framework gives a general path to finite-sample Type II guarantees for worst-case invariance testing over compact group actions. To the best of our knowledge, this is the first methodology that provides such guarantees for general group invariances and probability distance metrics.

2. Background

Throughout this paper, we consider a complete metric space $\mathcal{X}$ and 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 continuously on $\mathcal{X}$. Specifically, each group element g corresponds to a continuous bijection on $\mathcal{X}$, with the group operation given by function composition. For any probability measure $\mu \in \mathcal{P}(\mathcal{X})$, let $g\mu \in \mathcal{P}(\mathcal{X})$ denote the pushforward measure under the action of g on $\mathcal{X}$. Similarly, we define μ^G as the distribution of gx when $x \sim \mu$ and $g \in G$ is drawn independently according to the left Haar (uniform) measure on G. A probability (pseudo)metric $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ is a (pseudo)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 $\mu, \nu \in \mathcal{P}(\mathcal{X})$.

3. Main Results

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

Theorem 3.1 (Computational intractability)**.** *There exist a shift-invariant pseudometric $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_{\geq 0}$, a finite group G, and a discrete probability measure μ such that the optimization problem*

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

is NP-hard.

The proof of Theorem 3.1 is presented in Section B.1. We carefully construct a shift-invariant pseudometric D, a finite group G, and a Dirac measure μ such that finding $\arg\sup_{g \in G} D(\mu, g\mu)$ solves a special variant of the Travelling Salesman Problem (TSP), which we prove to be NP-complete. Theorem 3.1 implies that, even in the simplest settings, the optimization problem (1) is computationally intractable, not to mention the statistical challenges in estimating $\sup_{g \in G} D(\mu, g\mu)$ from observations. Next, we state our main theorem, which enables a randomized approximation for $\sup_{g \in G} D(\mu, g\mu)$ instead.

Theorem 3.2 (Probabilistic approximation (formal version of Theorem 1.1))**.** *Let $\mathcal{X}$ be a complete metric space and let $\mathcal{P}(\mathcal{X})$ denote the space of Borel probability measures on $\mathcal{X}$. Let G be a compact topological group acting continuously on $\mathcal{X}$, and let α denote its normalized left Haar measure. Suppose that*

$$D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_{\geq 0}$$

is G-invariant, meaning that

$$D(g\mu, g\nu) = D(\mu, \nu), \qquad \forall g \in G,\ \mu, \nu \in \mathcal{P}(\mathcal{X}),$$

and satisfies the one-sided triangle inequality

$$D(\mu, \nu) \leq D(\mu, \eta) + D(\eta, \nu), \qquad \forall \mu, \nu, \eta \in \mathcal{P}(\mathcal{X}).$$

Then, for every $\mu \in \mathcal{P}(\mathcal{X})$,

$$\mathbb{E}_{g \sim \alpha}\left[D(\mu, g\mu)\right] \leq \sup_{g \in G} D(\mu, g\mu) \leq 2\,\mathbb{E}_{g \sim \alpha}\left[D(\mu, g\mu)\right].$$

Note that the theorem does not require D to be symmetric. It is enough for D to be nonnegative, G-invariant, and to

satisfy the one-sided triangle inequality

$$D(\mu,\nu) \le D(\mu,\eta) + D(\eta,\nu), \qquad \forall \mu,\nu,\eta \in \mathcal{P}(\mathcal{X}).$$

Thus, the result applies not only to probability pseudometrics, but also to asymmetric discrepancies satisfying this directed triangle inequality.

The proof uses only the triangle inequality, the invariance of D, and the invariance of Haar measure. Fix $\mu \in \mathcal{P}(\mathcal{X})$ and $h \in G$. For any $g \in G$, the one-sided triangle inequality gives

$$D(\mu, h\mu) \le D(\mu, g\mu) + D(g\mu, h\mu).$$

By G-invariance of D, the second term can be rewritten as

$$D(g\mu, h\mu) = D(\mu, g^{-1}h\mu).$$

Averaging over $g \sim \alpha$, and using the fact that the map $g \mapsto g^{-1}h$ preserves Haar measure, we obtain

$$D(\mu, h\mu) \le 2\,\mathbb{E}_{g\sim\alpha} D(\mu, g\mu).$$

Since this holds for every $h \in G$, taking the supremum over h yields

$$\sup_{h\in G} D(\mu, h\mu) \le 2\,\mathbb{E}_{g\sim\alpha} D(\mu, g\mu).$$

The reverse inequality follows immediately because an average is bounded above by the supremum.

Remark 3.3. The shift invariance of D with respect to the group G in Theorem 3.2 is a general assumption satisfied in many settings, including the Wasserstein distance with any isometry group G; Sobolev Integral Probability Metrics (IPMs) with any isometry group G; Total Variation (TV) distance with any group G; Maximum Mean Discrepancy (MMD) distance with shift-invariant kernels; and energy distance with any isometry group G. The list continues beyond these examples.

One might argue that the left Haar (uniform) measure over the compact group G is not accessible in specific applications. Nevertheless, we can show that our results still extend to non-uniform distributions with full support, which we formalize below.

Corollary 3.4 (Probabilistic approximation with a non-uniform distribution)**.** *In light of Theorem 3.2, let α be the (uniform) left Haar measure on G, assumed to be inaccessible. Let β be an accessible but not necessarily uniform distribution on G, and suppose $\left|\frac{d\alpha}{d\beta}\right| \le B$ for some constant B. Then,*

$$\mathbb{E}_{g\sim\beta}[D(\mu, g\mu)] \le \sup_{g\in G} D(\mu, g\mu) \le 2B\,\mathbb{E}_{g\sim\beta}[D(\mu, g\mu)],$$

where the expectation is over β on G.

The proof of this corollary follows directly from a simple change of measure and Theorem 3.2. This result ensures that our constant-factor approximation for this class of intractable problems remains valid even when the uniform measure over G is inaccessible. Any measure β satisfying $|d\beta/d\alpha| \ge 1/B$ suffices for the randomized algorithm.

3.1. Half-Probability Certificate

Next, we complement the average-to-worst-case comparison with a probabilistic certificate. While the previous result relates to the average discrepancy over the group to the maximum discrepancy, the next theorem shows that this comparison also has a direct random-sampling interpretation: a single Haar-random group element detects a constant fraction of the worst-case asymmetry with constant probability.

Theorem 3.5 (Half-probability certificate for distance to asymmetry)**.** *Let G be a compact group with the normalized Haar measure α. Let $\mathcal{X}$ be a complete metric space and let $\mathcal{Y} = \mathcal{P}(\mathcal{X})$. Suppose G acts continuously on $\mathcal{Y}$, and let $D : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}_{\ge 0}$ be a G-invariant pseudometric, meaning that*

$$D(g\mu, g\nu) = D(\mu,\nu)$$

for all $g \in G$ and $\mu,\nu \in \mathcal{Y}$. Then, for every $\mu \in \mathcal{Y}$,

$$\mathbb{P}_{g\sim\alpha}\left(D(\mu, g\mu) \ge \frac{1}{2}\sup_{h\in G} D(\mu, h\mu)\right) \ge \frac{1}{2}.$$

Thus, asymmetry is not only detectable through the full group average or the worst-case group element, but also through random sampling. In particular, if μ is far from being invariant in the worst-case sense, then a Haar-random transformation witnesses at least half of this displacement with probability at least $1/2$. This provides a simple justification for randomized certificates of measure asymmetry.

As an immediate consequence, if $g_1, \ldots, g_\kappa \overset{\text{i.i.d.}}{\sim} \alpha$, then

$$\mathbb{P}\left(\max_{1\le i\le\kappa} D(\mu, g_i\mu) \ge \frac{1}{2}\sup_{h\in G} D(\mu, h\mu)\right) \ge 1 - 2^{-\kappa}.$$

Thus $\kappa = \lceil \log_2(1/\delta) \rceil$ random group samples suffice to obtain a half-optimal asymmetry certificate with probability at least $1-\delta$.

