Hierarchical Integral Probability Metrics: A distance on random probability measures with low sample complexity

Marta Catalano [*1] **Hugo Lavenant** [*2]

Abstract

Random probabilities are a key component to many nonparametric methods in Statistics and Machine Learning. To quantify comparisons between different laws of random probabilities several works are starting to use the elegant Wasserstein over Wasserstein distance. In this paper we prove that the infinite dimensionality of the space of probabilities drastically deteriorates its sample complexity, which is slower than any polynomial rate in the sample size. We propose a new distance that preserves many desirable properties of the former while achieving a parametric rate of convergence. In particular, our distance 1) metrizes weak convergence; 2) can be estimated numerically through samples with low complexity; 3) can be bounded analytically from above and below. The main ingredient are integral probability metrics, which lead to the name *hierarchical IPM*.

1. Introduction

In this work we discuss distances between laws of random probabilities, that is, probability measures on spaces of probability measures. Our motivation comes from nonparametric methods in Statistics and Machine Learning, often referred to as distribution-free in that they do not rely on strong assumptions on the underlying distribution of the data. The need for flexibility has inspired a wealth of both frequentist and Bayesian nonparametric models, which often share the use of random probability measures to make inference on the distribution of the data. The randomness, however, is of different nature. Frequentist methods build estimators whose randomness is only due to the randomness of the data. Standard examples include, e.g., empirical processes, histograms, bootstrapping, kernel density estimators, splines, and wavelets (see, e.g, Wasserman (2006) and references therein). In a Bayesian setting the distribution of the data is treated as the unknown parameter, and thus modeled as a random probability $\tilde{P}$. The posterior is the distribution of $\tilde{P}$ *conditionally* on the observed data, thus in principle its randomness does not stem from the data but from the uncertainty about the parameter. The most popular random probability in Bayesian nonparametrics is the Dirichlet process (Ferguson, 1973), though many useful generalizations are available in the literature, including species sampling processes, normalized completely random measures, stick-breaking processes, and kernel mixtures (see, e.g, Ghosal & van der Vaart (2017) and references therein).

Several interesting theoretical and applied findings in nonparametric inference may be framed as approximations of some distribution of interest. These include approximation to a ground truth, as in the study of consistency, to the distribution of another population, as in two-sample tests, to another posterior, as in the merging of opinions, or to the exact but unattainable posterior, as in (Markov chain) Monte-Carlo or other finite-dimensional approximations. In most of these problems to quantify the quality of the approximation one needs a distance between laws of random probabilities, a.k.a., probability distributions on spaces of probability distributions. Such a distance should ideally satisfy the following properties:

1. *Metrization of weak convergence.* Many of the most common random probabilities, such as empirical processes and Dirichlet processes, are almost surely discrete and are typically mutually singular. Weak convergence provides meaningful comparisons between distributions with different supports.
2. *Numerical estimation through samples.* Frequently in applications the exact distribution of the random probability is not known and thus the distance must be approximated numerically through samples, with both low sample complexity (the approximation error as the sample size increases) and low computational complexity (the speed of the algorithm as the sample size increases).

[*]Equal contribution [1]Luiss University, Rome, Italy [2]Bocconi University, Milan, Italy. Correspondence to: Marta Catalano <mcatalano@luiss.it>, Hugo Lavenant <hugo.lavenant@unibocconi.it>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

3. *Analytical upper and lower bounds.* To prove asymptotic statements and to perform theoretical comparisons with numerical estimations, one needs analytical expressions that capture the most meaningful behavior of the distance.

A natural candidate is the Wasserstein distance on $\mathcal{P}(\mathcal{P}(\mathbb{X}))$, the space of laws of random probabilities on the sample space $\mathbb{X}$. Indeed, the Wasserstein distance on a Polish space $\mathbb{X}$ makes the space of probabilities $\mathcal{P}(\mathbb{X})$ a Polish space (cfr. Remark 7.1.7 in Ambrosio et al. (2008)). Thus, one can build a Wasserstein distance on $\mathcal{P}(\mathcal{P}(\mathbb{X}))$ using the Wasserstein distance on $\mathcal{P}(\mathbb{X})$ as ground metric. We refer to the induced metric as *Wasserstein over Wasserstein distance*, also known under the name *Hierarchical Optimal Transport* distance in the optimal transport community. This elegant and intuitive metric has been recently independently defined and used in many different contexts, including the analysis of convergence of the mixing measure in Bayesian hierarchical mixture models (Nguyen, 2016), to measure the similarity measure between documents in topic models (Yurochkin et al., 2019; Bing et al., 2022b;a), for joint clustering of observations and their distributions (Ho et al., 2017), as training loss in generative adversarial networks on images (Dukler et al., 2019), and to measure the discrepancy between datasets for classification tasks (Alvarez-Melis & Fusi, 2020).

In this work we show that though it is indeed possible to estimate the Wasserstein over Wasserstein distance numerically through samples, the infinite dimensionality of $\mathcal{P}(\mathbb{X})$ drastically deteriorates its sample complexity, which can be slower than any polynomial rate of the sample size (Theorem 4.1). A classical workaround to the curse of dimensionality in optimal transport is to use entropic regularization (Mena & Niles-Weed, 2019) but it is unclear if it would be enough here, as the base space $\mathcal{P}(\mathbb{X})$ is infinite-dimensional. Since in practice a very large sample size may not be available or, more often, it may lead to an insurmountable computational burden, we propose a new distance between the laws of random probabilities that can achieve a parametric rate of convergence (Definition 2.4 and Theorem 4.4). One of the main ingredients we use are integral probabilities metrics (Müller, 1997), and for this reason we call our distance *hierarchical IPM*. The hierarchical IPM is dominated by the Wasserstein over Wasserstein distance, and thus retains all analytical upper bounds of the former. We also provide a general strategy to compute lower bounds and use them to recover closed-form expressions of the Wasserstein over Wasserstein distance when applied to empirical measures, to Dirichlet processes, and to the more general species sampling processes (Theorem 3.5). Moreover, in Theorem 3.1 we show that our distance metrizes the weak convergence on compact spaces. This completes our list of desirable properties for a distance on the laws of random probabilities. Note that the analysis we conduct here is valid for a fairly general class of IPMs, but the choice of the underlying function space affects the computational and statistical properties of our distance and may be chosen depending on its intended use.

The Dirichlet process is the cornerstone of Bayesian nonparametrics. It provides the law for a random probability that has been effectively used to build priors and derive posteriors for infinite-dimensional parameters (see Ghosal & van der Vaart (2017) and references therein). Its infinite dimensionality provides elegant and interpretable analytical properties that are less prone to computational algorithms. For this reason, a wealth of finite-dimensional approximations has been developed, each of which recovers the Dirichlet process as a limit. Whereas limiting behaviours are typically well-studied, approximation errors at the level of the random probability are rarely available. To our knowledge, these have been only carried out for almost-sure truncations of the random measures (see, e.g., Muliere & Tardella (1998); Ishwaran & Zarepour (2000); Ishwaran & James (2001); Arbel et al. (2019)) but not for distributional approximations, where the error is usually studied in terms of the L_1 distance between the marginal densities of the data distribution induced by specific Bayesian models (see, e.g., Ishwaran & Zarepour (2000; 2002); Campbell et al. (2019); Lijoi et al. (2020); Nguyen et al. (2023)). As an application of our findings, we use both the Wasserstein over Wasserstein distance and our HIPM to investigate the quality of some of the most common finite-dimensional approximations, namely the Dirichlet multinomial process, the truncated stick-breaking, and the hierarchical empirical measure. We provide numerical estimation of the approximation error and are able to compute non-asymptotic analytical error bounds that provide new rules-of-thumb for deciding the best one to use in practice.

In summary, the main contributions of our work are i) drawing attention on the need for a distance between the laws of random probabilities; ii) showing that the sample complexity of the Wasserstein over Wasserstein distance $\mathcal{W}_{\mathcal{W}}$ can be slower than any polynomial rate in the sample size (Section 4); iii) proposing a new distance, the hierarchical IPM, which is strongly related to $\mathcal{W}_{\mathcal{W}}$ but with better sample complexity (Sections 2 and 4) ; iv) using our distance to provide closed-form expressions for $\mathcal{W}_{\mathcal{W}}$ on empirical measures and on species sampling processes (Section 3); iv) proposing a gradient ascent algorithm to compute our distance (Section 5); v) applying our distance to select the most suitable finite-dimensional approximation of the Dirichlet process in Bayesian nonparametrics (Section 6).

