Two-sample goodness-of-fit tests on the flat torus based on Wasserstein distance and their relevance to structural biology

Javier González-Delgado, Alberto González-Sanz, Juan Cortés, Pierre Neuvial

Two-sample goodness-of-fit tests on the flat torus based on Wasserstein distance and their relevance to structural biology

Javier González-Delgado[1,2*‡], **Alberto González-Sanz**[1,3*†], **Juan Cortés**[2†] **and Pierre Neuvial**[1]

e-mail: javier.gonzalez-delgado@math.univ-toulouse.fr

e-mail: alberto.gonzalez_sanz@math.univ-toulouse.fr

e-mail: juan.cortes@laas.fr

e-mail: pierre.neuvial@math.univ-toulouse.fr

[1] *Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS.*

[2] *LAAS-CNRS, Université de Toulouse, CNRS.*

[3] *ImUva, Universidad de Valladolid.*

Abstract: This work is motivated by the study of local protein structure, which is defined by two variable dihedral angles that take values from probability distributions on the flat torus. Our goal is to provide the space $\mathcal{P}(\mathbb{R}^2/\mathbb{Z}^2)$ with a metric that quantifies local structural modifications due to changes in the protein sequence, and to define associated two-sample goodness-of-fit testing approaches. Due to its adaptability to the space geometry, we focus on the Wasserstein distance as a metric between distributions.

We extend existing results of the theory of Optimal Transport to the d-dimensional flat torus $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$, in particular a Central Limit Theorem. Moreover, we assess different techniques for two-sample goodness-of-fit testing for the two-dimensional case, based on the Wasserstein distance. We provide an implentation of these approaches in R. Their performance is illustrated by numerical experiments on synthetic data and protein structure data.

Keywords and phrases: Optimal Transport, Flat Torus, Wasserstein distance, Central Limit Theorem, Goodness-of-fit test, Structural biology, Intrinsically disordered proteins.

Contents

1 Introduction 2
2 Optimal transport in $\mathbb{R}^d/\mathbb{Z}^d$ 4

*Equal contribution.

†Research partially supported by the AI Interdisciplinary Institute ANITI, which is funded by the French "Investing for the Future – PIA3" program under the Grant agreement ANR-19-PI3A-0004.

‡Research supported by the ANR LabEx CIMI (grant ANR-11-LABX-0040) within the French State Programme "Investissements d'Avenir".

2.1 Description of the solutions . . . 5
2.2 Asymptotic behaviour . . . 7
2.3 Asymptotic normality . . . 7
3 Two-sample goodness-of-fit tests . . . 8
3.1 Marginal projections into $\mathbb{R}/\mathbb{Z}$ . . . 9
3.2 p-value upper bounding . . . 9
3.3 Asymptotic behaviour under the alternative . . . 11
4 Simulations and results . . . 11
4.1 Simulated data . . . 12
4.2 Protein structure data . . . 13
5 Summary . . . 15
6 Discussion . . . 15
Code availability . . . 16
A Proofs . . . 16
A.1 Proofs of Section 2 . . . 16
A.2 Proofs of Section 3 . . . 20
Supplementary Material . . . 23
References . . . 25

1. Introduction

When it comes to measure the distance between two probability distributions, the well known Wasserstein distance, derived from the theory of Optimal Transport (OT), provides both strong theoretical guarantees –it metrizes weak convergence [35]– and attractive empirical performance [21]. Most of the applications of such theory are related to the very active field of machine learning, notably in the framework of generative networks [2], robustness [30] or fairness [9], among others.

From a statistical point of view, one of the main caveats of the theory of OT comes from the *curse of dimensionality*: the rate of convergence of the empirical Wasserstein distance decreases as $n^{-1/d}$ with the dimension [12]. Another important issue is the asymptotic behavior of the fluctuations of the empirical optimal transport cost. For probabilities supported in $\mathbb{R}^d$, it has been proved, using Efron–Stein's inequality that, for the cost L^2, the difference $\sqrt{n}(\mathcal{W}_2^2(P_n, Q) - \mathbb{E}\mathcal{W}_2^2(P_n, Q))$ is asymptotically Gaussian [10]. Recently, the proofs have been extended to some general costs in $\mathbb{R}^d$, including the cost L^p, for $p > 1$ [8]. Concerning statistical goodness-of-fit tests based on Wasserstein distance, the one sample case has already been addressed in [14] and, when the probabilities are defined over $\mathbb{R}$, two-sample tests can be derived from [7].

In this paper, we focus on the d-dimensional flat torus $\mathbb{T}^d := \mathbb{R}^d/\mathbb{Z}^d$ where, even from the purely theoretical point of view, OT has not been explicitly addressed. However, this space appears naturally when the probability measures are periodic (e.g. for distributions of angles). The main objective of this work is to extend existing OT results to the space of probability measures on the flat

torus $\mathcal{P}(\mathbb{T}^d)$, specially a Central Limit Theorem (CLT), and to address in particular the two-dimensional case, constructing some two-sample goodness-of-fit testing techniques based on Wasserstein distance.

Our motivation for extending the theory of OT to $\mathbb{T}^2$ comes from the investigation of proteins. Understanding the relationships between protein sequence, structure and function is the main goal of Structural Biology. In addition to its scientific importance, a better understanding of these relationships is essential for applications in diverse areas, such as biomedicine and biotechnology. The conformational state of a protein can be defined by a vector of angles, corresponding to rotations around the chemical bonds between the atoms that constitute its "backbone". This vector contains two values per amino-acid, ϕ and ψ, which follow a certain distribution, and which are usually represented using the so-called Ramachandran plots [24] (see also Figure 5). The analysis of these distributions has several important applications, such as the validation or refinement of protein structures determined from biophysical techniques [19, 16], the prediction of some biophysical measurements to complement experiments [31], and the development of potential energy models or scoring methods for protein structure modeling, prediction and design [4, 26, 33].

In this context, the definition of a suitable distance between distributions on $\mathbb{T}^2$ is essential. This would allow to quantify the expected magnitude of structural effects associated with local changes in the sequence, and therefore to develop improved versions of the aforementioned modeling and prediction techniques. Nevertheless, this has not been done satisfactorily in previous works. For example, in [26] and [31] significant differences between two laws are stated after visual comparison of two empirical distributions, and in [33] the Hellinger distance is used to compare distributions on a non-periodic $[-\pi, \pi] \times [-\pi, \pi]$. Efficient statistical tests remain to be defined and implemented in order to state such differences, being based on a metric that takes geometry into consideration. As many other commonly-used metrics, Hellinger distance ignores the underlying geometry of the space. Here, we propose to use the Wasserstein distance, whose advantageous geometrical and mathematical properties are described in [21], [34] and [35], to define goodness-of-fit testing techniques for two measures on $\mathbb{T}^2$, allowing a more accurate study of protein local conformation distributions.

The paper is organized as follows:

- Section 2 starts by introducing the general framework of measures on the flat torus in general dimension, followed by the precise formulation of the optimal transport problem. Here we will set the notations used through the paper. It is divided into different subsections: Subsection 2.1 is devoted to the study of the shape of the solutions, recalling that they are the gradient of a periodic convex functions and showing the uniqueness of the potential in Corollary 2.2. Section 2.2 proves through Theorem 2.4 that the optimal transport potentials converge, up to additive constant, when the measures converge weakly. This result implies that the method of [10]

based on Efron–Stein's inequality can be applied to derive a Central Limit Theorem, see Theorem 2.5 in Subsection 2.3.

- Section 3 shows how this theory can be applied to perform two-sample goodness-of-fit tests in the two dimensional flat torus. The main result of the section is Theorem 3.1, which gives a concentration inequality primordial for the construction of statistical tests, together with faster convergence rates for the expectation.
- Section 4 reports numerical experiments supporting these theoretical results, first with synthetic data and then with real data from protein structures, showing that our methods behave well in both cases.

To facilitate reading, the proofs are relegated to the Appendix, but in some cases the intuitions of the proofs are provided in the main text for clarity.

2. Optimal transport in $\mathbb{R}^d/\mathbb{Z}^d$

Let $\mathbb{T}^d := \mathbb{R}^d/\mathbb{Z}^d$ be defined as the quotient space derived from the equivalence relation $x\mathcal{R}y$ if $x-y \in \mathbb{Z}^d$. For each $x \in \mathbb{R}^d$ we denote as $\bar{x} \in \mathbb{T}^d$ its equivalence class and reserve the notation τ for the canonical projection map $x \mapsto \tau(x) = \bar{x}$. The topology of the quotient space is defined as the finest one that makes τ continuous. With this topology the space $\mathbb{T}^d$ is a Polish space with the distance derived from the Euclidean norm $||\cdot||$,

$$d(\bar{x},\bar{y}) := \inf_{p\in\mathbb{Z}} ||x-y+p||.$$

Note that the last claim is true since the projection map τ is in fact a metric identification, $(\mathbb{R}^d, ||\cdot||)$ is a Banach space and $\mathbb{Z}^d$ is a closed subspace, then it is complete, metrizable through d and separable.

For two probability measures $P, Q \in \mathcal{P}(\mathbb{T}^d)$, a probability measure $\pi \in \mathcal{P}(\mathbb{T}^d \times \mathbb{T}^d)$ is said to be an *optimal transport plan for the cost* d^2 between P and Q if it solves

$$\mathcal{T}_2(P,Q) := \inf_{\gamma\in\Pi(P,Q)} \int_{\mathbb{T}^d\times\mathbb{T}^d} d^2(\bar{x},\bar{y}) d\gamma(\bar{x},\bar{y}), \tag{2.1}$$

where $\Pi(P,Q)$ is the set of probability measures $\gamma \in \mathcal{P}(\mathbb{T}^d \times \mathbb{T}^d)$ such that $\gamma(A\times\mathbb{R}^d) = P(A)$ and $\gamma(\mathbb{T}^d \times B) = Q(B)$ for all A, B measurable subsets of $\mathbb{T}^d$.