The half-probability certificate is stronger than what follows from the average-to-worst-case comparison alone. Indeed, the average comparison implies

$$\mathbb{E}_{g\sim\alpha} D(\mu, g\mu) \ge \frac{1}{2} M_\mu, \qquad M_\mu := \sup_{h\in G} D(\mu, h\mu).$$

Fix $\kappa \in [0, 1/2)$. Applying Markov's inequality to $M_\mu -$

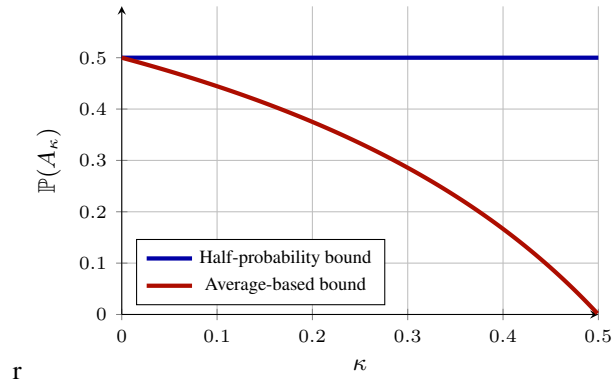

Figure 1. Comparison of lower bounds on $\mathbb{P}(A_\kappa)$, where $A_\kappa = \{D(\mu, g\mu) \geq \kappa M_\mu\}$ and $M_\mu = \sup_{h\in G} D(\mu, h\mu)$. The average-based guarantee degenerates as $\kappa \uparrow 1/2$, while the half-probability theorem remains at $1/2$.

$D(\mu, g\mu)$, we obtain

$$\begin{aligned}
&\mathbb{P}_{g\sim\alpha}\left(D(\mu, g\mu) < \kappa M_\mu\right) \\
&\quad = \mathbb{P}_{g\sim\alpha}\left(M_\mu - D(\mu, g\mu) > (1-\kappa)M_\mu\right) \\
&\quad \leq \frac{\mathbb{E}_{g\sim\alpha}\left[M_\mu - D(\mu, g\mu)\right]}{(1-\kappa)M_\mu} \leq \frac{1}{2(1-\kappa)}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}_{g\sim\alpha}\left(D(\mu, g\mu) \geq \kappa M_\mu\right) \\
&\quad \geq 1 - \frac{1}{2(1-\kappa)} = \frac{\frac{1}{2}-\kappa}{1-\kappa}.
\end{aligned}$$

This bound is strictly smaller than $1/2$ for every $\kappa \in (0, 1/2)$, and it vanishes as $\kappa \uparrow 1/2$.

By contrast, Theorem 3.5 gives directly

$$\mathbb{P}_{g\sim\alpha}\left(D(\mu, g\mu) \geq \frac{1}{2}M_\mu\right) \geq \frac{1}{2}.$$

Consequently, for every $\kappa \in [0, 1/2]$,

$$\mathbb{P}_{g\sim\alpha}\left(D(\mu, g\mu) \geq \kappa M_\mu\right) \geq \frac{1}{2}.$$

Thus, the probabilistic theorem gives a stronger random-certificate statement: it remains nontrivial at the half-optimal threshold, whereas the guarantee obtained only from the average comparison degenerates at that point.

3.2. Finite-Sample Randomized Testing

We now turn the random-witness theorem into a finite-sample test. Draw $g_1, \ldots, g_m$ independently from the normalized Haar measure on G, independently of the sample used to form $\widehat{\mu}$. The randomized plug-in test rejects invariance when

$$\max_{1\leq j\leq m} D(\widehat{\mu}, g_j\widehat{\mu}) > \frac{\epsilon}{4}.$$

The point is that the random search over the group and the statistical estimation error separate. The first is controlled by the half-probability certificate; the second is exactly the rate at which $D(\widehat{\mu}, \mu)$ converges to zero.

Proposition 3.6 (Finite-sample error decomposition)**.** *Let D be a G-invariant pseudometric. For the randomized plug-in test above, the following two bounds hold.*

If μ is invariant, then

$$\mathbb{P}_\mu(\text{reject}) \leq \mathbb{P}_\mu\left(D(\widehat{\mu}, \mu) > \frac{\epsilon}{16}\right).$$

If $\sup_{g\in G} D(\mu, g\mu) \geq \epsilon$, then

$$\mathbb{P}_\mu(\text{fail to reject}) \leq \mathbb{P}_\mu\left(D(\widehat{\mu}, \mu) > \frac{\epsilon}{16}\right) + 2^{-m}.$$

All probabilities are over both the sample and the independent Haar (uniform) draws from the group.

Proof. For every $g \in G$, the triangle inequality and G-invariance of D give

$$\begin{aligned}
D(\widehat{\mu}, g\widehat{\mu}) &\leq D(\widehat{\mu}, \mu) + D(\mu, g\mu) + D(g\mu, g\widehat{\mu}) \\
&= D(\mu, g\mu) + 2D(\widehat{\mu}, \mu).
\end{aligned}$$

Exchanging the roles of μ and $\widehat{\mu}$ gives

$$D(\mu, g\mu) \leq D(\widehat{\mu}, g\widehat{\mu}) + 2D(\widehat{\mu}, \mu).$$

Hence, for every $g \in G$,

$$|D(\widehat{\mu}, g\widehat{\mu}) - D(\mu, g\mu)| \leq 2D(\widehat{\mu}, \mu).$$

Thus, on the event $D(\widehat{\mu}, \mu) \leq \epsilon/16$, all sampled empirical discrepancies are within $\epsilon/8$ of their population counterparts. This is the only place where sample complexity enters.

We first control Type I error. If μ is invariant, then $D(\mu, g\mu) = 0$ for every $g \in G$. Therefore, on the event $D(\widehat{\mu}, \mu) \leq \epsilon/16$,

$$\max_{1\leq j\leq m} D(\widehat{\mu}, g_j\widehat{\mu}) \leq \frac{\epsilon}{8} < \frac{\epsilon}{4}.$$

A false rejection can therefore occur only when $D(\widehat{\mu}, \mu) > \epsilon/16$, which proves the first bound.

We next control Type II error. Suppose that

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

By Theorem 3.5, a Haar-random g satisfies

$$D(\mu, g\mu) \geq \frac{\epsilon}{2}$$

Discrepancy D	**Setting**	**Sufficient sample size** n
TV	Finite alphabet of size K. Nonatomic laws require smoothing or discretization, since $\mathrm{TV}(\widehat{\mu}, \mu) = 1$ a.s.	$n \gtrsim \epsilon^{-2}\big(K + \log(1/\delta)\big)$ (Weissman et al., 2003).
Kolmogorov	One-dimensional laws, $D(\widehat{\mu}, \mu) = \sup_t \lvert\widehat{F}(t) - F(t)\rvert$.	$n \gtrsim \epsilon^{-2}\log(1/\delta)$ (Massart, 1990).
MMD	Unsquared MMD with RKHS kernel $k(x,x) \le \kappa^2$.	$n \gtrsim \kappa^2\epsilon^{-2}\log(1/\delta)$ (Gretton et al., 2012; Sriperumbudur et al., 2012).
Energy	Square-root energy distance on a bounded negative-type metric space.	$n \gtrsim \epsilon^{-2}\log(1/\delta)$ (Sejdinovic et al., 2013).
Wasserstein W_p	$\mathcal{X} \subset [0,1]^d$, Euclidean metric.	$n \gtrsim \begin{cases} \epsilon^{-2p}\log(1/\delta), & d < 2p, \\ \epsilon^{-2p}\log^2(1/\epsilon) + \epsilon^{-2p}\log(1/\delta), & d = 2p, \\ \epsilon^{-d} + \epsilon^{-2p}\log(1/\delta), & d > 2p \end{cases}$ (Fournier & Guillin, 2015; Weed & Bach, 2019).
Sobolev IPM	$D(\mu,\nu) = \sup_{\lVert f\rVert_{W^{s,\infty}} \le 1} \left\lvert \int f\,d\mu - \int f\,d\nu \right\rvert$ on $[0,1]^d$.	$n \gtrsim \begin{cases} \epsilon^{-2}\log(1/\delta), & s > d/2, \\ \epsilon^{-2}\log^2(1/\epsilon) + \epsilon^{-2}\log(1/\delta), & s = d/2, \\ \epsilon^{-d/s} + \epsilon^{-2}\log(1/\delta), & 0 < s < d/2 \end{cases}$ (Dudley, 1967; Vaart & Wellner, 1997; Nickl & Pötscher, 2007).
VC-type IPM	$D(\mu,\nu) = \sup_{f \in \mathcal{F}} \left\lvert \int f\,d\mu - \int f\,d\nu \right\rvert$, with $\lvert f\rvert \le 1$, and $\mathrm{VC}(\mathcal{F}) = V$.	$n \gtrsim \epsilon^{-2}\big(V\log(1/\epsilon) + \log(1/\delta)\big)$ (Vaart & Wellner, 1997).