2. Background and main definition

Let $(\mathbb{X}, d_{\mathbb{X}})$ be a Polish (metric, complete and separable) space with a bounded diameter $\mathrm{diam}(\mathbb{X}) = \sup_{x,y\in\mathbb{X}} d_{\mathbb{X}}(x,y)$, e.g. a compact subset of $\mathbb{R}^d$ endowed with the Euclidean distance. We focus on probability distributions on the space $\mathcal{P}(\mathbb{X})$ of probabilities on $\mathbb{X}$, denoted by $\mathcal{P}(\mathcal{P}(\mathbb{X}))$. Elements of $\mathcal{P}(\mathbb{X})$ will usually be denoted as P, whereas elements of $\mathcal{P}(\mathcal{P}(\mathbb{X}))$ will usually be denoted by $\mathbb{Q}$. Random elements on these spaces are distinguished by a $\sim$, e.g. $\tilde{P}$ denotes a random probability. For any measurable function f on $\mathbb{X}$, $P(f) = \int_{\mathbb{X}} f \,\mathrm{d}P$. For a $\mathbb{X}$-valued random variable X we denote its law by $\mathcal{L}(X) \in \mathcal{P}(\mathbb{X})$. We write $\overset{\mathrm{d}}{=}$ for equality in distribution of random variables, that is, $X \overset{\mathrm{d}}{=} Y$ if $\mathcal{L}(X) = \mathcal{L}(Y)$. Eventually, we denote by $\mathbb{E}$ the expectation of random variables, and we use $\mathbb{E}_X$ to emphasize the source of randomness.

We define distances between the laws of random probabilities by using two baseline ingredients: integral probability metrics (Zolotarev, 1984; Müller, 1997) and the Wasserstein distance on a generic bounded Polish space $(\mathbb{Y}, d_{\mathbb{Y}})$, which in this work will be either $\mathbb{X}$ or the space of probabilities $\mathcal{P}(\mathbb{X})$ with a suitable metric.

Definition 2.1. Let $\mathcal{F}$ be a class of $\mathbb{R}$-valued bounded measurable functions on a Polish space $(\mathbb{Y}, d_{\mathbb{Y}})$. The integral probability metric (IPM) between $P_1, P_2 \in \mathcal{P}(\mathbb{Y})$ is

$$\mathcal{I}_{\mathcal{F}}(P_1, P_2) = \sup_{f\in\mathcal{F}} |P_1(f) - P_2(f)|.$$

A finite IPM is a distance whenever $\mathcal{F}$ separates probabilities on $\mathbb{Y}$, i.e. if $P_1(f) = P_2(f)$ for every $f \in \mathcal{F}$ implies that $P_1 = P_2$. IPMs encompass many well-established distances between probability measures, including the total variation distance, the Maximum Mean Discrepancy (MMD) for a characteristic kernel and the Wasserstein distance of order 1, which is recovered when $\mathcal{F} = \mathrm{Lip}_1(\mathbb{Y}, \mathbb{R})$ is the class of 1-Lipschitz functions on $(\mathbb{Y}, d_{\mathbb{Y}})$ (Sriperumbudur et al., 2012). By duality (see Remark 6.5 in (Villani, 2008)) the Wasserstein distance can also be expressed as an infimum over all couplings as follows. We recall that a coupling between two probabilities P_1, P_2 on a Polish space $(\mathbb{Y}, d_{\mathbb{Y}})$ is the law of any random vector (Y_1, Y_2) on the product space $\mathbb{Y} \times \mathbb{Y}$ such that $Y_1 \sim P_1$ and $Y_2 \sim P_2$. We denote by $\Gamma(P_1, P_2)$ the set of couplings between P_1 and P_2.

Definition 2.2. The Wasserstein distance of order p between $P_1, P_2 \in \mathcal{P}(\mathbb{Y})$ is

$$\mathcal{W}_p(P_1, P_2)^p = \inf_{\gamma\in\Gamma(P^1, P^2)} \mathbb{E}_{(Y_1,Y_2)\sim\gamma}(d_{\mathbb{Y}}(Y_1, Y_2)^p).$$

We focus on $p = 1$ because of its link with IPMs and for simplicity we denote the Wasserstein of order 1 on $\mathbb{X}$ as $\mathcal{W}$. When referring to the Wasserstein distance between the laws of two random variables Y_1, Y_2 we will sometimes omit the law $\mathcal{L}$ and write $\mathcal{W}(Y_1, Y_2) = \mathcal{W}(\mathcal{L}(Y_1), \mathcal{L}(Y_2))$.

When the space of probabilities $\mathcal{P}(\mathbb{X})$ endowed with an IPM is a metric space, the Wasserstein distance over this IPM naturally defines a distance between the laws of random probabilities. For an IPM $\mathcal{I}_{\mathcal{F}}$ we denote by $\mathrm{Lip}_1(\mathcal{I}_{\mathcal{F}}) = \mathrm{Lip}_1((\mathcal{P}(\mathbb{X}), \mathcal{I}_{\mathcal{F}}), \mathbb{R})$ the class of $\mathcal{I}_{\mathcal{F}}$-Lipschitz functions $h : \mathcal{P}(\mathbb{X}) \to \mathbb{R}$ s.t. $|h(P_1) - h(P_2)| \leq \mathcal{I}_{\mathcal{F}}(P_1, P_2)$ for every $P_1, P_2 \in \mathcal{P}(\mathbb{X})$.

Definition 2.3. Let $\mathcal{I}_{\mathcal{F}}$ be an IPM on $\mathbb{X}$. Then for any $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{P}(\mathcal{P}(\mathbb{X}))$ we define

$$\begin{aligned}\mathcal{W}_{\mathcal{F}}(\mathbb{Q}_1, \mathbb{Q}_2) &= \inf_{\gamma\in\Gamma(\mathbb{Q}_1,\mathbb{Q}_2)} \mathbb{E}_{(\tilde{P}_1,\tilde{P}_2)\sim\gamma}(\mathcal{I}_{\mathcal{F}}(\tilde{P}_1, \tilde{P}_2)) \\ &= \sup_{h\in\mathrm{Lip}_1(\mathcal{I}_{\mathcal{F}})} |\mathbb{Q}_1(h) - \mathbb{Q}_2(h)|\end{aligned}$$

We call $\mathcal{W}_{\mathcal{F}}$ the Wasserstein over IPM distance.

For $\mathcal{F} = \mathrm{Lip}_1(\mathbb{X}, \mathbb{R})$, $\mathrm{Lip}_1(\mathcal{I}_{\mathcal{F}}) = \mathrm{Lip}_1(\mathcal{W})$ is the class of Wasserstein-Lipschitz functions on $\mathcal{P}(\mathbb{X})$. We thus recover the Wasserstein over Wasserstein distance, which we denote as $\mathcal{W}_{\mathcal{W}}$. As discussed in the introduction and investigated in Section 4, the statistical properties of $\mathcal{W}_{\mathcal{W}}$ are not satisfactory and for this reason we propose a new, yet related, distance. The main idea is that for a random probability $\tilde{P}$, the laws of the integrals $\tilde{P}(f)$, where f is a continuous bounded function, are probabilities on $\mathbb{R}$ and they are enough to characterize the law of $\tilde{P}$ (cfr. Theorem 4.11 in Kallenberg (2017)).

Definition 2.4. Let $\mathcal{I}_{\mathcal{F}}$ be an IPM on $\mathbb{X}$. Then for any $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{P}(\mathcal{P}(\mathbb{X}))$ we define

$$d_{\mathcal{F}}(\mathbb{Q}_1, \mathbb{Q}_2) = \sup_{f\in\mathcal{F}} \mathcal{W}(\mathcal{L}(\tilde{P}_1(f)), \mathcal{L}(\tilde{P}_2(f))),$$

where $\tilde{P}_i \sim \mathbb{Q}_i$ for $i = 1, 2$. We call $d_{\mathcal{F}}$ the hierarchical integral probability metric (HIPM).