The Kantorovich problem (2.1) can be formulated in a dual form, as follows

$$\mathcal{T}_2(P,Q) = \sup_{(f,g)\in\Phi_2(P,Q)} \int_{\mathbb{T}^d} f(\bar{x}) dP(\bar{x}) + \int_{\mathbb{T}^d} g(\bar{y}) dQ(\bar{y}), \tag{2.2}$$

where

$$\Phi_2(P,Q) = \{(f,g) \in L_1(P)\times L_1(Q) :\ f(\bar{x}) + g(\bar{y}) \le d^2(\bar{x},\bar{y})\}.$$

$\psi \in L_1(P)$ is said to be an *optimal transport potential from P to Q for the cost d^2* if there exists $\varphi \in L_1(Q)$ such that the pair (ψ, φ) solves (2.2). Recall from [35] that the solutions of (2.2) are pairs (f, f^{d^2}) of d^2 conjugate d^2-concave functions. This means that there exists two sets $\mathcal{Y} \subset \mathbb{T}^d$ and $\Lambda \subset \mathbb{R}$ such that

$$f(x) = \inf_{y \in \mathcal{Y},\ \alpha \in \Lambda} \{d(x,y)^2 - \alpha\} \text{ and } \ f^{d^2}(y) = \inf_{x \in \mathbb{T}^d} \{d(x,y)^2 - f(x)\}.$$

Moreover, since $\mathbb{T}^d$ is a Polish space, then Theorem 4.1 in [35] implies that there exists a solution π^* of (2.1). Additionally Theorem 5.10 in [35] establishes that $\text{supp}(\pi^*)$ is d^2-*cyclically monotone*. This means that for every finite sequence $\{(x_k, y_k)\}_{k=1}^n \subset \text{supp}(\pi^*)$ and every bijection $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ the following inequality holds:

$$\sum_{k=1}^{n} d^2(x_k, y_k) \leq \sum_{k=1}^{n} d^2(x_k, y_{\sigma(k)}).$$

The concept of d^2-cyclical monotonicity is the generalization, to other spaces and costs, of the concept of cyclical monotonicity in convex analysis, described in [27]. A set $A \subset \mathbb{R}^d \times \mathbb{R}^d$ is *cyclically monotone* if for every finite sequence $\{(x_k, y_k)\}_{k=1}^n \subset A$ and every bijection $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ it holds that

$$\sum_{k=1}^{n} \langle x_k, y_k \rangle \geq \sum_{k=1}^{n} \langle x_k, y_{\sigma(k)} \rangle.$$

In some cases, that we will study latter on, there exists some measurable map T such that optimal transport plan π satisfies $\pi = (I \times T)\,\#P$, where I denotes the identity map. Therefore, the problem becomes equivalent to the following *Monge* formulation:

$$\mathcal{T}_2(P, Q) = \inf_{T \# P = Q} \int_{\mathbb{T}^d} d^2(\bar{x}, T(\bar{x}))\, dP(\bar{x}), \tag{2.3}$$

where the symbol $T\#P$ denotes the *push forward measure* of P through T, which is defined by $T\#P(A) := P(T^{-1}(A))$, for all measurable $A \subset \mathbb{T}^d$. The support of a probability Q is usually defined as the closed set $R_Q \subset \mathbb{T}^d$ composed by $\bar{x} \in \mathbb{T}^d$ such that for any neighborhood $\mathcal{U}_{\bar{x}}$ of $\bar{x}$ it is satisfied that $Q(\mathcal{U}_{\bar{x}}) > 0$. Yet in our case, for convenience, we will consider the interior of R_Q which we denote as

$$\text{supp}(Q) := \text{int}\,(R_Q)\,. \tag{2.4}$$

2.1. Description of the solutions

This section begins by setting the notation we will follow throughout the paper. Then, Theorem 2.1 adapts a result of [6] to claim the existence of solutions of the Monge problem (2.3), which is characterized in Lemma A.2. As a consequence,

Theorem 2.2 guarantees under certain assumptions of regularity on P that the solution of (2.2) is unique up to an additive constant.

Note that in practice a probability $P \in \mathcal{P}(\mathbb{T}^d)$ defines a periodic measure $\mu_P \in \mathcal{M}(\mathbb{R}^d)$ w.r.t. any $p \in \mathbb{Z}^d$. In other words, $T_p \# \mu_P = \mu_P$, for all $p \in \mathbb{Z}^d$, where $T_p : \mathbb{R}^d \to \mathbb{R}^d$ is the shift operator $x \mapsto x + p$. If $\mu_P, \mu_Q \in \mathcal{M}(\mathbb{R}^d)$ are absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^d$, denoted as $\mu_P, \mu_Q \ll \ell_d$, [6] establishes that there exists a convex function φ such that $\nabla\varphi \# \mu_P = \mu_Q$. Theorem 1.25 in [28] entails that there is a unique solution of the Monge problem in the torus, described by the relation $T = x - \nabla f$, where the sum is to be intended modulo $\mathbb{Z}^d$ and f is an optimal transport potential. The next Theorem proves that in fact such a map T satisfies $T \circ \tau = \tau \circ \nabla\varphi$, where φ is defined some lines above. This intrinsic characterization of the optimal transport map is at least surprising, meaning that we can relate the optimal transport map in the torus with the unique gradient of a convex function pushing forward their respective periodic measures. The proof starts by realizing that since $\mathbb{T}^d$ is a Polish space, then Theorem 4.1 in [35] implies that there exists a solution π^* of (2.1). Furthermore, Theorem 5.10 in [35] establishes that $\text{supp}(\pi^*)$ is d^2-cyclically monotone, which implies that, by Proposition 2 in [6], the set

$$\Gamma = \{(x+p, y+p) : \ (\bar{x}, \bar{y}) \in \text{supp}(\pi^*), \ x \in [0,1]^d, \ d(\bar{x}, \bar{y}) = ||x-y|| \text{ and } p \in \mathbb{Z}^d\} \tag{2.5}$$

is cyclically monotone. Since every cyclically monotone set is contained in the subdifferential of a convex function (Theorem 12.25 in [22]), then we have a candidate of convex function that plays the main role in the following theorem.

Theorem 2.1. *Let $P, Q \in \mathcal{P}(\mathbb{T}^d)$ be probability measures such that their associated periodic measures satisfy $\mu_P, \mu_Q \ll \ell_d$. Then there exists an unique solution T of (2.3). Moreover, there exists an unique convex function $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ such that*

- *the relation $T \circ \tau = \tau \circ (\nabla\varphi)$ holds μ_P-almost surely,*
- *and $\nabla\varphi \# \mu_P = \mu_Q$.*

The following result gives the uniqueness, up to additive constants, of the optimal transport potential. The proof investigates the intrinsic relation between the optimal transport potentials and the previously described φ, which serves to use general results for convex functions which have the same gradient a.e. in a connected domain of R^d.

Theorem 2.2. *Let $P, Q \in \mathcal{P}(\mathbb{T}^d)$ be probability measures such that their associated periodic measures satisfy $\mu_P, \mu_Q \ll \ell_d$. Then there exists an unique, up to an additive constant, d^2-concave function f solution of (2.2).*

The importance of Theorem 2.2 mainly lies on the study of the asymptotic behavior of the potential, allowing us to apply Arzelá-Ascoli like reasoning.

2.2. Asymptotic behaviour

This section deals with the asymptotic properties of the transport map and potentials. We consider two sequences of probabilities $\{\alpha_n\}_{n\in\mathbb{N}}, \{\beta_n\}_{n\in\mathbb{N}} \subset \mathcal{P}(\mathbb{T}^d)$ convergering weakly to P and Q respectively,

$$\alpha_n \xrightarrow{w} P \quad \text{and} \quad \beta_n \xrightarrow{w} Q.$$

Here the weak convergence is in the sense that $\int h(\bar{x})d\alpha_n(\bar{x}) \to \int h(\bar{x})dP(\bar{x})$, for every continuous function $h \in \mathcal{C}(\mathbb{T}^d)$. Note that the classic definition of weak convergence of probabilities involves only continuous functions with compact support, but the compactness of $\mathcal{T}$ allows us to relax that hypothesis. Once again thanks to that compactness the existence of moments of order 2 is always fulfilled by any $P \in \mathcal{P}(\mathbb{T}^d)$. In consequence Theorem 7.12 in [34] implies that $\alpha_n \xrightarrow{w} P$ if and only if the *quadratic Wasserstein distance* $\mathcal{W}_2(\alpha_n, P) := \sqrt{\mathcal{T}_2(\alpha_n, P)}$ tends to 0.

The idea of this section is to take advantage that any d^2-concave function f is continuous whereby it is finite. Moreover, it has bounded continuity modulus, so we can apply Arzelá-Ascoli's Theorem by fixing the constants.

Lemma 2.3. *Every d^2-concave function f is Lipchitz with constant 2 and with respect to the metric d.*

The proof of the next Theorem firstly proceeds by choosing the sequence $\{a_n\}_{n\in\mathbb{N}}$ to guarantee the uniform boundedness of the sequence $\{(f_n, g_n)\}_{n\in\mathbb{N}}$ of solutions of (2.2). This, together with Lemma 2.3 and Arzelá-Ascoli's Theorem, implies that $\{(f_n, g_n)\}_{n\in\mathbb{N}}$ is relatively compact. The uniqueness of solutions of (2.2), described in Theorem 2.2, allows us to conclude.