Table 1. Sufficient sample sizes for the randomized plug-in invariance test. Each row ensures $D(\widehat{\mu}, \mu) \le \epsilon/16$ with probability at least $1 - \delta/2$, under the stated assumptions. The number of Haar draws is always $m = \lceil \log_2(2/\delta) \rceil$. Constants may depend on fixed regularity parameters, the domain diameter, and p, s, d, but not on n, ϵ, δ.

with probability at least $1/2$. Since the draws $g_1, \dots, g_m$ are independent,

$$\mathbb{P}\left(\max_{1 \le j \le m} D(\mu, g_j\mu) < \frac{\epsilon}{2}\right) \le 2^{-m}.$$

On the complementary event, there is some j such that $D(\mu, g_j\mu) \ge \epsilon/2$. If also $D(\widehat{\mu}, \mu) \le \epsilon/16$, then

$$\begin{aligned} D(\widehat{\mu}, g_j\widehat{\mu}) &\ge D(\mu, g_j\mu) - 2D(\widehat{\mu}, \mu) \\ &\ge \frac{\epsilon}{2} - \frac{\epsilon}{8} = \frac{3\epsilon}{8} > \frac{\epsilon}{4}. \end{aligned}$$

Thus the test fails to reject only if the empirical measure is inaccurate or if the random group search misses all half-optimal witnesses. This gives the second bound. □

The proposition immediately gives the usual testing guarantees. Define the Type I and Type II errors at separation ϵ by

$$\begin{aligned} \mathrm{Err}_{\mathrm{I}} &= \sup_{\mu:\, \mu = g\mu\ \forall g \in G} \mathbb{P}_\mu(\text{reject}), \\ \mathrm{Err}_{\mathrm{II}}(\epsilon) &= \sup_{\mu:\, \sup_{g \in G} D(\mu, g\mu) \ge \epsilon} \mathbb{P}_\mu(\text{fail to reject}). \end{aligned}$$

If

$$m \ge \lceil \log_2(2/\delta) \rceil$$

and the sample size is large enough that

$$\sup_\mu \mathbb{P}_\mu\left(D(\widehat{\mu}, \mu) > \frac{\epsilon}{16}\right) \le \frac{\delta}{2} \tag{2}$$

over the null and alternative distributions under consideration, then

$$\mathrm{Err}_{\mathrm{I}} \le \delta, \qquad \mathrm{Err}_{\mathrm{II}}(\epsilon) \le \delta.$$

Thus, $O(\log(1/\delta))$ random group elements are enough. The remaining sample complexity is precisely the sample complexity of estimating μ in the discrepancy D.

Concrete sample sizes. We now specialize the condition

$$D(\widehat{\mu}, \mu) \le \epsilon/16$$

from (2) to several common choices of discrepancy. The bounds in Table 1 are sufficient conditions for the randomized invariance test to have both Type I and Type II errors at most δ, together with $m \ge \lceil \log_2(2/\delta) \rceil$. Constants may depend on fixed regularity parameters of the distribution μ, the diameter of the domain $\mathcal{X}$, and the order p (for the Wasserstein distance), but not on n, ϵ, δ.

We conclude this section with several remarks. First, total variation behaves differently from weak metrics in the sense that for continuous distributions, the empirical measure is singular with respect to μ, so TV plug-in estimation is inconsistent without additional structure. Second, kernel and smooth IPM metrics have parametric sample complexity once the discriminator class is sufficiently regular. Third, Wasserstein-type distances pay the familiar dimension-dependent price. These rates are not caused by the group search. They are exactly the rates for estimating

the underlying distribution in the chosen discrepancy. The group contribution remains logarithmic in $1/\delta$, i.e., the error level for Type I and II gurantees.

3.3. Beyond Pseudometric Assumption

The preceding result relies on the triangle inequality for the pseudometric D. It is natural to ask to what extent this assumption can be weakened. In particular, if it still satisfy a metric-like inequality after taking a suitable power, possibly up to a constant factor. The next proposition records that our average-to-worst-case comparison remains valid in this more general setting, with constants depending only on some relaxation factors for the general functions D.

Proposition 3.7 (Average-to-worst-case comparison under a powered relaxed triangle inequality)**.** *Let G be a compact group with normalized Haar measure α, and let $\mathcal{Y} = \mathcal{P}(\mathcal{X})$. Suppose that $D : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is G-invariant and that there exist constants $p \in (0, \infty)$ and $B \geq 1$ such that*

$$D(\eta, \xi)^p \leq B\big(D(\eta, \zeta)^p + D(\zeta, \xi)^p\big)$$

for all $\eta, \xi, \zeta \in \mathcal{Y}$. Then, for every $\mu \in \mathcal{Y}$,

$$\begin{aligned}(\mathbb{E}_{g\sim\alpha}\left[D(\mu, g\mu)^p\right])^{1/p} &\leq \sup_{h\in G} D(\mu, h\mu) \\ &\leq (2B)^{1/p}\left(\mathbb{E}_{g\sim\alpha}\left[D(\mu, g\mu)^p\right]\right)^{1/p}.\end{aligned}$$

Moreover, if $p \geq 1$, then the ordinary average also satisfies

$$\mathbb{E}_{g\sim\alpha}\left[D(\mu, g\mu)\right] \leq \sup_{h\in G} D(\mu, h\mu) \leq 2B\,\mathbb{E}_{g\sim\alpha}\left[D(\mu, g\mu)\right].$$

This proposition shows that the argument is not tied to the exact pseudometric structure of D. What is needed is only enough triangle-inequality structure to run the comparison after a powered transformation of the discrepancy. Thus, even beyond genuine pseudometrics, the average discrepancy over the group still controls the worst-case discrepancy up to an explicit constant.

Example: powered Wasserstein discrepancies. A natural example beyond genuine metrics is the powered Wasserstein discrepancy

$$D(\mu, \nu) := W_r(\mu, \nu)^r, \qquad r \geq 1.$$

Although D is generally not a metric for $r > 1$, its $1/r$-power is the Wasserstein metric W_r. Applying Theorem 3.2 to W_r gives

$$\sup_{h\in G} W_r(\mu, h\mu) \leq 2\,\mathbb{E}_{g\sim\alpha} W_r(\mu, g\mu).$$

Raising both sides to the r-th power yields

$$\sup_{h\in G} W_r(\mu, h\mu)^r \leq 2^r\left(\mathbb{E}_{g\sim\alpha} W_r(\mu, g\mu)\right)^r.$$

By Jensen's inequality,

$$\left(\mathbb{E}_{g\sim\alpha} W_r(\mu, g\mu)\right)^r \leq \mathbb{E}_{g\sim\alpha} W_r(\mu, g\mu)^r.$$

Therefore,

$$\mathbb{E}_{g\sim\alpha} D(\mu, g\mu) \leq \sup_{h\in G} D(\mu, h\mu) \leq 2^r\,\mathbb{E}_{g\sim\alpha} D(\mu, g\mu).$$

Thus, even though W_r^r is not itself a metric, its worst-case value over the group is still controlled, up to the explicit factor 2^r, by its average value over the group.

Example: powers of MMD. Let k be a positive definite kernel with RKHS $\mathcal{H}$, and consider the maximum mean discrepancy

$$\mathrm{MMD}_{\mathcal{H}}(\mu, \nu) := \sup_{\|f\|_{\mathcal{H}}\leq 1} \left|\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]\right|.$$

Since $\mathrm{MMD}_{\mathcal{H}}$ is a pseudometric, its q-th power

$$D(\mu, \nu) := \mathrm{MMD}_{\mathcal{H}}(\mu, \nu)^q, \qquad q \geq 1,$$

is generally not a pseudometric. However, $D^{1/q} = \mathrm{MMD}_{\mathcal{H}}$ is a pseudometric, so the powered relaxed triangle condition holds with $p = 1/q$ and $B = 1$. Therefore, the previous proposition gives

$$\begin{aligned}\left(\mathbb{E}_{g\sim\alpha}\mathrm{MMD}_{\mathcal{H}}(\mu, g\mu)\right)^q &\leq \sup_{h\in G} \mathrm{MMD}_{\mathcal{H}}(\mu, h\mu)^q \\ &\leq 2^q\left(\mathbb{E}_{g\sim\alpha}\mathrm{MMD}_{\mathcal{H}}(\mu, g\mu)\right)^q.\end{aligned}$$