Conditions guaranteeing that $d_{\mathcal{F}}$ is a distance, together with a characterization of the topology it generates, can be found below in Theorem 3.1. In the following we will devote much attention to $\mathcal{F} = \mathrm{Lip}_1(\mathbb{X}, \mathbb{R})$. The corresponding metric $d_{\mathcal{F}}$ will be compactly denoted d_{Lip} and we refer to it as the Lipschitz HIPM.

Note that an analogy can be made with the max-sliced Wasserstein distance (Deshpande et al., 2019) as $d_{\mathcal{F}}$ is the maximal one-dimensional Wasserstein distance between the "projections" of $\mathbb{Q}_1$, $\mathbb{Q}_2$ onto $\mathcal{P}(\mathbb{R})$ via the maps $\mathbb{Q} \mapsto \mathcal{L}(\tilde{P}(f))$, with $\tilde{P} \sim \mathbb{Q}$. Another point of view is to see $d_{\mathcal{F}}$ as an IPM on $\mathcal{P}(\mathbb{X})$: by expressing the Wasserstein distance on $\mathbb{R}$ as an IPM we observe that

$$d_{\mathcal{F}}(\mathbb{Q}_1, \mathbb{Q}_2) = \sup_{h\in\mathfrak{F}} |\mathbb{Q}_1(h) - \mathbb{Q}_2(h)|,$$

where here the class of functions $\mathfrak{F}$ from $\mathcal{P}(\mathbb{X})$ to $\mathbb{R}$ is

$$\mathfrak{F} = \{P \mapsto g(P(f)) \ : \ f \in \mathcal{F} \text{ and } g \in \mathrm{Lip}_1(\mathbb{R})\}.$$

Thus both $\mathcal{W}_{\mathcal{F}}$ and $d_{\mathcal{F}}$ are special cases of IPMs on $\mathcal{P}(\mathbb{X})$, whose corresponding classes of functions, $\mathrm{Lip}_1(\mathcal{I}_{\mathcal{F}})$ and $\mathfrak{F}$ respectively, satisfy $\mathfrak{F} \subset \mathrm{Lip}_1(\mathcal{I}_{\mathcal{F}})$. Intuitively, the class $\mathfrak{F}$ is much smaller than $\mathrm{Lip}_1(\mathcal{I}_{\mathcal{F}})$, and this is explains the better sample complexity of d_{Lip}. This will be precisely quantified in terms of Rademacher complexity in Section 4. On the other hand, under reasonable assumptions both distances metrize the same topology, as discussed in Section 3.

We conclude this section by recalling the paradigmatic example of a random probability: the Dirichlet process (Ferguson, 1973), which will be useful both in the study of the sample complexity of the Wasserstein over Wasserstein distance in Section 4, and in Section 6. We introduce it through its stick-breaking representation (Sethuraman, 1994).

Definition 2.5. A random probability $\tilde{P}$ has a Dirichlet process distribution with concentration $\alpha > 0$ and base probability $P_0 \in \mathcal{P}(\mathbb{X})$, written $\tilde{P} \sim \mathrm{DP}(\alpha, P_0)$, if

$$\tilde{P} \overset{\mathrm{d}}{=} \sum_{i=1}^{+\infty} J_i \delta_{X_i},$$

where $X_i \overset{\text{iid}}{\sim} P_0$ are independent of $V_i \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha)$, and $J_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$. The $\{J_i\}_i$ are termed the stick-breaking weights.

3. Topological and metric properties

In this section we prove that our distance metrizes the desired topology and explain its relation to the more-established Wasserstein over Wasserstein distance. We write $(C_b(\mathbb{X}), \mathrm{L}_\infty)$ for the space of continuous and bounded functions endowed with the supremum norm. We recall that a sequence $(P_n)_n$ of probability laws on a Polish space is said to converge *weakly* to P if for every $f \in C_b(\mathbb{X})$, $P_n(f)$ converges pointwise to $P(f)$. This turns $\mathcal{P}(\mathbb{X})$ into a Polish space, see Remark 7.1.7 in Ambrosio et al. (2008). Thus we can endow $\mathcal{P}(\mathcal{P}(\mathbb{X}))$ with the weak convergence when $\mathcal{P}(\mathbb{X})$ itself is endowed with weak convergence. Remarkably, one can prove (cfr. Prohorov (1961); von Waldenfels (1968); Harris (1971) and Theorem 4.11 in Kallenberg (2017)) that a sequence $(\mathbb{Q}_n)_n$ of laws of random probabilities converges weakly to $\mathbb{Q}$ if and only if for every $f \in C_b(\mathbb{X})$, the sequence of real-valued random variables $\tilde{P}_n(f)$, for $\tilde{P}_n \sim \mathbb{Q}_n$, converges weakly to $\tilde{P}(f)$, for $\tilde{P} \sim \mathbb{Q}$.

Theorem 3.1. *If $\mathcal{I}_{\mathcal{F}}$ is bounded and metrizes the weak convergence on $\mathcal{P}(\mathbb{X})$ then $\mathcal{W}_{\mathcal{F}}$ metrizes the weak convergence on $\mathcal{P}(\mathcal{P}(\mathbb{X}))$. If in addition $\{af + b \,:\, a, b \in \mathbb{R},\, f \in \mathcal{F}\}$ is dense in $(C_b(\mathbb{X}), \mathrm{L}_\infty)$ then $d_{\mathcal{F}}$ also metrizes the weak convergence on $\mathcal{P}(\mathcal{P}(\mathbb{X}))$.*

In the compact case the requirement of density can be checked with the following lemma.

Lemma 3.2. *Assuming $\mathbb{X}$ compact, let $\mathcal{F}$ a class of functions which is absolutely convex (that is, convex and such that $af \in \mathcal{F}$ for any $|a| \leq 1$ and $f \in \mathcal{F}$) and such that $\mathcal{I}_{\mathcal{F}}$ is a distance. Then $\{af + b \,:\, a, b \in \mathbb{R},\, f \in \mathcal{F}\}$ is dense in $(C_b(\mathbb{X}), \mathrm{L}_\infty)$.*

Note that we can always inflate $\mathcal{F}$ so that it becomes absolutely convex but the IPM $\mathcal{I}_{\mathcal{F}}$ stays the same: see Theorem 3.3 in Müller (1997). As examples, d_{Lip} metrizes weak convergence if $\mathbb{X}$ is compact and so does $d_{\mathcal{F}}$ if $\mathcal{I}_{\mathcal{F}}$ is a MMD distance for a c-universal kernel (by definition of universality, see Sriperumbudur et al. (2011)).

We move on to the study of analytical upper and lower bounds of these distances. For a random probability $\tilde{P}$ we denote by $\mathbb{E}(\tilde{P})$ the deterministic mean measure which satisfies $\mathbb{E}(\tilde{P})(A) = \mathbb{E}(\tilde{P}(A))$ for any measurable set A.

Proposition 3.3. *Let $\mathcal{W}_{\mathcal{F}}$ denote the Wasserstein over $\mathcal{I}_{\mathcal{F}}$ distance. Then if $\tilde{P}_i \sim \mathbb{Q}_i$ for $i = 1, 2$,*

$$\mathcal{I}_{\mathcal{F}}(\mathbb{E}(\tilde{P}_1), \mathbb{E}(\tilde{P}_2)) \leq d_{\mathcal{F}}(\mathbb{Q}_1, \mathbb{Q}_2) \leq \mathcal{W}_{\mathcal{F}}(\mathbb{Q}_1, \mathbb{Q}_2).$$

It is easy to see that the inequalities are tight. Indeed, when $\mathbb{Q}_1 = \delta_{P_1}$, and $\mathbb{Q}_2 = \delta_{P_2}$ for some (deterministic) probability measures P_1, P_2,

$$\mathcal{I}_{\mathcal{F}}(P_1, P_2) = d_{\mathcal{F}}(\mathbb{Q}_1, \mathbb{Q}_2) = \mathcal{W}_{\mathcal{F}}(\mathbb{Q}_1, \mathbb{Q}_2).$$

In particular, for $\mathcal{F} = \mathrm{Lip}_1(\mathbb{X}, \mathbb{R})$ our distance d_{Lip} is always a lower bound to the Wasserstein over Wasserstein distance $\mathcal{W}_{\mathcal{W}}$. The next result provides a general condition for equality to hold. We recall that $T : \mathbb{X} \to \mathbb{X}$ is an optimal transport map if its graph is a c-cyclically monotone subset of $\mathbb{X} \times \mathbb{X}$, in the sense of Definition 5.1 in Villani (2008). Moreover, if $P \in \mathcal{P}(\mathbb{X})$, $T_{\#}P(A) = P(T^{-1}(A))$ is the push-forward measure of P by the map T. In the following we write $\mathcal{W}_{\mathcal{W}}(\tilde{P}_1, \tilde{P}_2)$ in place of $\mathcal{W}_{\mathcal{W}}(\mathcal{L}(\tilde{P}_1), \mathcal{L}(\tilde{P}_2))$.