Theorem 2.4. *Let $P, Q \in \mathcal{P}(\mathbb{T}^d)$ be probabilities with connected supports such that their associated periodic measures satisfy $\mu_P, \mu_Q \ll \ell_d$. Let $\{\alpha_n\}_{n\in\mathbb{N}}$ and $\{\beta_n\}_{n\in\mathbb{N}} \subset \mathcal{P}(\mathbb{T}^d)$ be two sequences of probabilities converging weakly to P and Q respectively. Denote by (f_n, g_n) (resp. (f, g)) the solution of the dual problem between α_n and β_n (resp. P and Q). Then there exists a sequence of real numbers $\{a_n\}_{n\in\mathbb{N}}$ such that $f_n + a_n \to f$ uniformly on $\mathbb{T}^d$.*

2.3. Asymptotic normality

This section is devoted to prove a Central Limit Theorem (CLT) for the fluctuations of the empirical optimal transport cost. Recall that the previous section proves that, under certain regularity assumptions, there exists a unique optimal transport potential from P to Q. Let φ be such a potential. We will use Efron-Stein's inequality to derive that

$$\sqrt{n}\left(\mathcal{T}_p(P_n, Q) - \mathbb{E}\mathcal{T}_p(P_n, Q)\right) \xrightarrow{w} N(0, \sigma_p^2(P, Q)),$$

with

$$\sigma^2(P, Q) = \mathrm{Var}(\varphi(X)), \tag{2.6}$$

where φ is a transport potential from P to Q and $X \sim P$. Then we will see that the same holds in the two sample case. The idea is not new: it has already been used with the same goal in [10] for the quadratic cost in $\mathbb{R}^d$, and in its extension to general costs in [8]. Moreover, when using regularized optimal transport, [18] showed that the same technique can be applied.

Theorem 2.5. *Let $P, Q \in \mathcal{P}(\mathbb{T}^d)$ be probabilities with connected supports and negligible boundary such that their associated periodic measures satisfy $\mu_P, \mu_Q \ll \ell_d$. Then*

$$\sqrt{n}\left(\mathcal{T}_p(P_n, Q) - \mathbb{E}\mathcal{T}_p(P_n, Q)\right) \xrightarrow{w} N(0, \sigma^2(P, Q)),$$

and if $\frac{nm}{n+m} \to \lambda \in (0,1)$ as $n, m \to \infty$,

$$\sqrt{\tfrac{nm}{n+m}}\left(\mathcal{T}_p(P_n, Q_m) - \mathbb{E}\mathcal{T}_p(P_n, Q_m)\right) \xrightarrow{w} N\left(0, (1-\lambda)\sigma^2(P, Q) + \lambda\sigma^2(Q, P)\right),$$

with $\sigma^2(P, Q)$ and $\sigma^2(Q, P)$ are defined in (2.6) and satisfy

$$\sqrt{\tfrac{nm}{n+m}}\mathrm{Var}(\mathcal{T}_p(P_n, Q_m)) \longrightarrow (1-\lambda)\sigma^2(P, Q) + \lambda\sigma^2(Q, P). \tag{2.7}$$

3. Two-sample goodness-of-fit tests

Goodness-of-fit testing based in Wasserstein distance is still an open problem. The one-sample case in $\mathbb{R}^d$ has recently been addressed in [14], but approaches for two-sample testing in arbitrary dimension, and for measures on more general spaces, have not already been proposed to the best of our knowledge. The intrinsic difficulty of characterizing the distribution of $\mathcal{W}_p(P_n, Q_m)$ accounts for the lack of solutions, specially when dimension is higher than one. Our aim here is to present some goodness-of-fit testing techniques based on the statistic $\mathcal{T}_p(P_n, Q_m)$, allowing the assessment of the null hypothesis $H_0 : P = Q$, for measures on $\mathbb{T}^2$. All of our approaches are based on the extension of results for measures on $\mathbb{R}^d$. Therefore, they can also be adapted to the Euclidean space of general dimension. If we denote by $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ two simple random samples of laws $P, Q \in \mathcal{P}(\mathbb{T}^2)$ respectively, and P_n, Q_m their corresponding empirical probability measures, we aim to test the hypothesis $H_0 : P = Q$ via the definition of the p-value of the form

$$p = \mathbb{P}_{H_0}(\mathcal{T}_p(P_n, Q_m) \geq t_{nm}), \tag{3.1}$$

where t_{nm} denotes the statistic realization for the given samples.

As already mentioned, knowing the distribution of the statistic under the null remains an open (and maybe unfeasible) problem. Therefore, the three presented approaches are not exact but based on: the projection of the problem to a one-dimesional space, the upper bounding of p-values and the asymptotic behaviour of the statistic under the alternative hypothesis.

3.1. Marginal projections into $\mathbb{R}/\mathbb{Z}$

A first approach when testing the equality of two measures P, Q on $\mathbb{R}^2/\mathbb{Z}^2$ is to test the equality of their corresponding marginals P_x and P_y on the unit circle $\mathbb{R}/\mathbb{Z}$. This bypasses the dimension problem and allows the implementation of testing techniques based on Wasserstein distance for one-dimensional spaces. Optimal Transportation on the circle has been recently studied in detail in [15], where the limit laws of the one and two-sample empirical Wasserstein distance for measures on $\mathbb{R}/\mathbb{Z}$ are derived. However, the considered statistics are not distribution-free and only one-sample goodness-of-fit tests can therefore be performed. Here, we propose a partical alternative to test the equality of two marginal laws supported on the circle.

As shown in [23], given two pairwise different empirical probability measures of equal size P_n, Q_n on $\mathbb{R}/\mathbb{Z}$, the circle can be "cut" and laid out on the real line so that computing the optimal transport between P_n and Q_n is equivalent in both spaces. This cutpoint can be found in practice and therefore computation of Wasserstein distance on the circle can be reduced to its counterpart in $\mathbb{R}$. Once the empirical measures have been relocated on the real line, carrying out a goodness-of-fit test based on Wasserstein distance is equivalent to directly performing such a test on $\mathcal{P}(\mathbb{R}/\mathbb{Z})$.

The distribution-free Wasserstein test for distributions on $\mathbb{R}$ introduced in [25] therefore allows for testing $H_0^x : P_x = Q_x$ and $H_0^y : P_y = Q_y$. The main idea of this approach is to use Wasserstein distance to compare $G_m(F_n^{-1})$ to the uniform distribution, where F_n and G_m are the empirical cumulative distribution functions associated to P_n and Q_m, respectively. The corresponding statistic is distribution-free under the null. In our setting, a p-value for the test $H_0 : P = Q$ is given by twice the minimum of the p-values obtained for H_0^x and H_0^y, after multiplicity correction. It is obvious that sensitivity when testing H_0 may not be satisfactory, as two different measures in $\mathbb{R}^2/\mathbb{Z}^2$ can have the same marginals. Nevertheless, rejection of H_0^x or H_0^y implies rejection of H_0, so the test can be always performed as a first move to detect differences, non-rejections of H_0 being further analyzed with an alternative procedure such as the one described in the next section.

3.2. p-value upper bounding

A second approach to build a test of H_0 is to find an upper bound for p-values in (3.1). Such an upper bound will itself be a valid p-value for H_0 if it controls type I error (they remain with probability $1 - \alpha$ over a fixed signification level α under the null). For the sake of efficiency, we will also ask that power tends to one under the alternative.

One first strategy may consist in making use of concentration inequalities, whose goal is to upper bound

$$\mathbb{P}(\mathcal{T}_p(P_n, P) \geq t), \tag{3.2}$$

and adapt them to the two-sample case under the null. If $\phi(n,t)$ is a bound for (3.2) that is decreasing in its second argument, we can derive

$$\mathbb{P}_{H_0}(\mathcal{T}_p(P_n,Q_m)\geq t)\leq \int_0^1 \phi\left(n,\left(t^{\frac{1}{p}}-\phi^{-1}(m,s)^{\frac{1}{p}}\right)^p\right)ds, \tag{3.3}$$

where $\phi^{-1}(m,s)$ denotes the inverse function of $\phi(m,t)$ with respect to t. The proof of (3.3) is stated in the Appendix. A second strategy may be to directly bound $\mathcal{T}_p(P_n,Q_m)$ under the null. Using McDiarmid's inequality, we have the following result for the quadratic cost.

Theorem 3.1. *Let $P,Q\in\mathcal{P}(\mathbb{T}^2)$ and P_n,Q_m be two empirical probability measures of laws P, Q respectively. Then, for all $t\in\mathbb{R}$, we have*

$$P\left(\mathcal{T}_2(P_n,Q_m)-\mathbb{E}\mathcal{T}_2(P_n,Q_m)>t\right)\leq \exp\left(-\frac{nm}{n+m}8t^2\right). \tag{3.4}$$

The previous inequality upper bounds deviations from the mean. However, the expectation in (3.4) could be neglected under the null if its convergence speed when measures come from the same law is proved to be fast. This would directly provide a two-sample concentration inequality and thus an upper bound for p-values (3.1). To study the speed of convergence of the two-sample expectation under the null, we adapt the existing results for the one-sample case. Using directly the results exposed in [12], only bounds of order

$$\mathbb{E}\mathcal{T}_2(P_n,P)=O\left(n^{-\frac{1}{2}}\right) \tag{3.5}$$

can be expected. Another convergence result was proved in [1], when the unknown probability is in fact the uniform one $\mu_{\mathbb{T}^2}$ in $\mathbb{T}^2$. They show, using arguments based on partial differential equations, that

$$\mathbb{E}\mathcal{T}_2(P_n,\mu_{\mathbb{T}^2})\approx\frac{\log(n)}{4\pi n},\quad \text{if } X_i\sim\mu_{\mathbb{T}^2} \text{ for } i=1,\dots,n\ . \tag{3.6}$$

We can observe that if the following assumption holds, the convergence of the mean becomes faster.

Assumption 1. *There exists a one to one map T such that $T\#\mu_{\mathbb{T}^2}=P$ which is Lipschitz in $\mathbb{T}^2$.*

Such improvement of the speed of the convergence of the mean with respect to (3.5) is formalized in the next result.