By Jensen's inequality, this also implies the coarser but often more directly usable bound

$$\begin{aligned}\mathbb{E}_{g\sim\alpha}\mathrm{MMD}_{\mathcal{H}}(\mu, g\mu)^q &\leq \sup_{h\in G} \mathrm{MMD}_{\mathcal{H}}(\mu, h\mu)^q \\ &\leq 2^q\,\mathbb{E}_{g\sim\alpha}\mathrm{MMD}_{\mathcal{H}}(\mu, g\mu)^q.\end{aligned}$$

Example: powers of integral probability metrics. More generally, let $\mathcal{F}$ be a class of test functions and define the integral probability metric

$$d_{\mathcal{F}}(\mu, \nu) := \sup_{f\in\mathcal{F}} \left|\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]\right|.$$

Whenever $d_{\mathcal{F}}$ is finite, it is a pseudometric on $\mathcal{P}(\mathcal{X})$. Thus, for any $q \geq 1$, the powered discrepancy

$$D(\mu, \nu) := d_{\mathcal{F}}(\mu, \nu)^q$$

is generally not a pseudometric, but satisfies the powered relaxed triangle condition with $p = 1/q$ and $B = 1$, since $D^{1/q} = d_{\mathcal{F}}$. Applying the previous proposition yields

$$\begin{aligned}\left(\mathbb{E}_{g\sim\alpha} d_{\mathcal{F}}(\mu, g\mu)\right)^q &\leq \sup_{h\in G} d_{\mathcal{F}}(\mu, h\mu)^q \\ &\leq 2^q\left(\mathbb{E}_{g\sim\alpha} d_{\mathcal{F}}(\mu, g\mu)\right)^q.\end{aligned}$$

By Jensen's inequality, we also obtain

$$\mathbb{E}_{g\sim\alpha} d_{\mathcal{F}}(\mu, g\mu)^q \le \sup_{h\in G} d_{\mathcal{F}}(\mu, h\mu)^q \le 2^q\, \mathbb{E}_{g\sim\alpha} d_{\mathcal{F}}(\mu, g\mu)^q.$$

Thus, powers of IPMs provide a broad class of non-metric discrepancies for which the average-to-worst-case comparison still holds up to an explicit constant.

Example: squared MMD. Let k be a positive definite kernel with RKHS $\mathcal{H}$. Since $\mathrm{MMD}_{\mathcal{H}}$ is a pseudometric, its square satisfies the relaxed triangle inequality with $p = 1$ and $B = 2$. Hence,

$$\mathbb{E}_{g\sim\alpha}\mathrm{MMD}^2_{\mathcal{H}}(\mu, g\mu) \le \sup_{h\in G} \mathrm{MMD}^2_{\mathcal{H}}(\mu, h\mu) \le 4\, \mathbb{E}_{g\sim\alpha}\mathrm{MMD}^2_{\mathcal{H}}(\mu, g\mu).$$

Moreover, squared MMD has the closed form

$$\mathrm{MMD}^2_{\mathcal{H}}(\mu, g\mu) = \mathbb{E}k(X, X') + \mathbb{E}k(gX, gX') - 2\mathbb{E}k(X, gX'),$$

where $X, X' \sim \mu$ independently. If k is G-invariant, meaning that $k(gx, gx') = k(x, x')$ for all $g \in G$ and all $x, x' \in \mathcal{X}$, this simplifies to

$$\mathrm{MMD}^2_{\mathcal{H}}(\mu, g\mu) = 2\mathbb{E}k(X, X') - 2\mathbb{E}k(X, gX').$$

Consequently, in the G-invariant kernel case,

$$\begin{aligned}&\mathbb{E}k(X, X') - \mathbb{E}_{g\sim\alpha}\mathbb{E}k(X, gX') \\ &\quad\le \sup_{h\in G} \{\mathbb{E}k(X, X') - \mathbb{E}k(X, hX')\} \\ &\quad\le 4\left(\mathbb{E}k(X, X') - \mathbb{E}_{g\sim\alpha}\mathbb{E}k(X, gX')\right),\end{aligned}$$

where $X, X' \sim \mu$ independently in each inner expectation. Therefore, the proposition gives an explicit average-to-worst-case comparison for the closed-form squared-MMD discrepancy along the group orbit of μ.

3.4. Comparing Distributions Along Group Orbits

We now record a simple extension in which the two arguments of the discrepancy come from different group orbits. Instead of measuring the displacement $D(\mu, g\mu)$ of a single distribution under the group action, we consider $D(\mu, g\nu)$, which compares μ to transformed copies of another distribution ν. This setting captures how the discrepancy between two distributions changes as one of them is moved along its orbit. The next result shows that the worst-case cross-orbit discrepancy is still controlled by its average over random group elements, now with a universal factor 3.

Proposition 3.8 (Cross-orbit average-to-worst-case comparison)**.** *Let G be a compact group with normalized Haar measure α, acting on $\mathcal{Y} = \mathcal{P}(\mathcal{X})$. Let $D : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\ge 0}$ be a G-invariant pseudometric. Then, for every $\mu, \nu \in \mathcal{Y}$,*

$$\mathbb{E}_{g\sim\alpha} D(\mu, g\nu) \le \sup_{h\in G} D(\mu, h\nu) \le 3\, \mathbb{E}_{g\sim\alpha} D(\mu, g\nu).$$

The cross-orbit quantity $D(\mu, g\nu)$ compares μ with transformed copies of another distribution ν. Thus, while $D(\mu, g\mu)$ measures non-invariance of a single distribution, $D(\mu, g\nu)$ measures how the discrepancy between two distributions changes along the orbit of ν. In particular,

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

measures the largest increase in discrepancy obtained by moving ν along its group orbit. It is small when the baseline discrepancy $D(\mu, \nu)$ is representative of the entire orbit, and large when some transformed copy of ν is substantially farther from μ. The proposition shows that this worst-case cross-orbit discrepancy is controlled, up to a universal factor, by the average discrepancy over random group transformations. Consequently,

$$\sup_{g\in G} D(\mu, g\nu) - D(\mu, \nu) \le 3\, \mathbb{E}_{g\sim\alpha} D(\mu, g\nu) - D(\mu, \nu).$$

Therefore, random group samples can be used not only to certify non-invariance of a single distribution, but also to detect orbit-induced amplification of the discrepancy between two distributions.

4. Conclusion

We proposed a unified framework for statistical testing of invariance under general compact group actions. Our results show that, under mild assumptions, the worst-case discrepancy along a group orbit is controlled up to a universal constant by the average discrepancy under a Haar-random group transformation. We also showed that a single random group element provides a half-probability certificate of non-invariance.

This reduction separates the difficulty of searching over the group from the statistical difficulty of estimating the chosen discrepancy from samples. Once the empirical measure approximates the population measure in D, randomized group sampling yields finite-sample guarantees against worst-case invariance alternatives.

Future directions include sharpening the constants and rates in our comparisons, as well as extending the theory to non-compact group actions, where Haar averaging, tail behavior, and the formulation of suitable alternatives require additional care.

Acknowledgments

AS and PJ were partially supported by the ONR grant N00014-24-1-2470. BT and SJ were partially supported by the Office of Naval Research under grant N00014-20-1-2023 (MURI ML-SCOPE) and by the National Science Foundation under award CCF-2112665 (TILOS AI Institute). SJ also acknowledges support from an Alexander von Humboldt Professorship.