Lemma 3.4. *Let $T : \mathbb{X} \mapsto \mathbb{X}$ be an optimal transport map. If $T_{\#}\tilde{P}_1 \overset{d}{=} \tilde{P}_2$,*

$$\mathcal{W}_{\mathcal{W}}(\tilde{P}_1, \tilde{P}_2) = d_{\mathrm{Lip}}(\tilde{P}_1, \tilde{P}_2) = \mathcal{W}(\mathbb{E}(\tilde{P}_1), \mathbb{E}(\tilde{P}_2)).$$

We now state a corollary of Lemma 3.4 that shows its far reach. Let $\tilde{P}_1$ and $\tilde{P}_2$ be two discrete random probabilities

$$\tilde{P}_1 = \sum_{j \geq 1} J_j^{(1)} \delta_{X_j^{(1)}}, \qquad \tilde{P}_2 = \sum_{j \geq 1} J_j^{(2)} \delta_{X_j^{(2)}},$$

such that (a) the jumps are independent of the atoms, i.e., $(J_j^{(i)})_{j \geq 1} \perp (X_j^{(i)})_{j \geq 1;}$ for $i = 1, 2$; (b) the atoms are i.i.d. from base distributions P_1, P_2 respectively, i.e. $X_j^{(i)} \overset{\text{iid}}{\sim} P_i$ for $i = 1, 2$; (c) the jump distribution of $\tilde{P}_1$ and $\tilde{P}_2$ are the same, i.e. $(J_j^{(1)})_{j \geq 1} \overset{\mathrm{d}}{=} (J_j^{(2)})_{j \geq 1.}$.

Theorem 3.5. *Let $\tilde{P}_1$ and $\tilde{P}_2$ be s.t. (a), (b), (c) hold. Then,*

$$\mathcal{W}_{\mathcal{W}}(\tilde{P}_1, \tilde{P}_2) = d_{\mathrm{Lip}}(\tilde{P}_1, \tilde{P}_2) = \mathcal{W}(P_1, P_2).$$

Theorem 3.5 holds both for probabilities with a finite and an infinite number of atoms. In particular, the assumptions (a) and (b) are the ones that define the notable class of *species sampling processes* (Pitman, 1996). This is a very general class of random probabilities which encompasses, e.g., empirical measures, the Dirichlet process, the Pitman-Yor process (Pitman & Yor, 1997), homogeneous normalized completely random measures (Regazzini et al., 2003): Theorem 3.5 applies to all these examples. The contribution is two-fold: on the one hand it shows that in many cases our distance coincides with $\mathcal{W}_{\mathcal{W}}$, making it a good alternative, on the other it provides the exact expression of $\mathcal{W}_{\mathcal{W}}$ in many interesting scenarios. We point out two useful corollaries.

Corollary 3.6. *Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} P_1$ and $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} P_2$. Then,*

$$\mathcal{W}_{\mathcal{W}}\left(\mathcal{L}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right), \mathcal{L}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{Y_i}\right)\right) = \mathcal{W}(P_1, P_2).$$

Corollary 3.6 highlights the crucial role of the dependence across samples $\{X_i\}_i$ and $\{Y_i\}_i$ in estimating $\mathcal{W}(P_1, P_2)$. Whereas it is known that with independent samples the bias goes slowly to zero when n increases and suffers the curse of dimensionality, considering perfectly coupled atoms removes the bias. We thus expect partial forms of dependence across samples to also reduce the bias.

The next result focuses on the Dirichlet process (Definition 2.5).

Corollary 3.7. *For $P_1, P_2 \in \mathcal{P}(\mathbb{X})$ and $\alpha > 0$,*

$$\mathcal{W}_{\mathcal{W}}(\mathrm{DP}(\alpha, \mathrm{P}_1), \mathrm{DP}(\alpha, \mathrm{P}_2)) = \mathcal{W}(\mathrm{P}_1, \mathrm{P}_2).$$

This recovers the prominent identity in Nguyen (2016, Lemma 3.1). The author addresses it as a 'remarkable identity of the Dirichlet process' - our result extends it to all species sampling models.

Another useful feature of $\mathcal{F} = \mathrm{Lip}_1(\mathbb{R}, \mathbb{R})$ is that the identity function $f(x) = x$ belongs to $\mathcal{F}$. Then a natural lower bound to d_{Lip}, and thus $\mathcal{W}_{\mathcal{W}}$, is the standard Wasserstein distance between the random means: for $\tilde{P}_i \sim \mathbb{Q}_i$,

$$\mathcal{W}\left(\int_{\mathbb{R}} x \,\mathrm{d}\tilde{P}_1(x), \int_{\mathbb{R}} x \,\mathrm{d}\tilde{P}_2(x)\right) \leq d_{\mathrm{Lip}}(\mathbb{Q}_1, \mathbb{Q}_2). \quad (1)$$

4. Sample complexity

Let $\tilde{P}_1, \dots, \tilde{P}_n \overset{\text{iid}}{\sim} \mathbb{Q}$ and consider the empirical estimator

$$\tilde{\mathbb{Q}}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\delta_{\tilde{P}_i}.$$

We have introduced a third level of randomness: $\tilde{\mathbb{Q}}_{(n)}$ is a random variable taking values in $\mathcal{P}(\mathcal{P}(\mathbb{X}))$, with the randomness coming from $\tilde{P}_1, \dots, \tilde{P}_n$. The extension of Glivenko-Cantelli theorem to general Polish spaces (Varadarajan, 1958) ensures that $\tilde{\mathbb{Q}}_{(n)}$ converges weakly to $\mathbb{Q}$ almost surely (a.s.). By Theorem 3.1, both $\mathcal{W}_{\mathcal{W}}(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q})$ and $d_{\mathrm{Lip}}(\mathbb{Q}_{(n)}, \mathbb{Q})$ converge to zero a.s. as the sample size n diverges. Yet, we prove that there is a crucial difference in their convergence rate: whereas the convergence of the former is slower than polynomial, the latter can achieve parametric rate of convergence.

Theorem 4.1. *Let $\mathbb{X}$ a bounded subset of $\mathbb{R}^d$. Then there exist constants C, N depending on d and $\mathrm{diam}(\mathbb{X})$ such that, for $n \geq N$, for any $\mathbb{Q} \in \mathcal{P}(\mathcal{P}(\mathbb{X}))$*

$$\mathbb{E}(\mathcal{W}_{\mathcal{W}}(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q})) \leq C\frac{\log(\log(n))}{\log(n)}.$$

Moreover let $\mathbb{Q} = \mathrm{DP}(\alpha, P_0)$ be the law of a Dirichlet process for $\alpha > 0$ and $P_0 \in \mathcal{P}(\mathbb{X})$ whose support has a non-empty interior. Then for every $\gamma > 0$ there exists $c_\gamma > 0$ such that, at least for n large enough,

$$\mathbb{E}(\mathcal{W}_{\mathcal{W}}(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q})) \geq c_\gamma n^{-\gamma}.$$

On the other hand we show that $d_{\mathcal{F}}$ has parametric convergence rate when $\mathbb{X} \subseteq \mathbb{R}$ under some uniformly bounded condition on $\mathcal{F}$ with respect to the supremum norm. We give a slightly weaker condition in terms of $\mathcal{F}^* = \{f^* = f - f(x_0) \text{ s.t. } f \in \mathcal{F}\}$ for a fixed $x_0 \in \mathbb{X}$. This will allow one to treat the case $\mathcal{F} = \mathrm{Lip}_1(\mathbb{X}, \mathbb{R})$, as $\mathrm{Lip}_1^*(\mathbb{X}, \mathbb{R})$ is uniformly bounded whereas $\mathrm{Lip}_1(\mathbb{X}, \mathbb{R})$ contains all constant functions.