Lemma 3.2. *Let $P\in\mathcal{P}(\mathbb{T}^2)$ be such that Assumption 1 holds with constant L, then we have that*

$$\mathbb{E}\mathcal{T}_2(P_n,P)=O\left(\frac{\log(n)}{n}\right). \tag{3.7}$$

Remark 3.3. *After Remark 4.25 in [11], Assumption 1 holds for all measures on $\mathbb{T}^2$ supported on an uniformly convex set and bounded away from zero and infinity on its support.*

If Assumption 1 holds, the convergence speed of the two-sample expectation under the null also improves, thanks to the next result.

Lemma 3.4. *Let $P = Q \in \mathcal{P}(\mathbb{T}^2)$. Then, we have that*

$$\mathbb{E}\mathcal{T}_2(P_n, Q_m) \leq \mathbb{E}\mathcal{T}_2(P_n, P) + \mathbb{E}\mathcal{T}_2(Q_m, Q) + \sqrt{\mathbb{E}\mathcal{T}_2(P_n, P)\,\mathbb{E}\mathcal{T}_2(Q_m, Q)}. \quad (3.8)$$

Thus, the convergence speed of the two-sample expectation under the null has improved after the last result and (3.7). As it is substantiallly faster than the one of $\mathcal{T}_2(P_n, Q_m)$ (that we can derive from (3.5) with a triangular inequality), we can neglect the two-sample expectation and derive

$$\mathbb{P}_{H_0}(\mathcal{T}_2(P_n, Q_m) \geq t) \lesssim \exp\left(-\frac{nm}{n+m} 8t^2\right). \quad (3.9)$$

This second upper bound is sharper than the first one (3.3) with the concentration inequality stated in [36]. Thus, inequality (3.9) will be used for testing. Recalling Remark 3.3, this testing procedure can be implemented for all measures on $\mathbb{T}^2$ that are supported on an uniformly convex set and bounded away from zero and infinity.

3.3. Asymptotic behaviour under the alternative

As no asymptotic distribution is known for the statistic $\mathcal{T}_p(P_n, Q_m)$ when both empirical measures follow the same law, we can try to detect similarities by testing $H_0 : P \neq Q$. A similar approach has been adressed in [9] for measures in $\mathbb{R}$, but with the advantage that the distance between the true laws $\mathcal{T}_p(P, Q)$ appears in the asymptotic result, allowing testing of $\mathcal{T}_p(P, Q) \geq \Delta_0$, for a given threshold Δ_0. In the same way, the earlier work [13] also introduced such an asymptotic test for assessing similarities based on the trimmed Wasserstein distance, for measures on the real line whose samples can be dependent. This can not be applied for measures in $\mathbb{T}^2$, where the derived CLT 2.5 only states gaussian deviations from the mean. If we use (2.7), we could consider the statistic

$$\frac{\mathcal{T}_p(P_n, Q_m) - \mathbb{E}\mathcal{T}_p(P_n, Q_m)}{\mathrm{Var}(\mathcal{T}_p(P_n, Q_m))} \overset{P \neq Q}{\approx} N(0, 1), \quad (3.10)$$

where, in practice, the variance and expectation can be estimated by bootstrapping the given samples. However, this result won't allow efficient testing as the given statistic tends faster to zero when $P = Q$, so the null and the alternative cannot be distinguished under the equality of measures. As it will be illustrated in simulations, no relevant results can be obtained from its implementation. Further discussion about this issue can be found in Section 6.

4. Simulations and results

The three introduced testing techniques would allow appropriate goodness-of-fit testing as long as they control type I error (sensitivity) at a fixed level α and

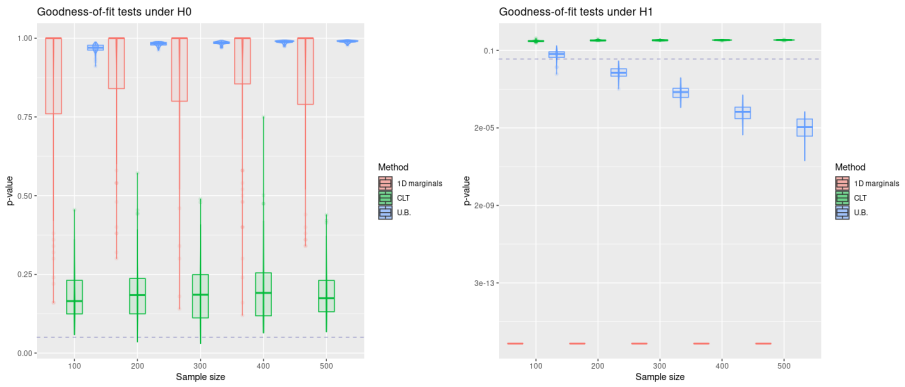
they are consistent under fixed alternatives, i.e. power (specificity) tends to one as sample sizes tend to infinity. The validity of both conditions will be assessed via their implementation on simulated and real protein structure data.

4.1. Simulated data

The three introduced approaches have first been applied to simulated data on $\mathbb{T}^2$. Under the null $H_0 = Q$, we first considered two uniform laws $\mu_{\mathbb{T}}^2$ on $\mathbb{T}^2$. Under the alternative, we consider the following three cases:

$$H_1 : P \sim \mu_{\mathbb{T}}^2 \text{ and } Q \sim N(\mathbf{0.5}, \Sigma_1),$$
$$H_2 : P \sim N(\mathbf{0.25}, \Sigma_2) \text{ and } Q \sim N(\mathbf{0.75}, \Sigma_2),$$
$$H_3 : P \sim N(\mathbf{0.25}, \Sigma_3) \text{ and } Q \sim N(\mathbf{0.75}, \Sigma_3),$$

where $\Sigma_k^{ij} = 10^{-4}$ for $k = 1, 2, 3$ and $i \neq j$, $\Sigma_1^{ii} = 0.01$, $\Sigma_4^{ii} = 0.05$ and $\Sigma_5^{ii} = 0.005$ for $i = 1, 2$. Uniform and gaussian distributions satisfy Assumption 1 as their support is uniformly convex, so inequality (3.9) can be applied. Results for H_0 and H_1 are depicted in Figure 1.



(a) $H_0 : P = Q$ (b) $H_1 : P \neq Q$

FIGURE 1. *p-value distributions under H_0 (a) and H_1 (b) for the three proposed testing methods and for different sample sizes. Values below machine precision (2.2e-16) are set to this value (1D marginals in (b)). The dashed line indicates an arbitrary level of significancy of $\alpha = 0.05$. Figure shows superposed boxplots and violin plots.*

As it was expected, the CLT-based test doesn't provide any significant result under the alternative, so this method should be discarded. Conversely, the two remaining techniques show satisfactory levels of sensitivity and specificity: the marginal test is efficient even for small (~ 100) sample sizes, and the power of the upper bounding test tends to zero when sample size increases, completely rejecting the null at level $\alpha = 0.05$ for more than 200 individuals. Results follow the same trend for the two other alternatives considered.

As the one-dimensional method compares marginals on $\mathbb{R}/\mathbb{Z}$, different measures on $\mathbb{R}^2/\mathbb{Z}^2$ with equal marginals become a special case of interest. An additional alternative regarding that case is considered:

$$H_4 : P \sim N(\mathbf{0.5}, \Sigma_4) \text{ and } Q \sim N(\mathbf{0.5}, \Sigma_4'),$$

where $\Sigma_4^{ii} = \Sigma_4'^{ii} = 0.02$ for $i = 1, 2$, and $\Sigma_4^{ij} = -\Sigma_4'^{ij} = 0.019$ for $i \neq j$. This allows the assessment of whether the upper bounding technique efficiently retrieves significant results in scenarios where the one-dimensional method is ineffective. The corresponding p-values have been simulated and results are depicted in Figure 2.

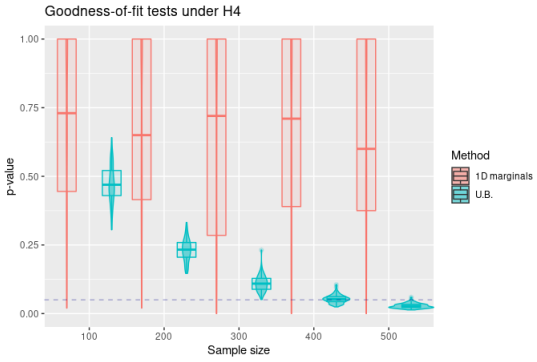


FIGURE 2. *p-value distributions under H_4 for the two first proposed testing methods and for different sample sizes. The dashed line indicates an arbitrary level of significancy of $\alpha = 0.05$. Figure shows superposed boxplots and violin plots.*

As expected, the one-dimensional approach does not retrieve significant differences between two different laws with equal marginals. However, the upper bounding approach does reject the null for big-enough sample sizes, as illustrated in Figure 2. This shows how using the upper bound (3.9) complements the first technique when the latter is ineffective (for example here, where the only difference between both laws is an opposite correlation). This, and conclusions obtained after Figure 1, can be restated with another possible visualization, as the one depicted in Figure 4, where the empirical cumulative distribution functions of p-values corresponding to hypothesis H_0, H_1 and H_4 are compared for the two considered approaches. Further discussion about the convenience of using one approach or the other depending on the situation can be found in section 6, where computational issues are also discussed.