References

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

Bellot, A. and van der Schaar, M. Application of kernel hypothesis testing on set-valued data. In *Uncertainty in Artificial Intelligence*, pp. 194–204. PMLR, 2021. 12

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4): 18–42, 2017. 1

Caselles-Dupré, H., Garcia Ortiz, M., and Filliat, D. Symmetry-based disentangled representation learning requires interaction with environments. *Advances in Neural Information Processing Systems*, 32, 2019. 12

Chiu, K. and Bloem-Reddy, B. Hypothesis tests for distributional group symmetry with applications to particle physics. In *NeurIPS 2023 AI for Science Workshop*, 2023. 12

Dehmamy, N., Walters, R., Liu, Y., Wang, D., and Yu, R. Automatic symmetry discovery with lie algebra convolutional network. *Advances in Neural Information Processing Systems*, 34:2503–2515, 2021. 12

Dobriban, E. Consistency of invariance-based randomization tests. *The Annals of Statistics*, 50(4):2443–2466, 2022. 12

Domingo-Enrich, C., Dwivedi, R., and Mackey, L. Cheap permutation testing. *arXiv preprint arXiv:2502.07672*, 2025. 1

Dudley, R. M. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967. 6

Eden, T. and Yates, F. 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

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

Fisher, R. The design of experiments. 1935. 1

Fisher, R. A., Fisher, R. A., Genetiker, S., Fisher, R. A., Genetician, S., Britain, G., Fisher, R. A., and Généticien, S. *The design of experiments*, volume 21. Springer, 1966. 1

Fournier, N. and Guillin, A. On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3):707–738, 2015. 6

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012. 6

Hemerik, J. On the term "randomization test". *The American Statistician*, pp. 1–8, 2024. 1

Hemerik, J. and Goeman, J. J. Another look at the lady tasting tea and differences between permutation tests and randomisation tests. *International Statistical Review*, 89 (2):367–381, 2021. 1

Hinton, G. E., Krizhevsky, A., and Wang, S. D. 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*, pp. 44–51. Springer, 2011. 12

Kashlak, A. B. Asymptotic symmetry and group invariance for randomization. *arXiv preprint arXiv:2211.00144*, 2022. 12

Koning, N. W. More power by using fewer permutations. *Biometrika*, pp. asae031, 2024. 1, 12

Koning, N. W. and Hemerik, J. More efficient exact group invariance testing: using a representative subgroup. *Biometrika*, 111(2):441–458, 2024. 1, 12

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 12

Langsrud, Ø. Rotation tests. *Statistics and computing*, 15: 53–60, 2005. 1

Law, H. C., Yau, C., and Sejdinovic, D. Testing and learning on distributions with symmetric noise invariance. *Advances in Neural Information Processing Systems*, 30, 2017. 12

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

Lehmann, E. L., Romano, J. P., and Casella, G. *Testing statistical hypotheses*, volume 3. Springer, 1986. 1

Mallat, S. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012. 12

Marchetti, G. L., Tegnér, G., Varava, A., and Kragic, D. Equivariant representation learning via class-pose decomposition. In *International Conference on Artificial Intelligence and Statistics*, pp. 4745–4756. PMLR, 2023. 12

Massart, P. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pp. 1269–1283, 1990. 6

Moskalev, A., Sepliarskaia, A., Sosnovik, I., and Smeulders, A. Liegg: Studying learned lie group generators. *Advances in Neural Information Processing Systems*, 35: 25212–25223, 2022. 12

Nickl, R. and Pötscher, B. M. Bracketing metric entropy rates and empirical central limit theorems for function classes of besov-and sobolev-type. *Journal of Theoretical Probability*, 20(2):177–199, 2007. 6

Onghena, P. Randomization tests or permutation tests? a historical and terminological clarification. In *Randomization, masking, and allocation concealment*, pp. 209–228. Chapman and Hall/CRC, 2017. 1

Perry, P. O. and Owen, A. B. A rotation test to verify latent structure. *Journal of Machine Learning Research*, 11(2), 2010. 1

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 12

Quessard, R., Barrett, T., and Clements, W. Learning disentangled representations and group structure of dynamical environments. *Advances in Neural Information Processing Systems*, 33:19727–19737, 2020. 12

Ramdas, A., Barber, R. F., Candès, E. J., and Tibshirani, R. J. Permutation tests using arbitrary permutation distributions. *Sankhya A*, 85(2):1156–1177, 2023. 12

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008. 12

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The annals of statistics*, pp. 2263–2291, 2013. 6

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

Solari, A., Finos, L., and Goeman, J. J. Rotation-based multiple testing in the multivariate linear model. *Biometrics*, 70(4):954–961, 2014. 1

Soleymani, A., Tahmasebi, B., Jegelka, S., and Jaillet, P. Learning with exact invariances in polynomial time. In *Int. Conference on Machine Learning (ICML)*, 2025a. 12

Soleymani, A., Tahmasebi, B., Jegelka, S., and Jaillet, P. A robust kernel statistical test of invariance: Detecting subtle asymmetries. In *Int. Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025b. 12

Soleymani, A., Tahmasebi, B., Jaillet, P., and Jegelka, S. Efficient learning and symmetry discovery under exact invariances. In *Conference on Learning Theory (COLT)*, 2026. 12

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012. 6

Tahmasebi, B. and Jegelka, S. The exact sample complexity gain from invariances for kernel regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 12

Tahmasebi, B. and Jegelka, S. Sample complexity bounds for estimating probability divergences under invariances. In *Int. Conference on Machine Learning (ICML)*, 2024. 12

Tahmasebi, B. and Jegelka, S. Generalization bounds for canonicalization: A comparative study with group averaging. In *Int. Conference on Learning Representations (ICLR)*, 2025a. 12

Tahmasebi, B. and Jegelka, S. Regularity in canonicalized models: A theoretical perspective. In *Int. Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025b. 12

Tahmasebi, B. and Weber, M. Achieving approximate symmetry is exponentially easier than exact symmetry. In *Int. Conference on Learning Representations (ICLR)*, 2026a. 12

Tahmasebi, B. and Weber, M. Adaptive symmetry discovery for dynamical system identification. In *Int. Conference on Machine Learning (ICML)*, 2026b. 12

Tahmasebi, B., Weber, M., and Jegelka, S. Data augmentation: A fourier analysis perspective. In *Conference on Learning Theory (COLT)*, 2026. 12

Tusher, V. G., Tibshirani, R., and Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001. 1

Vaart, A. v. d. and Wellner, J. A. Weak convergence and empirical processes with applications to statistics. *Journal of the Royal Statistical Society-Series A Statistics in Society*, 160(3):596–608, 1997. 6

Weed, J. and Bach, F. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019. 6

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, pp. 125, 2003. 6

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

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

Wigner, E. P. Events, laws of nature, and invariance principles. *Science*, 145(3636):995–999, 1964. 1

Wu, D., Lim, E., Vaillant, F., Asselin-Labat, M.-L., Visvader, J. E., and Smyth, G. K. Roast: rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17): 2176–2182, 2010. 1

Yang, J., Dehmamy, N., Walters, R., and Yu, R. Latent space symmetry discovery. In *International Conference on Machine Learning*, 2023a. 12

Yang, J., Walters, R., Dehmamy, N., and Yu, R. Generative adversarial symmetry discovery. In *International Conference on Machine Learning*, pp. 39488–39508. PMLR, 2023b. 12

Yu, H.-X., Wu, J., and Yi, L. Rotationally equivariant 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1456–1464, 2022. 12

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. *Advances in neural information processing systems*, 30, 2017. 12

Zhou, A., Knowles, T., and Finn, C. Meta-learning symmetries by reparameterization. In *International Conference on Learning Representations*, 2021. 12

A. Related Work

Invariance testing has been approached from several directions. One line of work changes the discrepancy itself, for example by encoding invariance to additive symmetric noise or by adapting kernel tests to set-valued observations (Law et al., 2017; Bellot & van der Schaar, 2021). Another line studies randomization tests and approximate invariance in structured models. Dobriban (2022) analyzes invariance-based randomization tests in signal-plus-noise models, while Kashlak (2022) studies asymptotic symmetry under group actions.

Closer to our setting are works that ask which transformations should be used in the test. Koning & Hemerik (2024) replace random subsets of transformations by carefully chosen fixed subgroups, and Koning (2024) study how small coded subgroups can improve power while reducing computation. For permutation tests, Ramdas et al. (2023) show that valid randomized p-values can be obtained from arbitrary subsets and non-uniform distributions over permutations. A complementary viewpoint is due to Chiu & Bloem-Reddy (2023), who quantify non-invariance through the distance between a distribution and its orbit average.

Our goal is different. We test the worst-case alternative in (1), where non-invariance is certified by finding a group element that moves the distribution by at least ϵ. The main point is that this worst-case search over a compact group does not have to be solved directly. For shift-invariant discrepancies, a small number of Haar-random transformations is enough, and the remaining sample complexity is exactly the cost of estimating μ in the chosen discrepancy.

Learning and Symmetries. Incorporating symmetries into machine learning models can substantially improve sample complexity (Tahmasebi & Jegelka, 2023, 2024) while preserving computational feasibility (Soleymani et al., 2025a, 2026). Subsequently, the design of invariant machine learning models 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 (Krizhevsky et al., 2012; Mallat, 2012). Deep Sets were developed to handle set-structured data (Zaheer et al., 2017), and PointNets were proposed for point cloud data that are permutation-invariant (Qi et al., 2017). Graph Neural Networks (GNNs) (Scarselli et al., 2008) were designed for graph-structured data. Beyond these models, model-agnostic methods for exploiting invariance in machine learning, such as canonicalization and data augmentation, are also relevant in this area; see, for example, (Tahmasebi & Jegelka, 2025a,b; Tahmasebi et al., 2026).