Lemma 4.2. *Let $\mathcal{F}^* = \{f^* = f - f(x_0) \text{ s.t. } f \in \mathcal{F}\}$ be uniformly bounded by K in the supremum norm, for a fixed $x_0 \in \mathbb{X}$. Then for every $\mathbb{Q} \in \mathcal{P}(\mathcal{P}(\mathbb{X}))$,*

$$\mathbb{E}(d_{\mathcal{F}}(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q})) \leq \frac{320\log(2)K}{\sqrt{n}} + \inf_{\epsilon>0}\left\{4\epsilon + \frac{64}{\sqrt{n}}\int_{\epsilon/4}^{K}\sqrt{\log N\left(\frac{\delta}{2}; \mathcal{F}^*, \mathrm{L}_\infty\right)}\,\mathrm{d}\delta\right\},$$

where $N(\cdot; \mathcal{F}^*, \mathrm{L}_\infty)$ is the covering number of $\mathcal{F}^*$ with respect to the supremum norm.

Lemma 4.2 reduces the convergence rate of the HIPM to the convergence rate of the corresponding IPM (the infimum in the right hand side) as the first term coming from the "hierarchical" part of the distance goes to zero at the parametric rate $1/\sqrt{n}$. The infimum is already well-studied for a variety of classes, in particular for $\mathcal{F} = \mathrm{Lip}_1(\mathbb{X}, \mathbb{R})$ when $\mathbb{X}$ is a subset of $\mathbb{R}^d$.

Theorem 4.3. *Let $\mathcal{F} = \mathrm{Lip}_1(\mathbb{X}, \mathbb{R})$ with $\mathbb{X}$ a bounded subset of $\mathbb{R}^d$. Then there exists a constant $C_d > 0$ depending*

on d and $\mathrm{diam}(\mathbb{X})$ but not on n, such that, at least for n large enough,

$$\mathbb{E}(d_{\mathrm{Lip}}(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q})) \leq \begin{cases} C_1 n^{-1/2} & \text{if } d = 1, \\ C_2 n^{-1/2}\log(n) & \text{if } d = 2, \\ C_d n^{-1/d} & \text{if } d > 2. \end{cases}$$

In practice, to compute the HIPM distance between two empirical distributions we need that the realizations $\tilde{P}_1, \dots, \tilde{P}_n \overset{\text{iid}}{\sim} \mathbb{Q}$ are almost surely discrete and with a finite number of atoms. When this is not the case we may approximate $\tilde{P}_i$ through the empirical distribution of an exchangeable sequence whose de Finetti measure is $\tilde{P}_i$, namely $X_1^{(i)}, \dots, X_m^{(i)} | \tilde{P}_i \overset{\text{iid}}{\sim} \tilde{P}_i$. By de Finetti's theorem $\tilde{P}_{i,(m)} = \frac{1}{m}\sum_{j=1}^m \delta_{X_j^{(i)}} \to \tilde{P}_i$ weakly almost surely. A hierarchical empirical estimator can then be defined as

$$\tilde{\mathbb{Q}}_{(n,m)} = \frac{1}{n}\sum_{i=1}^n \delta_{\tilde{P}_{i,(m)}}. \tag{2}$$

Theorem 4.4. *Let $\mathcal{F} = \mathrm{Lip}_1(\mathbb{X}, \mathbb{R})$ with $\mathbb{X}$ a bounded subset of $\mathbb{R}^d$. Then there exist a constant $C_d > 0$, not depending on n or m, such that $\mathbb{E}(d_{\mathrm{Lip}}(\tilde{\mathbb{Q}}_{(n,m)}, \mathbb{Q}))$ is smaller or equal to*

$$\begin{cases} C_1(n^{-1/2} + m^{-1/2}) & \text{if } d = 1, \\ C_2(n^{-1/2}\log(n) + m^{-1/2}\log(m)) & \text{if } d = 2, \\ C_d(n^{-1/d} + m^{-1/d}) & \text{if } d > 2. \end{cases}$$

5. Numerical estimation of the distances

In this section we focus on random measures that are fully discrete. That is, we consider $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{P}(\mathcal{P}(\mathbb{X}))$ such that

$$\mathbb{Q}_k = \frac{1}{n}\sum_{i=1}^n \delta_{P_i^k},$$

for $k = 1, 2$. Here n is the number of distinct probabilities in the support of $\mathbb{Q}_1$, $\mathbb{Q}_2$. Each of the probabilities P_i^k is also discrete and reads

$$P_i^k = \frac{1}{m}\sum_{i=1}^m \omega_{i,j}^k \delta_{X_{i,j}^k} \qquad \text{with } \sum_{j=1}^m \omega_{i,j}^k = 1. \tag{3}$$

Thus m is the number of atoms in each probability and we allow for non-uniform weights $\omega_{i,j}^k$ for each atom as it will be useful in Section 6. The measures $\mathbb{Q}_1$, $\mathbb{Q}_2$ can each be summarized as a $n \times m \times 2$ array containing the atoms $X_{i,j}^k$ and the weights $\omega_{i,j}^k$. Extensions to different values of n for $\mathbb{Q}_1$ and $\mathbb{Q}_2$ and to a number m of atoms which may depend on $i \in \{1, \dots, n\}$ are possible, but we stick to this setting for the sake of simplicity. In this section we discuss how to output (an approximation of) $\mathcal{W}_{\mathcal{W}}(\mathbb{Q}_1, \mathbb{Q}_2)$ or $d_{\mathrm{Lip}}(\mathbb{Q}_1, \mathbb{Q}_2)$.

Wasserstein over Wasserstein. We follow Definition 2.3: we first need to compute the pairwise Wasserstein distance between $P_{i_1}^1$ and $P_{i_2}^2$ for any i_1, i_2, and then solve the $n \times n$ optimal transport problem whose cost matrix is given by the pairwise Wasserstein distances. This was already used e.g. in Yurochkin et al. (2019) and we refer to Appendix B for the pseudocode. The only requirement is a package to solve the optimal transport problem. Assuming that the space $\mathbb{X}$ is a subset of $\mathbb{R}$, so that the inner optimal transport problems are just sorting problems, solving the $n \times n$ optimal transport exactly requires $O(n^3\log(n))$ operations (Orlin, 1997). Thus this algorithm outputs the exact distance in $O(n^2 m\log(m) + n^3\log(n))$ operations.

Lipschitz HIPM in dimension one. We approximate our new distance $d_{\mathrm{Lip}}(\mathbb{Q}_1, \mathbb{Q}_2)$, as in Definition 2.4 with $\mathcal{F} = \mathrm{Lip}_1(\mathbb{X}, \mathbb{R})$. We need to find a supremum over the space of Lipschitz functions $\mathrm{Lip}_1(\mathbb{X}, \mathbb{R})$, and we will resort to a gradient ascent algorithm. We restrict $\mathbb{X}$ to be one-dimensional, so that $\mathbb{X} = [a, b]$. Higher dimensions would require a more careful parametrization of $\mathrm{Lip}_1(\mathbb{X}, \mathbb{R})$, e.g. with neural networks, but we leave this investigation for future work.

To evaluate $P_i^k(f)$ for $k = 1, 2$ and $i = 1, \dots, n$, we only need the values of f on $\{X_{i,j}^k\}$, that is on $2nm$ points. The requirement $f \in \mathrm{Lip}(\mathbb{X}, \mathbb{R})$ corresponds to $2nm - 1$ constraints of the form $|f(X) - f(X')| \leq |X - X'|$ for consecutive atoms X, X'. As we will run a gradient ascent with the function f as unknown, this can become quickly prohibitive. For this reason, we will rather project the atoms of each measure on a fixed grid of stepsize Δx. Let M be the number of grid points and $Y_1, \dots, Y_M$ be the grid points over $[a, b]$. We consider P_i^k given by

$$P_i^k = \sum_{q=1}^M \omega_{i,q}^k \delta_{Y_q}, \qquad \text{with } \sum_{q=1}^M \omega_{i,q}^k = 1, \tag{4}$$

which is a specific instance of (3) with common atoms. Each random measure is now described by a nM array of weights $\omega_{i,q}^1$ and $\omega_{i,q}^2$. If P_i^k is given by the form (3), by projecting each atom $X_{i,j}^k$ onto the closest point on the grid $Y_1, \dots, Y_M$, it is possible to approximate it with a probability measure of the form (4), up to an error Δx in Wasserstein distance. As we would anyway have a statistical error of size $n^{-1/2} + m^{-1/2}$ if we are in the setting of Section 4 (see Theorem 4.4), the rationale is that adding an error Δx is reasonable if $\Delta x \asymp n^{-1/2} + m^{-1/2}$, that is, $M \asymp \min(n^{1/2}, m^{1/2})$.