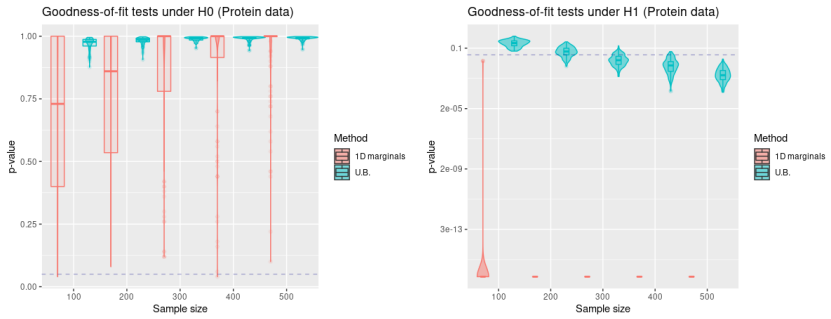
4.2. Protein structure data

A method to accurately compare local structural preferences in conformational ensemble models of proteins is extremely useful to understand the sequence-structure-function relationships, allowing, for instance, assessing structural effects of sequence mutations. The local structure of a protein is determined by

two dihedral angles, ϕ and ψ, which describe the conformational state of each amino acid residue along the sequence. For most amino acid types (for all excepting proline and glycine), the distribution of ϕ and ψ angles is supported on the same subset of $\mathbb{T}^2$, which, even if there exist some physically forbidden regions, is uniformly convex. This is illustrated in Figure 5. We can also assume that density is continuous and strictly positive in its support, so conditions in Remark 3.3 are satisfied. Both the marginal and the upper bounding test can therefore be implemented, and sensitivity and specificity of both approaches will be again simulated for protein structure data.

For the analysis presented here, we used a structural database of three-residue fragments (also called tripeptides) extracted from experimentally-determined high-resolution protein structures [20]. The reason to consider tripeptides instead of single amino acid residues is that the distribution of the ϕ and ψ angles does not depend only on the amino acid type, but also on the sequence context, and particularly on the closest neighbors.

Under H_0, we will consider structural data corresponding to the central amino acid residue of the Ala-Ala-Ala tripeptide, which is a protein fragment having three consecutive alanine residues. To simulate under the alternative, two tripeptides with significant different structural properties have to be chosen. We will again consider Ala-Ala-Ala, which has a high propensity to form helices, and Leu-His-Leu (a fragment of leucine, histidine and leucine), which shows a low helical ppropensity. The corresponding empirical distributions are depicted in Figure 6. Before implementation, data have to be rescaled to $[0,1] \times [0,1]$ to correctly apply the upper bound. Results are shown in Figure 3.



(a) $H_0 : P = Q$ (b) $H_1 : P \neq Q$

FIGURE 3. *p-value distributions under H_0 (a) and H_1 (b) for the two first proposed testing methods and for different sample sizes of protein structure data. The dashed line indicates an arbitrary level of significance of $\alpha = 0.05$. Figure shows superposed boxplots and violin plots.*

Both the marginal and the upper bounding test show satisfactory levels of sensitivity and specificity. Type I error is controlled at level $\alpha = 0.05$ for both techniques and all sample sizes, and power tends to one when sample size increases for the two procedures. The power of the marginal test equals one (at level $\alpha = 0.05$) for all the simulated sample sizes, and the upper bounding

technique controls power when ~ 500 individuals form the sample. The lack of individuals should not prevent from performing this second procedure, as the available structural datasets are large-enough for most of the tripeptides.

Finally, even if the Central Limit Theorem 2.5 does not allow goodness-of-fit testing, we can, to conclude, illustrate the asymptotic behavior of Wasserstein distance using protein structure data. For this, we used data from two very different tripeptides from the structural point of view: Ala-Ala-Ala and Ala-Gly-Ala, where the glycine in the middle changes abruptly the configuration. We simulated their squared Wasserstein distance distribution for different sample sizes and represent its normalized deviation from the mean. Results are depicted in Figure 7, where all the resulting distributions are significantly standard normals according to a Kolmogorov-Smirnov test.

5. Summary

Important results of Optimal Transport Theory have been extended to the d-dimensional flat torus $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}$, specially the Central Limit Theorem 2.5, which states asymptotic Gaussian deviations from the mean. The particular case of $d = 2$ has been addressed in detail, with the aim of defining goodness-of-fit tests for measures on such space. For the first considered technique, the equality of marginals on $\mathbb{R}/\mathbb{Z}$ is tested by transferring the problem to the real line in such a way that the computation of Wasserstein distance is equivalent in both spaces. This is made using the work in [23] and the one-dimensional distribution free statistic introduced in [25]. The second approach consists on upper bounding p-values (3.9). This is possible thanks to the derived concentration inequalities (3.4) for the two-sample empirical Wasserstein distance with the quadratic cost, and to the improved convergence speed of its expectation (3.2).

The presented techniques should be of great interest for the Structural Biology community, as they represent a mathematically efficient solution for the problem of comparison of local protein structures. Both approaches are built under statistical guarantees and are based on the geometry of the underlying space, which is fundamental when a physical problem is assessed. We believe that the goodness-of-fit tests defined in this paper constitute a relevant building block for the study of the sequence-structure-function relationship in proteins, and in particular for Intrinsically Disordered Proteins (IDPs), allowing their structural investigation with mathematical guarantees.

6. Discussion

When implementing the two testing approaches, some practical and numerical considerations should be taken into account. The first technique shows good empirical performance for all sample sizes, but it is by construction unable to detect differences when both laws have equal marginals. This sensitivity issue can be solved by applying the second approach when the first one is not conclusive. The upper bound test shows good empirical performance across all the

simulation settings considered, but requires larger sample sizes to be powerful. Both approaches are therefore complementary, and we suggest the following guidelines. First, when only small samples are available, the marginal approach must be used. If sample sizes are large enough to apply the upper bound, we recommend to perform the marginal approach as a first move in all cases, as rejections are more abrupt under the alternative. However, this technique also requires substantially more computing time, as simulation of the statistic under the null is required, and the "cutpoints" on the circle must be found to lay the problem on the real line. This may slow down computation if a large number of tests has to be performed. The computation of the upper bound is substantially faster, as only the values of the statistic need to be computed. In any case, this approach should be always performed after non-rejections of the one-dimensional test, to detect any possible significant difference between laws with equal marginals.

The results presented in Section 2 fulfill the fundamental study of Optimal Transport on the d-dimensional flat torus. The issue of two-sample goodness of fit testing studied in Section 3 remains largely open. Our contribution in this respect is to propose easily implementable goodness-of-fit testing approaches that are built on top of state-of-the-art tools in Optimal Transport. Finding the exact or asymptotic distribution of the Wasserstein statistic in general dimension remains one of the main unsolved problems of the theory of Optimal Transport, preventing the construction of more efficient two-sample goodness-of-fit tests. An asymptotic approach for measures supported on a finite set has been presented in [32] and, in the one-dimensional case, [3] have obtained a CLT under the null $P = Q$ for deviations of $W_p(P_n, Q_n)$ from the true distance $W_p(P, Q)$ (instead of $\mathbb{E}(W_p(P_n, Q_n))$). The results of [3] are already quite challenging mathematically, and extensions to higher dimensions are clearly beyond the scope of the present paper. Altogether, we believe that the techniques here presented can already be of great interest in and beyond the Structural Biology community, as they allow solving the goodness-of-fit testing problem for two distributions lying on general periodic spaces, which appears in many other domains of application.

Code availability

The R code implementing the statistical test approaches presented in this work is available at https://github.com/gonzalez-delgado/wgof_torus. For the computation of empirical Wasserstein distances, we made use of the R package `transport` [29].

Appendix A: Proofs

A.1. Proofs of Section 2

Proof of Theorem 2.1. Set Γ defined as in (2.5). Since, after Proposition 2 in [6], Γ is cyclically monotone, Theorem 12.25 in [22] states that there exists a

convex function φ such that $\Gamma \subset \partial\varphi$. Moreover, note that if $(x, y) \in \Gamma$ then $\bar{x} \in \text{supp}(P)$ and $x + p \in \text{supp}(\mu_P)$, for all $p \in \mathbb{Z}^d$, which implies that

$$\text{supp}(\mu_P) \subset \text{dom}(\varphi). \tag{A.1}$$

The set of differentiable points of φ, denoted as $\text{dom}(\nabla\varphi)$ is of full Lebesgue measure in its domain (Theorem 25.5 in [27]). This, together with (A.1) implies that φ is μ_P-almost everywhere differentiable.

Let $x \in \text{supp}(\mu_P)$ be such that $x + p \in \text{dom}(\nabla\varphi)$, for all $p \in \mathbb{Z}^d$. Then, there exists y such that $(x, y) \in \Gamma \subset \partial\varphi$. Since $x \in \text{dom}(\nabla\varphi)$ then $y = \nabla\varphi(x)$. For any $p \in \mathbb{Z}^d$ we have that $x+p \in \text{supp}(\mu_P)$ and $\nabla\varphi(x)+p$ is the unique such that $(x + p, \nabla\varphi(x) + p) \in \Gamma$. Since the countable intersection of sets of full measure is of full measure, (A.1) implies the relation

$$\nabla\varphi(x + p) = \nabla\varphi(x) + p, \quad \mu_P\text{-almost everywhere.} \tag{A.2}$$

We denote as $\Delta_{P,\varphi}$ the set of x such that (A.2) holds. In consequence the map

$$\begin{aligned} \overline{\nabla\varphi} : \text{supp}(P) &\to \mathbb{T}^d \\ \bar{x} &\mapsto \overline{\nabla\varphi(x)} \end{aligned}$$

is P-a.s. well defined. The following Lemma is a consequence of the definition of Γ.