Recent efforts have explored alternative approaches, such as automatically discovering the underlying symmetries in data (Tahmasebi & Weber, 2026b; Zhou et al., 2021; Dehmamy et al., 2021; Moskalev et al., 2022; Yang et al., 2023b,a) or relaxing the invariance (Tahmasebi & Weber, 2026a). Another line of work focuses on learning equivariant representations given known symmetries (Hinton et al., 2011; Yu et al., 2022), particularly targeting symmetric disentangled representations (Caselles-Dupré et al., 2019; Quessard et al., 2020; Marchetti et al., 2023). Despite this extensive body of work, the problem of testing invariances (which is central to our work) remains relatively underexplored in the machine learning literature.

An earlier version of this paper focused on kernel distances for hypothesis testing (Soleymani et al., 2025b). In this paper, we provide a unified viewpoint for general probability divergences and improve upon some of those initial results.

B. Proofs of the Main Results

B.1. Proof of Theorem 3.1

Theorem 3.1 (Computational intractability)**.** *There exist a shift-invariant pseudometric $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_{\geq 0}$, a finite group G, and a discrete probability measure μ such that the optimization problem*

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

is NP-hard.

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 Theorem B.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 $g \in G$,

$$
\begin{aligned}
D(\mu, g\mu) &\coloneqq \sum_{i=1}^{d} d\Big(e^{-1}g(i), e^{-1}g(i+1 \bmod d)\Big) + \sum_{i=1}^{d} d\Big(g^{-1}e(i), g^{-1}e(i+1 \bmod d)\Big) \\
&= \sum_{i=1}^{d} d\Big(g(i), g(i+1 \bmod d)\Big) + \sum_{i=1}^{d} d\Big(g^{-1}(i), g^{-1}(i+1 \bmod d)\Big) \\
&= 2\sum_{i=1}^{d} d\Big(g(i), g(i+1 \bmod d)\Big), \qquad (3)
\end{aligned}
$$

where Equation (3) 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 B.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. □

B.2. Proof of Theorem 3.2

Theorem 3.2 (Probabilistic approximation (formal version of Theorem 1.1))**.** *Let $\mathcal{X}$ be a complete metric space and let $\mathcal{P}(\mathcal{X})$ denote the space of Borel probability measures on $\mathcal{X}$. Let G be a compact topological group acting continuously on $\mathcal{X}$, and let α denote its normalized left Haar measure. Suppose that*

$$D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_{\geq 0}$$

is G-invariant, meaning that

$$D(g\mu, g\nu) = D(\mu, \nu), \qquad \forall g \in G,\ \mu, \nu \in \mathcal{P}(\mathcal{X}),$$

and satisfies the one-sided triangle inequality

$$D(\mu, \nu) \leq D(\mu, \eta) + D(\eta, \nu), \qquad \forall \mu, \nu, \eta \in \mathcal{P}(\mathcal{X}).$$

Then, for every $\mu \in \mathcal{P}(\mathcal{X})$,

$$\mathbb{E}_{g\sim\alpha}\left[D(\mu, g\mu)\right] \leq \sup_{g\in G} D(\mu, g\mu) \leq 2\,\mathbb{E}_{g\sim\alpha}\left[D(\mu, g\mu)\right].$$

Proof. Proof. Fix $\mu \in \mathcal{Y} = \mathcal{P}(\mathcal{X})$, and define

$$M_\mu \coloneqq \sup_{h\in G} D(\mu, h\mu).$$

The lower bound is immediate, since an average is bounded above by the supremum:

$$\mathbb{E}_{g\sim\alpha} D(\mu, g\mu) \leq \sup_{h\in G} D(\mu, h\mu).$$

We prove the upper bound. Fix an arbitrary $h \in G$. For every $g \in G$, the one-sided triangle inequality gives

$$D(\mu, h\mu) \leq D(\mu, g\mu) + D(g\mu, h\mu).$$

Using G-invariance of D, the second term can be rewritten as

$$D(g\mu, h\mu) = D(\mu, g^{-1}h\mu).$$

Therefore, for every $g \in G$,

$$D(\mu, h\mu) \leq D(\mu, g\mu) + D(\mu, g^{-1}h\mu).$$

Averaging both sides with respect to the normalized Haar measure α, we obtain

$$D(\mu, h\mu) \leq \mathbb{E}_{g\sim\alpha} D(\mu, g\mu) + \mathbb{E}_{g\sim\alpha} D(\mu, g^{-1}h\mu).$$

Since the map $g \mapsto g^{-1}h$ preserves Haar measure on the compact group G, the second expectation equals the first:

$$\mathbb{E}_{g\sim\alpha} D(\mu, g^{-1}h\mu) = \mathbb{E}_{g\sim\alpha} D(\mu, g\mu).$$

Hence

$$D(\mu, h\mu) \leq 2\,\mathbb{E}_{g\sim\alpha} D(\mu, g\mu).$$

Since $h \in G$ was arbitrary, taking the supremum over h gives

$$\sup_{h \in G} D(\mu, h\mu) \leq 2\, \mathbb{E}_{g \sim \alpha} D(\mu, g\mu).$$

Combining this with the lower bound yields

$$\mathbb{E}_{g \sim \alpha} D(\mu, g\mu) \leq \sup_{h \in G} D(\mu, h\mu) \leq 2\, \mathbb{E}_{g \sim \alpha} D(\mu, g\mu).$$

This completes the proof. □

B.3. An Alternative Second-Order Proof of Theorem 3.2

We provide an alternative proof of Theorem 3.2 using a second-order probability construction. Although the direct proof in the main text is shorter, the second-order viewpoint is useful because it interprets the average asymmetry as a Wasserstein distance between a measure and its orbit law.

Recall that $\mathcal{X}$ is a compact metric space and that $\mathcal{P}(\mathcal{X})$ denotes the space of Borel probability measures on $\mathcal{X}$. Let

$$\mathcal{Y} := \mathcal{P}(\mathcal{X}), \qquad \mathcal{Z} := \mathcal{P}(\mathcal{Y}) = \mathcal{P}(\mathcal{P}(\mathcal{X})).$$

Thus, elements of $\mathcal{Y}$ are probability measures on $\mathcal{X}$, while elements of $\mathcal{Z}$ are probability measures over probability measures. We use nonbold symbols, such as μ, ν, for elements of $\mathcal{Y}$, and bold symbols, such as $\boldsymbol{\mu}, \boldsymbol{\nu}$, for elements of $\mathcal{Z}$.

Let

$$D : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$$

be the G-invariant pseudometric from Theorem 3.2. We define its Wasserstein lift to $\mathcal{Z}$ by

$$\mathbf{W}_D(\boldsymbol{\mu}, \boldsymbol{\nu}) := \inf_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{\mathcal{Y} \times \mathcal{Y}} D(\eta, \xi)\, d\boldsymbol{\pi}(\eta, \xi),$$

where $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ denotes the set of couplings of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. In particular, for Dirac masses $\boldsymbol{\delta}_\mu, \boldsymbol{\delta}_\nu \in \mathcal{Z}$,

$$\mathbf{W}_D(\boldsymbol{\delta}_\mu, \boldsymbol{\delta}_\nu) = D(\mu, \nu).$$

Thus, $\mathbf{W}_D$ extends the original pseudometric D from $\mathcal{Y}$ to the second-order space $\mathcal{Z}$.

The point of this construction is that the orbit of μ can be represented as a distribution over measures. Let α be the normalized Haar measure on G, and define the second-order orbit measure

$$\boldsymbol{\mu}^G := \int_G \boldsymbol{\delta}_{g\mu}\, d\alpha(g) \in \mathcal{Z}.$$

Equivalently, $\boldsymbol{\mu}^G$ is the pushforward of Haar measure under the orbit map $g \mapsto g\mu$. This should not be confused with the ordinary averaged measure

$$\int_G g\mu\, d\alpha(g) \in \mathcal{Y}.$$

The latter collapses the orbit into a single mixture, while $\boldsymbol{\mu}^G$ retains the full distribution of orbit points in $\mathcal{Y}$.