We now expand Definition 2.4 in this setting. By Birkhoff theorem we rewrite the optimal transport problem as an infimum over permutations, and the optimal permutation can be found very efficiently by sorting; see Remark 2.2.27

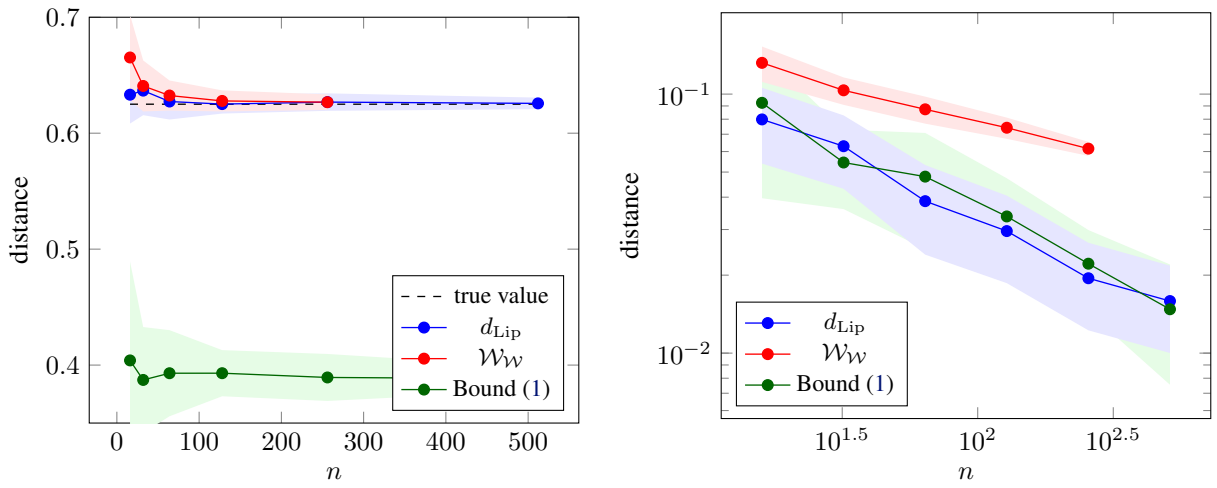

Figure 1. Distances between independent realizations of $\tilde{\mathbb{Q}}_{(n,m)}$ and $\tilde{\mathbb{Q}}'_{(n,m)}$ in (2). We fix $m = 5000$ and let n evolve. Computations are repeated over 24 realizations, with errors corresponding to one standard deviation. In the left plot $\mathbb{Q} = \mathrm{DP}(1, \mathrm{P}_1)$ and $\mathbb{Q}' = \mathrm{DP}(1, \mathrm{P}_2)$ with $P_1 = \mathrm{Unif}([-1/2, 1/2])$, $P_2 = 1/2\,\mathrm{Unif}([-1, -3/4]) + 1/2\,\mathrm{Unif}([3/4, 1])$ so that $\mathcal{W}_{\mathcal{W}}(\mathbb{Q}, \mathbb{Q}') = \mathcal{W}(P_1, P_2) = 5/8$. In the right plot (in log-log scale) $\mathbb{Q} = \mathbb{Q}' = \mathrm{DP}(1, \mathrm{Unif}[0, 1])$.

in Peyré & Cuturi (2019). We replace the function $f \in \mathrm{Lip}_1([a, b], \mathbb{R})$ by a vector $\mathbf{f} \in \mathbb{R}^M$ corresponding to $\mathbf{f}_q = f(Y_q)$ the evaluation of f on grid points. The distance can be rewritten as

$$\sup_{\mathbf{f} \in \mathbb{R}^M} \mathcal{G}(\mathbf{f})$$
$$\text{such that } |\mathbf{f}_{q+1} - \mathbf{f}_q| \leq \Delta x \ \forall q \in \{1, \dots, M-1\},$$

where, by denoting $\mathcal{S}(n)$ the permutations of $\{1, \dots, n\}$,

$$\mathcal{G}(\mathbf{f}) := \inf_{\sigma \in \mathcal{S}(n)} \frac{1}{n} \sum_{i=1}^{n} \left| \sum_{q=1}^{M} (\omega^1_{i,q} - \omega^2_{\sigma(i),q}) \mathbf{f}_q \right|.$$

The constraints correspond to $\mathbf{f}$ being the restriction of a function in $\mathrm{Lip}_1(\mathbb{X}, \mathbb{R})$ and they define a convex set. The function $\mathcal{G}$ we maximize is piece-wise linear. Being an infimum of convex functions, it is neither convex nor concave: we have left the realm of convex optimization. However, the gradient can be easily found, as explained in Appendix B. Thus we use a gradient ascent algorithm to perform maximization of $\mathcal{G}$, but we have no guarantee of finding a global maximizer (in practice, we start with different random initializations). As the function is piece-wise linear there is no canonical stepsize for the gradient ascent, thus we implemented the following: starting from the gradient as an ascent direction, we first project it orthogonally on the set of ascent directions that preserve the convex constraint. Along this new direction, we perform a line search to move enough for the function to increase, and we stop if the increase is too small. We also do a linear change of variables and rather parametrize the function $\mathbf{f}$ by its derivative $\mathbf{g}$ (that is $\mathbf{f}_q = \sum_{q' < q} \Delta x\, \mathbf{g}_{q'}$ for any q) to simplify the Lipschitz constraint in a box constraint. The details and pseudocode can be found in Appendix B.

Experiments. To illustrate our code and the results of Section 4 we consider the setting where the discrete measures $\mathbb{Q}_1$ and $\mathbb{Q}_2$ are realizations of the empirical hierarchical estimators $\tilde{\mathbb{Q}}_{(n,m)}$, $\tilde{\mathbb{Q}}'_{(n,m)}$ in (2) of two Dirichlet processes, $\mathbb{Q}$ and $\mathbb{Q}'$ respectively.

We first consider $\mathbb{Q} = \mathrm{DP}(P_1, \alpha)$ and $\mathbb{Q}' = \mathrm{DP}(P_2, \alpha)$ so that by Corollary 3.7 $\mathcal{W}_{\mathcal{W}}(\mathbb{Q}, \mathbb{Q}') = d_{\mathrm{Lip}}(\mathbb{Q}, \mathbb{Q}') = \mathcal{W}(P_1, P_2)$. We fix m large and check that, as n increases, the distances between realizations of $\tilde{\mathbb{Q}}_{(n,m)}$ and $\tilde{\mathbb{Q}}'_{(n,m)}$ indeed converge towards the true value. We also report the lower bound (1) for completeness, and we observe that in this scenario it is informative but it is not tight. The results are displayed in the left plot of Figure 1.

To further emphasize the difference between $\mathcal{W}_{\mathcal{W}}$ and d_{Lip}, we also consider the case where $\mathbb{Q}_1$ and $\mathbb{Q}_2$ are independent realizations of the same $\tilde{\mathbb{Q}}_{(n,m)}$, where $\mathbb{Q}$ is taken to be a Dirichlet process. We expect $\mathcal{W}_{\mathcal{W}}(\mathbb{Q}_1, \mathbb{Q}_2)$ and $d_{\mathrm{Lip}}(\mathbb{Q}_1, \mathbb{Q}_2)$ to go to zero. We fix m large and study the convergence as $n \to +\infty$. We see that convergence is faster for d_{Lip} than for $\mathcal{W}_{\mathcal{W}}$ as predicted by our theory of Section 4. We also report the lower bound (1), and observe that it almost coincides with d_{Lip} in this case. The results are reported in the right plot of Figure 1.