Lemma A.1. *If $(x, y) \in \Gamma$, then $(\bar{x}, \bar{y}) \in \text{supp}(\pi^*)$ and, if $(\bar{x}, \bar{y}) \in \text{supp}(\pi^*)$, then there exists $p, q \in \mathbb{Z}^d$ such that $(x + p, y + q) \in \Gamma$.*

Setting $x \in \Delta_{P,\varphi}$ and applying Lemma A.1 we have that $(\bar{x}, \overline{\nabla\varphi}(\bar{x})) \in \text{supp}(\pi^*)$. Moreover, if $(\bar{x}, \bar{y}) \in \text{supp}(\pi^*)$, Lemma A.1 implies that there exists $p, q \in \mathbb{Z}^d$ such that $(x+p, y+q) \in \Gamma$, which implies that $y+q = \nabla\varphi(x)+p$. Consequently, we have that $\bar{y} = \overline{\nabla\varphi}(\bar{x})$, which implies that for P-almost every $\bar{x}$ there exists a unique $\bar{y}$ such that $(\bar{x}, \bar{y}) \in \text{supp}(\pi^*)$ and, furthermore, $\bar{y} = \overline{\nabla\varphi}(\bar{x})$. This final statement concludes the proof. □

Proof of Theorem 2.2. For every $(x, y) \in \Gamma$, by definition we have that $d(\bar{x}, \bar{y}) = ||x-y||$. Since $(\bar{x}, \bar{y}) \in \text{supp}(\pi^*)$, Theorem 5.10 in [35] establishes that if (f, f^{d^2}) solves (2.2) then

$$f(\bar{z}) \leq f(\bar{x}) + d^2(\bar{z}, \bar{y}) - d^2(\bar{x}, \bar{y}) \quad \text{for all } \bar{z} \in \mathbb{T}^d. \tag{A.3}$$

More precisely, (A.3) can be replaced by

$$f(\bar{z}) \leq f(\bar{x}) + d^2(\bar{z}, \bar{y}) - ||x - y||^2 \quad \text{for all } \bar{z} \in \mathbb{T}^d.$$

Using that $d(\bar{z}, \bar{y}) \leq ||z - y||$, for all $(x, y) \in \Gamma$ we have that

$$f(\bar{z}) \leq f(\bar{x}) + ||z - y||^2 - ||x - y||^2 \quad \text{for all } z \in \mathbb{R}^d. \tag{A.4}$$

This implies that the function $\tilde{f}$ defined by the relation $\tilde{f}(x) = \frac{1}{2}\left(||x||^2 - f(\bar{x})\right)$ has non empty subgradient for every $x \in \mathcal{K}(P)$, where

$$\mathcal{K}(P) = \{x \in [0,1]^d \text{ such that } \bar{x} \in \text{supp}(P)\}. \tag{A.5}$$

Note that from (A.4) we have that, for all $z \in \mathbb{R}^d$ and $y \in \partial\tilde{f}(x)$,

$$\begin{aligned}
\tilde{f}(z) &= \frac{||z||^2 - f(\bar{z})}{2} \\
&\geq \frac{||z||^2 - f(\bar{x}) - ||z-y||^2 + ||x-y||^2}{2} \\
&= \frac{||z||^2 - f(\bar{x}) - ||z||^2 - ||y||^2 + 2\langle z, y\rangle + ||x||^2 + ||y||^2 - 2\langle x, y\rangle}{2} \\
&= \frac{-f(\bar{x}) + ||x||^2 + 2\langle z - x, y\rangle}{2} = \tilde{f}(x) + \langle z - x, y\rangle.
\end{aligned}$$

This subgradient is defined for each $x \in \mathcal{K}(P)$ as

$$\partial\tilde{f}(x) = \left\{y \in \mathbb{R}^d : \ (\bar{x}, \bar{y}) \in \text{supp}(\pi^*) \text{ and } d(\bar{x}, \bar{y}) = ||x - y||\right\}.$$

Let us define the convex function

$$g_f(x) = \sup_{z \in \mathcal{K}(P),\ y \in \partial\tilde{f}(z)} \{\tilde{f}(z) + \langle y, z - x\rangle\} \tag{A.6}$$

and realize that $g_f(x) = \tilde{f}(x)$ for every $x \in \mathcal{K}(P)$. Set $x \in \mathcal{K}(P)$ and note that, since it is a supremum,

$$g_f(x) \geq \tilde{f}(x), \tag{A.7}$$

and if $y \in \partial\tilde{f}(z)$ we have

$$\tilde{f}(z) + \langle y, x - z\rangle \leq \tilde{f}(x), \tag{A.8}$$

and the equality holds. Finally, the next Lemma concludes the proof.

Lemma A.2. *Let $P, Q \in \mathcal{P}(\mathbb{T}^d)$ be probability measures such that their associated periodic measures satisfy $\mu_P, \mu_Q \ll \ell_d$, $f : \mathbb{T}^d \to \mathbb{R} \cup \{-\infty\}$ be a d^2-concave solution of (2.2), φ be defined in Theorem 2.1, and $g_f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be the convex function defined by the relation*

$$g_f(x) = \sup_{z \in \mathcal{K}(P),\ y \in \partial\tilde{f}(z)} \{\frac{1}{2}\left(||x||^2 - f(\bar{x})\right) + \langle y, z - x\rangle\}. \tag{A.9}$$

If supp(P) is connected and with Lebesgue negligible boundary, then there exists a constant $C \in \mathbb{R}^d$ such that $g_f(x) = \varphi(x) + C$, for all $x \in \mathcal{K}(P)$.

Proof. Since both functions are convex then they are locally Lipischitz. We claim that there exists a set $D \subset \mathcal{K}(P) \cap \text{dom}(\nabla\varphi) \cap \text{dom}(\nabla g_f)$ such that

$$\ell_d(\mathcal{K}(P) \setminus D) = 0 \ \text{ and } \ \nabla g_f(x) = \nabla\varphi(x), \ \text{ for all } x \in D. \tag{A.10}$$

Once the claim is proved we can conclude by using Theorem 2.6 in [8]. To prove the claim note that, since both convex functions are finite in $\mathcal{K}(P)$, then $\mathrm{dom}(\nabla\varphi)\cap\mathrm{dom}(\nabla g_f)$ is of full Lebesgue measure in $\mathcal{K}(P)$. Setting $x\in\mathcal{K}(P)\cap\mathrm{dom}(\nabla\varphi)\cap\mathrm{dom}(\nabla g_f)$, and $y_x\in\mathbb{R}^d$ such that $(x,y_x)\in\Gamma$, then we deduce the following assertions:

- Since $\Gamma\subset\partial\varphi$ and $x\in\mathrm{dom}(\nabla\varphi)$, then $y_x=\nabla\varphi(x)$.
- By definition of $\partial\tilde{f}(x)$, that coincides with $\partial g_f(x)$ for $x\in\mathcal{K}(P)$, we have that $y\in\partial g_f(x)$ if $\bar{y}\in\partial^{d^2}f(\bar{x})$ and $d(\bar{x},\bar{y})=||x-y||$. Note that y_x satisfies that condition, since $x\in\mathrm{dom}(\nabla g_f)$. In consequence we have that $y_x=\nabla g_f(x)$.

Then (A.10) holds with $D=\mathrm{dom}(\nabla\varphi)\cap\mathrm{dom}(\nabla g_f)$. □

□

Proof of Lemma 2.3. Set $\bar{x}\in\mathrm{dom}(f)$. Then, the set

$$\partial^{d^2}f(\bar{x})=\{\bar{y}:\ f(\bar{z})\leq f(\bar{x})+d^2(\bar{z},\bar{y})-d^2(\bar{x},\bar{y}),\ \text{for all } \bar{z}\in\mathbb{T}^d\}$$

is non empty. Set $\bar{y}_x\in\partial^{d^2}f(x)$, by definition, for all $\bar{z}\in\mathbb{T}^d$

$$f(\bar{z})-f(\bar{x})\leq d^2(\bar{z},\bar{y}_x)-d^2(\bar{x},\bar{y}_x).$$

For every $\bar{z}\in\mathrm{dom}(f)$ we can repeat the previous reasoning and obtain that

$$|f(\bar{z})-f(\bar{x})|\leq\sup\left(|d^2(\bar{z},\bar{y}_x)-d^2(\bar{x},\bar{y}_x)|,|d^2(\bar{z},\bar{y}_z)-d^2(\bar{x},\bar{y}_z)|\right).$$

Finally, the relation $a^2-b^2=(a-b)(a+b)$ and the triangle inequality of the distance d lead to

$$\begin{aligned}|f(\bar{z})-f(\bar{x})|&\leq d(\bar{z},\bar{x})\sup\left(|d(\bar{z},\bar{y}_x)+d(\bar{x},\bar{y}_x)|,|d(\bar{z},\bar{y}_z)+d(\bar{x},\bar{y}_z)|\right)\\&\leq 2d(\bar{z},\bar{x})\sup_{\bar{z},\bar{x}\in\mathbb{T}^d}(d(\bar{z},\bar{x}))\leq 2d(\bar{z},\bar{x}).\end{aligned}$$

□

Proof of Theorem 2.4. Set $\bar{p}\in\mathrm{supp}(P)$ and assume that $f(\bar{p})=0$. Set $\epsilon_m\to 0$ and consider the sequence of balls $\mathbb{B}_{\epsilon_m}(\bar{p})\subset\mathrm{supp}(P)$, centered in p and radius ϵ_n. Since the ball is a continuity set of P, by Portmanteau Theorem, then $P_n\xrightarrow{w}P$ implies that for each m there exists a n_m such that P_n gives mass to $\mathbb{B}_{\epsilon_m}(\bar{p})$ for all $n\geq m_n$. Then, we can extract a sequence $\bar{p}_n\to\bar{p}$ such that $\bar{p}_n\in R_{P_n}$. In consequence, we have that $f_n(\bar{p}_n)\in\mathbb{R}$, we can set $a_n=-f_n(\bar{p}_n)$ and define $h_n=f_n+a_n$. Recall from Lemma 2.3 that all such functions are 2-Lipschitz in their respective domains. Kirszbraun Theorem implies that, without loss of generality, we can consider that h_n (resp. f) are 2-Lipschitz functions defined in the whole $\mathbb{T}^d$. The previous reasoning implies that $\{h_n\}_{n\in\mathbb{N}}$ is pointwise bounded for the compact sequence $\{\bar{p}_n\}_{n\in\mathbb{N}}$. Since all such functions are 2-Lipschitz, then Arzelá-Ascoli's Theorem concludes that every subsequence $\{h_{n_k}\}_{k\in\mathbb{N}}$ admits a convergent subsequence $\{h_{n_{k_j}}\}_{j\in\mathbb{N}}$. The uniqueness described in Theorem 2.2 and the fact that $\bar{p}_n\to\bar{p}$ and $h_n(\bar{p}_n)=f(\bar{p})=0$ conclude that f is the unique possible limit of such subsequences. □

Proof of Theorem 2.5. Note that as Theorem 2.4 holds, since probabilities are supported in a compact set, the torus, then the reasoning of [10] can be imitated. Here the main steps of the proof for the one sample case are given. For further details about the proof we refer to the original text.