We first record the convexity property of the Wasserstein lift. Although D itself need not be convex as a function of its measure arguments, $\mathbf{W}_D$ is convex in each argument. For example, for $a \in [0, 1]$ and $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\nu} \in \mathcal{Z}$,

$$\begin{aligned}\mathbf{W}_D\big(a\boldsymbol{\mu}_1 + (1-a)\boldsymbol{\mu}_2, \boldsymbol{\nu}\big) \leq\ & a\, \mathbf{W}_D(\boldsymbol{\mu}_1, \boldsymbol{\nu}) \\ & + (1-a)\, \mathbf{W}_D(\boldsymbol{\mu}_2, \boldsymbol{\nu}).\end{aligned}$$

Indeed, if $\boldsymbol{\pi}_i \in \Pi(\boldsymbol{\mu}_i, \boldsymbol{\nu})$ for $i = 1, 2$, then $a\boldsymbol{\pi}_1 + (1-a)\boldsymbol{\pi}_2$ is a coupling of $a\boldsymbol{\mu}_1 + (1-a)\boldsymbol{\mu}_2$ and $\boldsymbol{\nu}$. Taking the transportation cost of this coupling and then infimizing over $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2$ gives the claim. The same argument gives convexity in the second argument.

We now prove the comparison. Since D is G-invariant, the lifted distance $\mathbf{W}_D$ is invariant under the induced action on $\mathcal{Z}$. Namely, if $g_{\#}\boldsymbol{\nu}$ denotes the pushforward of $\boldsymbol{\nu} \in \mathcal{Z}$ under the map $\eta \mapsto g\eta$, then

$$\mathbf{W}_D(g_{\#}\boldsymbol{\mu}, g_{\#}\boldsymbol{\nu}) = \mathbf{W}_D(\boldsymbol{\mu}, \boldsymbol{\nu}).$$

Moreover, Haar invariance implies that

$$h_{\#}\boldsymbol{\mu}^G = \boldsymbol{\mu}^G, \qquad \forall h \in G.$$

Fix $h \in G$. Since $\mathbf{W}_D$ extends D on Dirac masses, and using the triangle inequality for $\mathbf{W}_D$,

$$\begin{aligned} D(\mu, h\mu) &= \mathbf{W}_D(\boldsymbol{\delta}_\mu, \boldsymbol{\delta}_{h\mu}) \\ &\leq \mathbf{W}_D(\boldsymbol{\delta}_\mu, \boldsymbol{\mu}^G) + \mathbf{W}_D(\boldsymbol{\mu}^G, \boldsymbol{\delta}_{h\mu}). \end{aligned}$$

The second term equals the first. Indeed, applying h^{-1} to both arguments and using invariance of $\mathbf{W}_D$, we get

$$\begin{aligned} \mathbf{W}_D(\boldsymbol{\mu}^G, \boldsymbol{\delta}_{h\mu}) &= \mathbf{W}_D\big((h^{-1})_{\#}\boldsymbol{\mu}^G, (h^{-1})_{\#}\boldsymbol{\delta}_{h\mu}\big) \\ &= \mathbf{W}_D(\boldsymbol{\mu}^G, \boldsymbol{\delta}_\mu) \\ &= \mathbf{W}_D(\boldsymbol{\delta}_\mu, \boldsymbol{\mu}^G), \end{aligned}$$

where the last equality uses symmetry of $\mathbf{W}_D$. Hence

$$D(\mu, h\mu) \leq 2\,\mathbf{W}_D(\boldsymbol{\delta}_\mu, \boldsymbol{\mu}^G).$$

Finally, by convexity of the Wasserstein lift in the second argument,

$$\begin{aligned} \mathbf{W}_D(\boldsymbol{\delta}_\mu, \boldsymbol{\mu}^G) &= \mathbf{W}_D\left(\boldsymbol{\delta}_\mu, \int_G \boldsymbol{\delta}_{g\mu}\, d\alpha(g)\right) \\ &\leq \int_G \mathbf{W}_D(\boldsymbol{\delta}_\mu, \boldsymbol{\delta}_{g\mu})\, d\alpha(g) \\ &= \int_G D(\mu, g\mu)\, d\alpha(g). \end{aligned}$$

Therefore, for every $h \in G$,

$$D(\mu, h\mu) \leq 2 \int_G D(\mu, g\mu)\, d\alpha(g).$$

Taking the supremum over $h \in G$ gives

$$\sup_{h \in G} D(\mu, h\mu) \leq 2 \int_G D(\mu, g\mu)\, d\alpha(g).$$

The reverse inequality follows because an average is bounded above by the supremum:

$$\int_G D(\mu, g\mu)\, d\alpha(g) \leq \sup_{g \in G} D(\mu, g\mu).$$

Combining the two inequalities completes the alternative proof of Theorem 3.2.

Interpretation of the second-order construction. The second-order proof gives a useful geometric interpretation of the average asymmetry. Since the first argument is a Dirac mass, the coupling between $\boldsymbol{\delta}_\mu$ and $\boldsymbol{\mu}^G$ is unique, and hence

$$\mathbf{W}_D(\boldsymbol{\delta}_\mu, \boldsymbol{\mu}^G) = \int_G D(\mu, g\mu)\, d\alpha(g).$$

Thus the average displacement appearing in Theorem 3.2 is exactly the Wasserstein distance, in the second-order space, between the point mass at μ and the Haar orbit law of μ. This perspective distinguishes the orbit law $\boldsymbol{\mu}^G$ from the ordinary invariant mixture $\int_G g\mu \, d\alpha(g)$, which may lose information about the variability of the orbit. It is therefore useful when studying finite-sample orbit approximations: if $g_1, \ldots, g_m \sim \alpha$, the empirical orbit law

$$\widehat{\boldsymbol{\mu}}_m = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\delta}_{g_i \mu}$$

is a natural approximation of $\boldsymbol{\mu}^G$ in $\mathcal{P}(\mathcal{P}(\mathcal{X}))$. In this sense, the second-order formulation is not needed for the scalar factor-two comparison, but it provides a convex geometric framework for reasoning about distributions over orbit points.

B.4. Proof of Theorem 3.5

Theorem 3.5 (Half-probability certificate for distance to asymmetry)**.** *Let G be a compact group with the normalized Haar measure α. Let $\mathcal{X}$ be a complete metric space and let $\mathcal{Y} = \mathcal{P}(\mathcal{X})$. Suppose G acts continuously on $\mathcal{Y}$, and let $D : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a G-invariant pseudometric, meaning that*

$$D(g\mu, g\nu) = D(\mu, \nu)$$

for all $g \in G$ and $\mu, \nu \in \mathcal{Y}$. Then, for every $\mu \in \mathcal{Y}$,

$$\mathbb{P}_{g \sim \alpha} \left(D(\mu, g\mu) \geq \frac{1}{2} \sup_{h \in G} D(\mu, h\mu) \right) \geq \frac{1}{2}.$$

Proof. Define

$$\Delta_\mu(g) \coloneqq D(\mu, g\mu), \qquad M_\mu \coloneqq \sup_{h \in G} D(\mu, h\mu).$$

We first record a simple subadditivity property of Δ_μ. For any $g_1, g_2 \in G$, the triangle inequality gives

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

By G-invariance of D,

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

Therefore

$$\Delta_\mu(g_1 g_2) \leq \Delta_\mu(g_1) + \Delta_\mu(g_2).$$

Let $h^\star \in G$ be such that

$$\Delta_\mu(h^\star) = M_\mu.$$

For any $g \in G$, apply the subadditivity bound with $g_1 = g$ and $g_2 = g^{-1} h^\star$. Since $g g^{-1} h^\star = h^\star$, we get

$$M_\mu = \Delta_\mu(h^\star) \leq \Delta_\mu(g) + \Delta_\mu(g^{-1} h^\star).$$

Hence, for every $g \in G$, at least one of the two quantities $\Delta_\mu(g), \Delta_\mu(g^{-1} h^\star)$ is at least $M_\mu/2$.