The algorithms were implemented in Julia. Computations were performed on the CPU of a standard 6-core laptop and we report execution time in the appendix. As computations

were repeated on multiple realizations of $\tilde{\mathbb{Q}}_{n,m}$ and $\tilde{\mathbb{Q}}'_{n,m}$, we only keep cases where the distance for a single realization can be computed in less than one minute. This explains why some values of n are missing for $\mathcal{W}_{\mathcal{W}}$ in Figure 1. The code is available at the following address:

https://github.com/HugoLav/HierarchicalIPM

6. Application: Approximation errors for the Dirichlet process

We consider three popular finite-dimensional approximations of the Dirichlet process (Definition 2.5). Each approximation is indexed by a finite number of atoms N and, as N goes to $+\infty$, it converges to a $\mathrm{DP}(\alpha, \mathrm{P}_0)$.

1. Dirichlet multinomial process

$$\tilde{P}_1 | X_1, \dots, X_N \sim \mathrm{DP}\left(\alpha, \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \delta_{\mathrm{X_i}}\right), \quad \mathrm{X_i} \overset{\text{iid}}{\sim} \mathrm{P}_0.$$

   It is equivalent to $\tilde{P}_1 = \sum_{i=1}^N J_i \delta_{X_i}$ where the jumps $(J_1, \dots, J_N) \in \mathbb{R}^N$ follow a Dirichlet distribution of parameter $(\alpha/N, \dots, \alpha/N)$ while the atoms are i.i.d. with law P_0.

2. Truncated stick-breaking process

$$\tilde{P}_2 = \sum_{i=1}^{N-1} J_i \delta_{X_i} + \left(1 - \sum_{i=1}^{N-1} J_i\right) \delta_{X_N},$$

   where $X_i \overset{\text{iid}}{\sim} P_0$ and $J_1, \dots, J_{N-1}$ are the first $N-1$ stick-breaking weights (see Definition 2.5).

3. The dependent or hierarchical empirical measure

$$\tilde{P}_3 = \frac{1}{N} \sum_{i=1}^{N} \delta_{Y_i},$$

   where $Y_1, \dots, Y_N | \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}$ and $\tilde{P} \sim \mathrm{DP}(\alpha, \mathrm{P}_0)$. We already considered it in Section 4.

The empirical measure is less established in the Bayesian nonparametric literature as a finite-dimensional approximation of the Dirichlet process. Its atoms are exchangeable but not independent and they can be sampled through the Pólya urn or Chinese restaurant process scheme (Blackwell & MacQueen, 1973). In the following we use both analytical and empirical arguments to show that in many regimes it provides similar approximation errors to the widely used multinomial Dirichlet process, which has been independently defined by many different authors (see Section 4 of Ishwaran & Zarepour (2002)). We state our results in terms of upper bounds of $\mathcal{W}_{\mathcal{W}}$ which, thanks to Proposition 3.3, are also upper bounds of our distance d_{Lip}.

Proposition 6.1. *If $\tilde{P} \sim \mathrm{DP}(\alpha, \mathrm{P}_0)$ with $P_0 \in \mathcal{P}(\mathbb{X})$,*

$$\mathcal{W}_{\mathcal{W}}(\tilde{P}, \tilde{P}_1) \leq \mathbb{E}\left(\mathcal{W}\left(P_0, \frac{1}{N}\sum_{i=1}^N \delta_{X_i}\right)\right),$$

$$\mathcal{W}_{\mathcal{W}}(\tilde{P}, \tilde{P}_2) \leq \mathbb{E}(d_{\mathbb{X}}(X_1, X_2))\left(\frac{\alpha}{\alpha+1}\right)^N,$$

$$\mathcal{W}_{\mathcal{W}}(\tilde{P}, \tilde{P}_3) \leq \mathbb{E}_{\tilde{P}}\left(\mathcal{W}\left(\tilde{P}, \frac{1}{N}\sum_{i=1}^N \delta_{Y_i}\right)\right),$$

where $X_1, \dots, X_N \overset{\text{iid}}{\sim} P_0$ and $Y_1, \dots, Y_N | \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}$.
In particular if $P_0 \in \mathcal{P}(\mathbb{R})$ and F_0 denotes its c.d.f.,

$$\mathcal{W}_{\mathcal{W}}(\tilde{P}, \tilde{P}_1) \leq \frac{1}{\sqrt{N}} \int_{-\infty}^{+\infty} \sqrt{F_0(x)(1-F_0(x))}\mathrm{d}x,$$

$$\mathcal{W}_{\mathcal{W}}(\tilde{P}, \tilde{P}_2) \leq 2\left(\frac{\alpha}{\alpha+1}\right)^N \int_{-\infty}^{+\infty} F_0(x)(1-F_0(x))\mathrm{d}x,$$

$$\mathcal{W}_{\mathcal{W}}(\tilde{P}, \tilde{P}_3) \leq \sqrt{\frac{\alpha}{N(\alpha+1)}} \int_{-\infty}^{+\infty} \sqrt{F_0(x)(1-F_0(x))}\mathrm{d}x.$$

Proposition 6.1 sheds light on a number of interesting properties. The Dirichlet multinomial process has polynomial convergence rate in the number of atoms N, whereas the stick-breaking has exponential convergence rate. However, this rate depends on α and for a diverging sequence of α it may fail to converge. The exponential convergence of the stick-breaking approximation can also be captured by moment summaries of the truncation error, as described e.g. in Ishwaran & Zarepour (2000), where it is also underlined that "there appears to be no simple method for assessing the adequacy" of the Dirichlet multinomial approximation. Moreover, the potentially critical regime for the stick-breaking ($\alpha \to +\infty$) is rarely mentioned in the literature. This is not an uncommon situation since if we consider n conditionally i.i.d. observations with de Finetti measure $\tilde{P}$, the posterior is a Dirichlet process with concentration parameter $\alpha + n$, which thus diverges $\asymp n$. In particular, when approximating the posterior arising from a sequence of conditionally i.i.d. observations with a Dirichlet process prior one should always choose N diverging faster than the number of observations n, which can be a problem with large sample sizes. On the contrary, the approximation of the Dirichlet multinomial and of the empirical measures do not show this high dependence on α but they are negatively affected by an increasing dimension of the space. Interestingly, the upper bounds for the Dirichlet multinomial and the hierarchical empirical measure are quite similar, and this similarity is also confirmed by the simulations.

Practical conclusions. Our study led to the following practical advice when using an approximation of the Dirichlet process. If one is approximating a Dirichlet processes with large concentration parameter, as in the case of posteriors for a large dataset arising from a conditionally i.i.d.

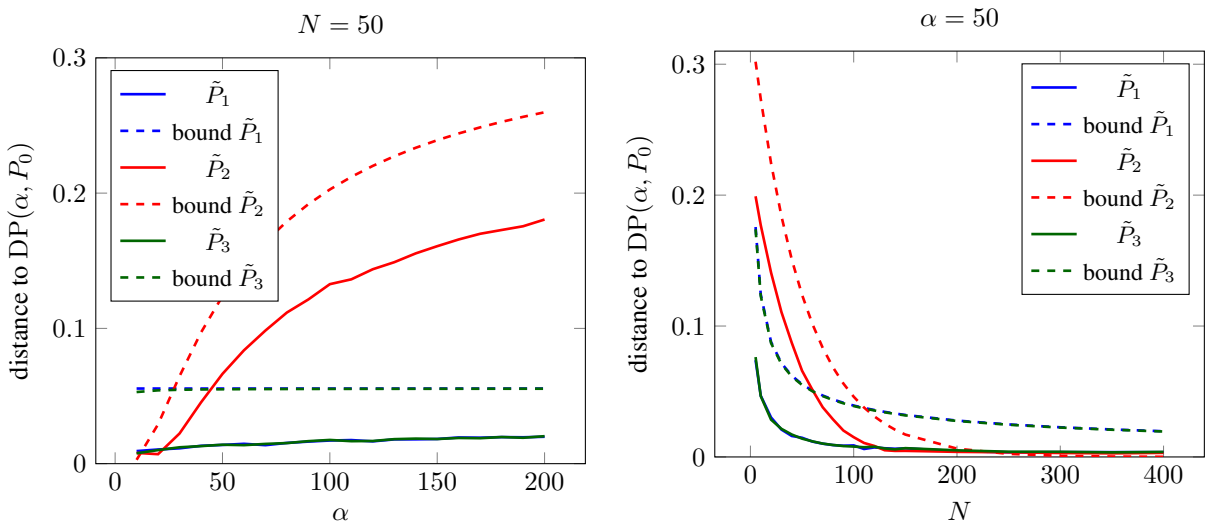

Figure 2. Distance from a $\mathrm{DP}(\alpha, \mathrm{P}_0)$ for finite-dimensional approximations with N atoms: Dirichlet multinomial ($\tilde{P}_1$), truncated stick breaking ($\tilde{P}_2$), and hierarchical empirical measure ($\tilde{P}_3$). The base measure P_0 is uniform over $[0, 1]$. In the left plot $N = 50$ is fixed, and α varies; in the right plot $\alpha = 50$ is fixed, and N varies. The solid line is the distance d_{Lip}, while the dashed lines are the upper bounds of Proposition 6.1.

model directed by a Dirichlet process, the Dirichlet multinomial process and the hierarchical empirical measure will provide similar and more reliable approximations than the truncated stick-breaking, as depicted in the left plot of Figure 2. On the other hand if the observations live in a high-dimensional space or if we can afford taking N consistently larger than the concentration parameter, the stick-breaking will tend to provide better approximations, as depicted in the right plot of Figure 2.