Efron-Stein inequality, see Chapter 3.1 in [5], states that if $(X_1', \dots, X_n')$ is an independent copy of $(X_1, \dots, X_n)$, then we have the bound

$$\mathrm{Var}(f(X_1, \dots, X_n)) \leq \sum_{i=1}^{n} \mathbb{E}(f(X_1, \dots, X_n) - f(X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n).)_+^2.$$

Moreover, if $X_1, \dots, X_n$ are i.i.d, such inequality can be written as

$$\mathrm{Var}(f(X_1, \dots, X_n)) \leq n\mathbb{E}(f(X_1, \dots, X_n) - f(X_1', \dots, X_n))_+^2.$$

Set the empirical measures $P_n = \frac{1}{n}\sum_{k=1}^n \delta_{X_k}$ and $P_n' = \frac{1}{n}(\delta_{X_1'} + \sum_{k=2}^n \delta_{X_k})$, and the values $R_n = \mathcal{T}_2(P_n, Q) - \int \varphi dP_n$ and $R_n' = \mathcal{T}_2(P_n', Q) - \int \varphi dP_n'$. Let φ_n and φ_n' be solutions of the dual problem (2.2) of $\mathcal{T}_2(P_n, Q)$ and $\mathcal{T}_2(P_n', Q)$ respectively. Then from (2.2) we derive that

$$\begin{aligned}(R_n - R_n')_+ \leq & \frac{1}{n}|\varphi_n(X_1) - \varphi(X_1) - \varphi_n(X_1') + \varphi(X_1')| \\ & + |\varphi_n'(X_1) - \varphi(X_1) - \varphi_n'(X_1') + \varphi(X_1)|,\end{aligned}$$

which together with Theorem 2.4 yields

$$n(R_n - R_n')_+ \xrightarrow{a.s.} 0.$$

Since the probabilities are supported in the torus, which is compact, then $n^2\mathbb{E}(R_n - R_n')_+^2 \to 0$. Finally, we conclude by the so called Efron-Stein's inequality.

□

A.2. Proofs of Section 3

Proof of Theorem 3.1. Note that $\mathcal{T}_2(P_n, Q_m) = \mathcal{T}(X_1, \dots, X_n, Y_1, \dots, Y_m)$ is a function of $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$. For each $x_1, \dots, x_n, y_n, \dots, y_m \in \mathbb{T}^d$ and $x' \in \mathbb{T}^d$ let π and π' be both joint measures such that

$$\begin{aligned}\mathcal{T} := \sum_{i,j} d(x_i - y_j)^2 \pi_{i,j} &= \mathcal{T}(x_1, \dots, x_n, y_1, \dots, y_m) \\ s.t. \quad \sum_{i,j} \pi_{i,j} &= \frac{1}{n}, \quad j = 1, \dots, m, \\ \sum_{i,j} \pi_{i,j} &= \frac{1}{m}, \quad i = 1, \dots, n,\end{aligned}$$

and

$$\begin{aligned}\mathcal{T}' := \sum_j d(x'_1 - y_j)^2 \pi'_{1,j} + \sum_{i>1,j} d(x_i - y_j)^2 \pi'_{i,j} = \mathcal{T}(x'_1, \dots, x_n, y_1, \dots, y_m)\\ s.t. \quad \sum_{i,j} \pi'_{i,j} = \frac{1}{n}, \quad j = 1, \dots, m,\\ \sum_{i,j} \pi'_{i,j} = \frac{1}{m}, \quad i = 1, \dots, n.\end{aligned}$$

Then we have that

$$\mathcal{T}' \leq \sum_j d(x'_1 - y_j)^2 \pi_{1,j} + \sum_{i,j} d(x_i - y_j)^2 \pi_{i,j},$$

which implies

$$\begin{aligned}\mathcal{T}' - \mathcal{T} &\leq \sum_j \left(d(x_i - y_j)^2 - d(x_1 - y_j)^2 \right) \pi_{1,j} \\ &\leq \sum_j \frac{1}{2} \pi_{1,j} = \frac{1}{2n},\end{aligned}$$

where the second inequality comes from the fact that $d^2(x, y) \leq 1/2$ in $\mathbb{T}^d$. By symmetry we also obtain the reverse inequality. Doing the same with y'_1 and y_1 we obtain the bound $\frac{1}{2m}$. By using McDiarmid's inequality, see [17], we derive that

$$P\left(\mathcal{T}_2(P_n, Q_m) - \mathbb{E}\mathcal{T}_2(P_n, Q_m) > t\right) \leq \exp\left(-\frac{nm}{n+m} 8t^2\right).$$

□

Proof of Lemma 3.2. Denote T the function of Assumption 1. We observe that $U_i = T^{-1}(X_i) \sim \mu_{\mathbb{T}^d}$, where $U_1, \dots U_n$ is a i.i.d. sample. Then, by (3.6) we have $\mathbb{E}\mathcal{W}_2^2(T^{-1}\#P_n, T^{-1}\#P) = O\left(\frac{\log(n)}{n}\right)$. Let H_n be the optimal transport map between $\mu_{\mathbb{T}^d}$ and $T^{-1}\#P_n$, then the map $S = T \circ H_n \circ T^{-1}$ satisfies that $S\#P = P_n$. Finally, we conclude by noticing that

$$\begin{aligned}\mathcal{W}_2^2(P_n, P) \leq \int |S(x) - x|^2 dP(x) = \int |T(H_n(x)) - T(x)|^2 dP(x)\\ \leq L\mathcal{W}_2^2(T^{-1}\#P_n, T^{-1}\#P).\end{aligned}$$

□

Proof of (3.3). Let $\mathbb{P}(\mathcal{T}_p(P_n, P) \geq t) \leq \phi(n, t)$ and $\phi^{-1}(n, s)$ denote the inverse

function of ϕ with respect to the second variable. Under the null, we have

$$\begin{aligned}\mathbb{P}(\mathcal{W}_p(P_n, Q_m) \geq t) \leq \mathbb{P}(\mathcal{W}_p(P_n, P) + \mathcal{W}_p(Q_m, Q) \geq t) = \\ = \mathbb{E}\left(\mathbb{P}(\mathcal{W}_p(Q_m, Q) \geq t - \mathcal{W}_p(P_n, P)) = \right. \\ = \int_0^1 \mathbb{P}\left(\mathbb{P}\left(\mathcal{W}_p(Q_m, Q) \geq t - \mathcal{W}_p(P_n, P)\right) > s\right) ds \leq \\ \leq \int_0^1 \mathbb{P}\left(\phi(m, (t - \mathcal{W}_p(P_n, P))^p) > s\right) ds = \int_0^1 \mathbb{P}\left(\mathcal{W}_p(P_n, P) \geq t - \phi^{-1}(m, s)^{\frac{1}{p}}\right) ds,\end{aligned}$$

as $\phi^{-1}(m, s)$ is a decreasing function. Finally, after upper bounding one last time:

$$\mathbb{P}_{H_0}(\mathcal{T}_p(P_n, Q_m) \geq t) \leq \int_0^1 \phi(n, \left(t^{\frac{1}{p}} - \phi^{-1}(m, s)^{\frac{1}{p}}\right)^p ds,$$

which concludes the proof. □

Proof of Lemma 3.4. To derive (3.8), it suffices to show that $E(\mathcal{T}_2(P_n, Q_m))$ converges, under the null, with the same speed as for the one-sample expectation when Assumption 1 holds (recall Lemma 3.2). When $P = Q$ we have

$$\begin{aligned}\mathbb{E}\mathcal{T}_2(P_n, Q_m) \leq \mathbb{E}\left((\mathcal{W}_2(P_n, P) + \mathcal{W}_2(Q_m, Q))^2\right) = \\ = \mathbb{E}\mathcal{T}_2(P_n, P) + \mathbb{E}\mathcal{T}_2(Q_m, Q) + \mathbb{E}(\mathcal{W}_2(P_n, P)\mathcal{W}_2(Q_m, Q)).\end{aligned}$$