Now define the good set

$$A \coloneqq \left\{ g \in G : \Delta_\mu(g) \geq \frac{1}{2} M_\mu \right\}.$$

The previous argument shows that for every $g \in G$,

$$g \in A \qquad \text{or} \qquad g^{-1} h^\star \in A.$$

Equivalently,

$$G = A \cup \{g \in G : g^{-1}h^\star \in A\}.$$

By Haar invariance, the map $g \mapsto g^{-1}h^\star$ preserves the Haar measure α. Therefore

$$\alpha\left(\{g \in G : g^{-1}h^\star \in A\}\right) = \alpha(A).$$

Thus

$$1 = \alpha(G) \leq \alpha(A) + \alpha\left(\{g \in G : g^{-1}h^\star \in A\}\right) = 2\alpha(A).$$

Hence

$$\alpha(A) \geq \frac{1}{2}.$$

Substituting back the definition of A, we obtain

$$\alpha\left(\left\{g \in G : D(\mu, g\mu) \geq \frac{1}{2}M_\mu\right\}\right) \geq \frac{1}{2}.$$

This completes the proof. □

B.5. Proof of Proposition 3.7

Proof. Define the powered cost

$$\rho(\eta, \xi) \coloneqq D(\eta, \xi)^p.$$

Then ρ is G-invariant and satisfies the relaxed triangle inequality

$$\rho(\eta, \xi) \leq B\big(\rho(\eta, \zeta) + \rho(\zeta, \xi)\big).$$

Let $\mathbf{W}_\rho$ denote the Wasserstein lift of ρ to the second-order space $\mathcal{Z} = \mathcal{P}(\mathcal{Y})$. Fix $\mu \in \mathcal{Y}$, and define the second-order orbit measure

$$\boldsymbol{\mu}^G \coloneqq \int_G \boldsymbol{\delta}_{g\mu}\, d\alpha(g) \in \mathcal{Z}.$$

For any $h \in G$, the lifted relaxed triangle inequality gives

$$\begin{aligned} D(\mu, h\mu)^p &= \mathbf{W}_\rho(\boldsymbol{\delta}_\mu, \boldsymbol{\delta}_{h\mu}) \\ &\leq B\Big(\mathbf{W}_\rho(\boldsymbol{\delta}_\mu, \boldsymbol{\mu}^G) + \mathbf{W}_\rho(\boldsymbol{\mu}^G, \boldsymbol{\delta}_{h\mu})\Big). \end{aligned}$$

By G-invariance of ρ, the Wasserstein lift $\mathbf{W}_\rho$ is also G-invariant. Moreover, by Haar invariance, the second-order orbit measure $\boldsymbol{\mu}^G$ is invariant under the induced action of G. Thus

$$\mathbf{W}_\rho(\boldsymbol{\mu}^G, \boldsymbol{\delta}_{h\mu}) = \mathbf{W}_\rho(\boldsymbol{\mu}^G, \boldsymbol{\delta}_\mu) = \mathbf{W}_\rho(\boldsymbol{\delta}_\mu, \boldsymbol{\mu}^G),$$

where the last equality uses symmetry. Hence

$$D(\mu, h\mu)^p \leq 2B\, \mathbf{W}_\rho(\boldsymbol{\delta}_\mu, \boldsymbol{\mu}^G).$$

By convexity of the Wasserstein lift in its second argument,

$$\begin{aligned} \mathbf{W}_\rho(\boldsymbol{\delta}_\mu, \boldsymbol{\mu}^G) &= \mathbf{W}_\rho\left(\boldsymbol{\delta}_\mu, \int_G \boldsymbol{\delta}_{g\mu}\, d\alpha(g)\right) \\ &\leq \int_G \mathbf{W}_\rho(\boldsymbol{\delta}_\mu, \boldsymbol{\delta}_{g\mu})\, d\alpha(g) \\ &= \mathbb{E}_{g\sim\alpha}\left[D(\mu, g\mu)^p\right]. \end{aligned}$$

Therefore, for every $h \in G$,

$$D(\mu, h\mu)^p \leq 2B\, \mathbb{E}_{g\sim\alpha}\left[D(\mu, g\mu)^p\right].$$

Taking the supremum over $h \in G$ gives

$$\sup_{h\in G} D(\mu, h\mu)^p \leq 2B\, \mathbb{E}_{g\sim\alpha}\left[D(\mu, g\mu)^p\right].$$

Equivalently,

$$\sup_{h\in G} D(\mu, h\mu) \leq (2B)^{1/p} \left(\mathbb{E}_{g\sim\alpha}\left[D(\mu, g\mu)^p\right]\right)^{1/p}.$$

The reverse inequality in the L^p-average comparison is immediate because $D(\mu, g\mu) \leq \sup_{h\in G} D(\mu, h\mu)$ for every $g \in G$. Hence

$$\left(\mathbb{E}_{g\sim\alpha}\left[D(\mu, g\mu)^p\right]\right)^{1/p} \leq \sup_{h\in G} D(\mu, h\mu).$$

This proves the first two-sided comparison.

It remains to prove the ordinary-average comparison when $p \geq 1$. Let

$$M_\mu := \sup_{h\in G} D(\mu, h\mu).$$

If $M_\mu = 0$, then the claim is immediate. Otherwise, from the L^p comparison above,

$$M_\mu^p \leq 2B\, \mathbb{E}_{g\sim\alpha}\left[D(\mu, g\mu)^p\right].$$

Since $D(\mu, g\mu) \leq M_\mu$ for every $g \in G$ and $p \geq 1$, we have

$$D(\mu, g\mu)^p \leq M_\mu^{p-1} D(\mu, g\mu).$$

Therefore,

$$M_\mu^p \leq 2B M_\mu^{p-1} \mathbb{E}_{g\sim\alpha}\left[D(\mu, g\mu)\right].$$

Dividing by $M_\mu^{p-1} > 0$ yields

$$M_\mu \leq 2B\, \mathbb{E}_{g\sim\alpha}\left[D(\mu, g\mu)\right].$$

The lower bound

$$\mathbb{E}_{g\sim\alpha}\left[D(\mu, g\mu)\right] \leq M_\mu$$

again follows because an average is bounded above by the supremum. Combining the two inequalities gives

$$\mathbb{E}_{g\sim\alpha}\left[D(\mu, g\mu)\right] \leq \sup_{h\in G} D(\mu, h\mu) \leq 2B\, \mathbb{E}_{g\sim\alpha}\left[D(\mu, g\mu)\right].$$

This completes the proof. □

B.6. Proof of Proposition 3.8

Proof. The lower bound

$$\mathbb{E}_{g\sim\alpha} D(\mu, g\nu) \leq \sup_{h\in G} D(\mu, h\nu)$$

is immediate, since an average is bounded above by the supremum.

We prove the upper bound. Fix $h \in G$. For every $g \in G$, the triangle inequality gives

$$D(\mu, h\nu) \leq D(\mu, g\nu) + D(g\nu, h\nu).$$

By G-invariance of D,

$$D(g\nu, h\nu) = D(\nu, g^{-1}h\nu).$$

Averaging over $g \sim \alpha$, and using the fact that the map $g \mapsto g^{-1}h$ preserves Haar measure, we obtain

$$D(\mu, h\nu) \leq \mathbb{E}_{g\sim\alpha} D(\mu, g\nu) + \mathbb{E}_{g\sim\alpha} D(\nu, g\nu).$$

It remains to control the second term by the cross-orbit average. Fix $a \in G$. For every $g \in G$, the triangle inequality gives

$$D(\nu, a\nu) \leq D(\nu, g^{-1}\mu) + D(g^{-1}\mu, a\nu).$$

Using symmetry and G-invariance of D, the first term satisfies

$$D(\nu, g^{-1}\mu) = D(g\nu, \mu) = D(\mu, g\nu),$$

and the second term satisfies

$$D(g^{-1}\mu, a\nu) = D(\mu, ga\nu).$$

Therefore,

$$D(\nu, a\nu) \leq D(\mu, g\nu) + D(\mu, ga\nu).$$

Averaging over $g \sim \alpha$, and using Haar invariance again, gives

$$D(\nu, a\nu) \leq 2\,\mathbb{E}_{g\sim\alpha} D(\mu, g\nu).$$

Finally, averaging this inequality over $a \sim \alpha$, we get

$$\mathbb{E}_{a\sim\alpha} D(\nu, a\nu) \leq 2\,\mathbb{E}_{g\sim\alpha} D(\mu, g\nu).$$

Substituting this into the previous bound on $D(\mu, h\nu)$, we obtain

$$D(\mu, h\nu) \leq 3\,\mathbb{E}_{g\sim\alpha} D(\mu, g\nu).$$

Since $h \in G$ was arbitrary, taking the supremum over $h \in G$ yields

$$\sup_{h\in G} D(\mu, h\nu) \leq 3\,\mathbb{E}_{g\sim\alpha} D(\mu, g\nu).$$

Combining this with the lower bound completes the proof. □

□