Finally, as the last term is in fact an inner product $\langle \mathcal{W}_2(P_n, P), \mathcal{W}_2(Q_m, Q)\rangle$, we have, using Cauchy-Schwarz inequality,

$$\mathbb{E}(\mathcal{W}_2(P_n, P)\mathcal{W}_2(Q_m, Q)) \leq \sqrt{\mathbb{E}\mathcal{T}_2(P_n, P)\,\mathbb{E}\mathcal{T}_2(Q_m, Q)},$$

which prooves the improved convergence speed of the expectation. □

Supplementary Material

Supplementary figures

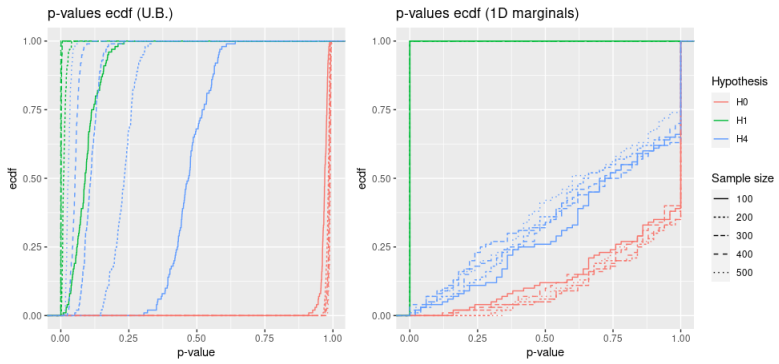
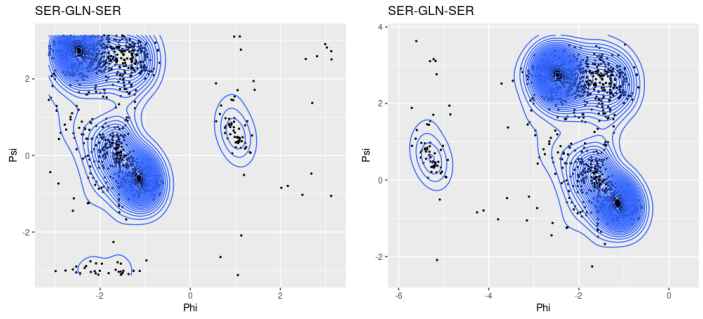
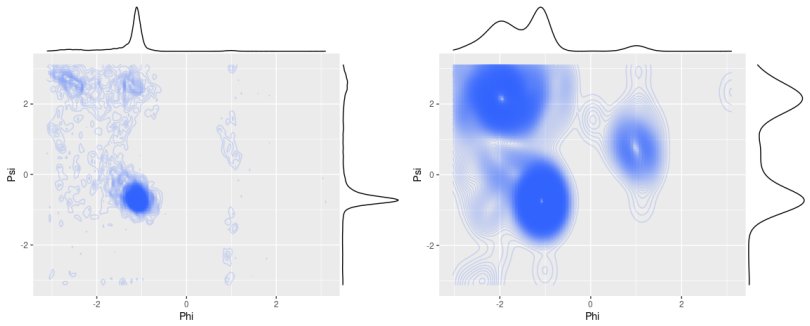


FIGURE 4. *Simulated empirical cumulative distribution functions for p-values corresponding to the upper bound (left) and one-dimensional (right) approaches, under the hypothesis H_0, H_1 and H_4 and for different sample sizes.*



(a) (b)

FIGURE 5. *The support of dihedral angles conformation is uniformely convex. When data is presented in the periodic square $[-\pi, \pi] \times [-\pi, \pi]$ (a), the band dividing the left ($\phi < 0$) and the right ($\phi > 0$) cluster is physically forbidenn, as well as the band separating the left upper ($\phi < 0$, $\psi > -2$) and the left bottom ($\phi < 0$, $\psi < -2$) cluster. However, one can translate points overacross the torus and show (b) that periodicity makes support uniformely convex. The depicted data correspond to tripeptide SER-GLN-SER, a fragment of a serine, a glutamine and a serine residue.*



(a) ALA-ALA-ALA (b) LEU-HIS-LEU

FIGURE 6. *Kernel density estimates of the distributions corresponding to tripeptides ALA-ALA-ALA (a) and LEU-HIS-LEU (b), and of their one-dimensional marginals.*

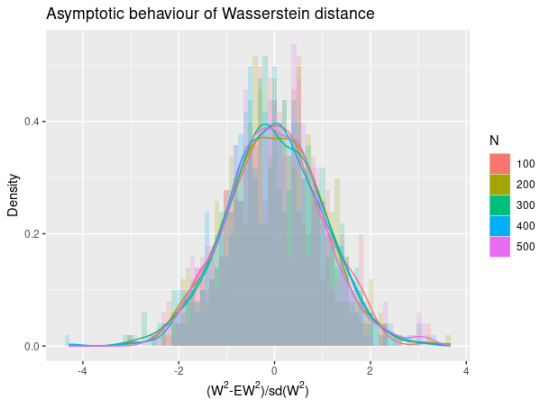


FIGURE 7. *Normalized asymptotic deviations from the mean of squared Wasserstein distance between two different empirical probability measures. Figures shows the corresponding histograms and the associated kernel density estimates, for different sample sizes.*

References

[1] Ambrosio, L., Glaudo, F. and Trevisan, D. (2019). On the optimal map in the 2-dimensional random matching problem. *Discrete & Continuous Dynamical Systems - A* **39** 7291-7308.

[2] Arjovsky, M., Chintala, S. and Bottou, L. (2017). Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.). *Proceedings of Machine Learning Research* **70** 214–223. PMLR.

[3] Berthet, P. and Fort, J.-C. (2019). Weak convergence of empirical Wasserstein type distances. arXiv:1911.02389v1.

[4] Betancourt, M. R. and Skolnick, J. (2004). Local Propensities and Statistical Potentials of Backbone Dihedral Angles in Proteins. *Journal of Molecular Biology* **342** 635 - 649.

[5] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford.

[6] Cordero-Erausquin, D. (1999). Sur le transport de mesures périodiques. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics* **329** 199 - 202.

[7] del Barrio, E., Cuesta-Albertos, J. A., Matran, C. and Rodriguez-Rodriguez, J. M. (1999). Tests of Goodness of Fit Based on the L2-Wasserstein Distance. *The Annals of Statistics* **27** 1230–1239.

[8] Del Barrio, E., González-Sanz, A. and Loubes, J.-M. (2021). Central Limit Theorems for General Transportation Costs. working paper or preprint.

[9] del Barrio, E., Gordaliza, P. and Loubes, J.-M. (2019). A central limit theorem for Lp transportation cost on the real line with application to fairness assessment in machine learning. *Information and Inference: A Journal of the IMA* **8**.

[10] del Barrio, E. and Loubes, J.-M. (2019). Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability* **47** 926 – 951.

[11] Figalli, A. (2017). *The Monge–Ampère Equation and Its Applications*. Zurich Lectures in Advanced Mathematics.

[12] Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields* **162** 707.

[13] Freitag, G., Czado, C. and Munk, A. (2007). A nonparametric test for similarity of marginals—With applications to the assessment of population bioequivalence. *Journal of Statistical Planning and Inference* **137** 697-711. Special Issue on Nonparametric Statistics and Related Topics: In honor of M.L. Puri.

[14] Hallin, M., Mordant, G. and Segers, J. (2021). Multivariate goodness-of-Fit tests based on Wasserstein distance. arXiv:2003.06684v3.

[15] Hundrieser, S., Klatt, M. and Munk, A. (2021). The Statistics of Circular Optimal Transport. arXiv:2103.15426v1.

[16] Lovell, S. C., Davis, I. W., Arendall III, W. B., de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S. and Richardson, D. C. (2003). Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. *Proteins: Structure, Function, and Bioinformatics* **50** 437-450.

[17] McDiarmid, C. (1989). *On the method of bounded differences.* In *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference. London Mathematical Society Lecture Note Series* 148–188. Cambridge University Press.

[18] Mena, G. and Weed, J. (2019). Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *NeurIPS*.

[19] Morris, A. L., MacArthur, M. W., Hutchinson, E. G. and Thornton, J. M. (1992). Stereochemical quality of protein structure coordinates. *Proteins: Structure, Function, and Bioinformatics* **12** 345-364.

[20] Estaña, A., Sibille, N., Delaforge, E., Vaisset, M., Cortés, J. and Bernadó, P. (2019). Realistic Ensemble Models of Intrinsically Disordered Proteins Using a Structure-Encoding Coil Database. *Structure* **27** 381-391.e2.

[21] Peyré, G. and Cuturi, M. (2019). Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning* **11** 355-607.

[22] R. Tyrrell Rockafellar, R. J. B. W. (1998). *Variational Analysis.* Springer, Berlin, Heidelberg.

[23] Rabin, J., Delon, J. and Gousseau, Y. (2009). Transportation Distances on the Circle. *Journal of Mathematical Imaging and Vision* **41**.

[24] Ramachandran, G. N., Ramakrishnan, C. and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology* **7** 95-99.

[25] Ramdas, A., Garcia, N. and Cuturi, M. (2015). On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests. *Entropy* **19**.

[26] Rata, I., Li, Y. and Jakobsson, E. (2010). Backbone statistical potential from local sequence-structure interactions in protein loops. *J Phys Chem B* **114** 1859-1869.

[27] Rockafellar, R. T. (1970). *Convex Analysis.* Princeton University Press.

[28] Santambrogio, F. (2015). Optimal Transport for Applied Mathematicians. Calculus of Variations, PDEs and Modeling.

[29] Schuhmacher, D., Bähre, B., Gottschlich, C., Hartmann, V., Heinemann, F. and Schmitzer, B. (2020). transport: Computation of Optimal Transport Plans and Wasserstein Distances R package version 0.12-2.

[30] Serrurier, M., Mamalet, F., González-Sanz, A., Boissin, T., Loubes, J.-M. and del Barrio, E. (2020). Achieving robustness in classification using optimal transport with hinge regularization. arXiv:2006.06520v3.

[31] Shen, Y., Roche, J., Grishaev, A. and Bax, A. (2018). Prediction

of nearest neighbor effects on backbone torsion angles and NMR scalar coupling constants in disordered proteins. *Protein Science* **27** 146-158.

[32] Sommerfeld, M. and Munk, A. (2018). Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 219-238.

[33] Ting, D., Wang, G., Shapovalov, M., Mitra, R., Jordan, M. and Dunbrack, R. (2010). Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. *PLoS computational biology* **6** e1000763.

[34] Villani, C. (2003). *Topics in Optimal Transportation*. American mathematical society, Providence, Rhode Island.

[35] Villani, C. (2008). *Optimal Transport: Old and New*. Springer-Verlag Berlin Heidelberg.

[36] Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli* **25** 2620 – 2648.