Limitations

Our work presents some limitations from the computational and theoretical point of view. Our numerical method is restricted to a one-dimensional space and one should be careful in drawing delicate conclusions based on the numerical evaluation of the distance, as it relies on gradient ascent over a non-concave objective and thus it does not come with a guarantee of convergence. From a theoretical perspective, the sample complexity is derived for probabilities on a space with a bounded metric. Finally, we still lack a good understanding of the qualitative difference between the Wasserstein over Wasserstein distance and the HIPM.

Acknowledgments

The authors thank Stefan Schrott and Giacomo Sodini for some helpful discussions as well as the anonymous referees for their suggestions which contrinute to improve the quality of the paper. The second author is also affiliated to the Bocconi Institute for Data Science and Analytics (BIDSA) and is partially supported by the MUR-Prin 2022-202244A7YL funded by the European Union - Next Generation EU.

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alvarez-Melis, D. and Fusi, N. Geometric dataset distances via optimal transport. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21428–21439. Curran Associates, Inc., 2020.

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Arbel, J., De Blasi, P., and Prünster, I. Stochastic Approximations to the Pitman–Yor Process. *Bayesian Analysis*, 14(4):1201 – 1219, 2019.

Armijo, L. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of mathematics*, 16(1):1–3, 1966.

Bing, X., Bunea, F., and Niles-Weed, J. Estimation and in-

ference for the Wasserstein distance between mixing measures in topic models. *arXiv preprint arXiv:2206.12768*, 2022a.

Bing, X., Bunea, F., Strimas-Mackey, S., and Wegkamp, M. Likelihood estimation of sparse topic distributions in topic models and its applications to Wasserstein document distance calculations. *The Annals of Statistics*, 50(6):3307 – 3333, 2022b.

Blackwell, D. and MacQueen, J. B. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2): 353–355, 1973.

Bobkov, S. G. and Ledoux, M. One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Memoirs of the American Mathematical Society*, 2019.

Boissard, E. and Gouic, T. L. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(2):539 – 563, 2014.

Campbell, T., Huggins, J. H., How, J. P., and Broderick, T. Truncated random measures. *Bernoulli*, 25(2):1256 – 1288, 2019.

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., and Schwing, A. G. Max-sliced Wasserstein Distance and its use for GANs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10648–10656, 2019.

Dudley, R. M. The speed of mean Glivenko-Cantelli Convergence. *The Annals of Mathematical Statistics*, 40(1): 40 – 50, 1969.

Dukler, Y., Li, W., Lin, A., and Montufar, G. Wasserstein of Wasserstein loss for learning generative models. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1716–1725. PMLR, 09–15 Jun 2019.

Ferguson, T. S. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209 – 230, 1973.

Fournier, N. and Guillin, A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738, 2015.

Ghosal, S. and van der Vaart, A. *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press, 2017.

Gibbs, A. L. and Su, F. E. On choosing and bounding probability metrics. *International statistical review*, 70 (3):419–435, 2002.

Harris, T. Random measures and motions of point processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 18:85–115, 1971.

Ho, N., Nguyen, X., Yurochkin, M., Bui, H. H., Huynh, V., and Phung, D. Multilevel clustering via Wasserstein means. In *International conference on machine learning*, pp. 1501–1509. PMLR, 2017.

Ishwaran, H. and James, L. F. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

Ishwaran, H. and Zarepour, M. Markov chain monte carlo in approximate dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390, 2000. ISSN 00063444.

Ishwaran, H. and Zarepour, M. Exact and approximate sum representations for the dirichlet process. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 30(2):269–283, 2002. ISSN 03195724.

Kallenberg, O. *Random Measures, Theory and Applications*. Probability Theory and Stochastic Modelling. Springer, 2017.

Kolmogorov, A. N. and Tikhomirov, V. M. ϵ-entropy and ϵ-capacity of sets in functional spaces. *American Mathematical Society Translations*, 2(17):277–364, 1961.

Lijoi, A. and Regazzini, E. Means of a Dirichlet process and multiple hypergeometric functions. *The Annals of Probability*, 32(2):1469–1495, 2004.

Lijoi, A., Prünster, I., and Rigon, T. The Pitman–Yor multinomial process for mixture modelling. *Biometrika*, 107 (4):891–906, 06 2020. ISSN 0006-3444.

Mena, G. and Niles-Weed, J. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in neural information processing systems*, 32, 2019.

Muliere, P. and Tardella, L. Approximating distributions of random functionals of Ferguson-Dirichlet priors. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 26(2):283–297, 1998.

Müller, A. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29 (2):429–443, 1997.

Nguyen, T. D., Huggins, J., Masoero, L., Mackey, L., and Broderick, T. Independent finite approximations for Bayesian nonparametric inference. *Bayesian Analysis*, pp. 1 – 38, 2023.

Nguyen, X. Borrowing strengh in hierarchical Bayes: Posterior concentration of the Dirichlet base measure. *Bernoulli*, 22(3):1535 – 1571, 2016.

Orlin, J. B. A polynomial time primal network simplex algorithm for minimum cost flows. *Mathematical Programming*, 78:109–129, 1997.

Peyré, G. and Cuturi, M. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Pitman, J. Some developments of the Blackwell-Macqueen urn scheme. *Lecture Notes-Monograph Series*, 30:245–267, 1996.

Pitman, J. and Yor, M. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855 – 900, 1997.

Prohorov, Y. Random measures on a compactum. *Soviet Mathematics Doklady*, 2:539–541, 1961.

Regazzini, E., Lijoi, A., and Prünster, I. Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, 31(2): 560–585, 2003.

Rudin, W. *Real and Complex Analysis*. Mathematics series. McGraw-Hill, 1987. ISBN 9780071002769.

Sethuraman, J. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639 – 650, 1994.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G. Non-parametric estimation of integral probability metrics. In *2010 IEEE International Symposium on Information Theory*, pp. 1428–1432, 2010.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*, 12(7), 2011.

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*, 6:550–1599, 2012.

Vaart, A. v. d. and Wellner, J. A. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

Varadarajan, V. S. On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 19(1/2):23–26, 1958.

Villani, C. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, 2008. ISBN 9783540710509.

von Waldenfels, W. Characteristische funktionale zufaelliger masse. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 10:279–283, 1968.

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Wasserman, L. *All of Nonparametric Statistics (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387251456.

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.

Yurochkin, M., Claici, S., Chien, E., Mirzazadeh, F., and Solomon, J. M. Hierarchical optimal transport for document representation. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Zolotarev, V. M. Probability metrics. *Theory of Probability and its Applications*, 28:278–302, 1984.

A. Proofs

Proof of Theorem 3.1

The distance $\mathcal{W}_{\mathcal{F}}$ metrizes weak convergence. That $\mathcal{W}_{\mathcal{F}}$ metrizes weak convergence over $\mathcal{P}(\mathcal{P}(\mathbb{X}))$ is a simple consequence of Remark 7.1.7 in Ambrosio et al. (2008) and the fact that $(\mathcal{P}(\mathbb{X}), \mathcal{I}_{\mathcal{F}})$ has a bounded diameter by assumption.

The distance $d_{\mathcal{F}}$ metrizes weak convergence. Let $\mathbb{Q}_n$ a sequence in $\mathcal{P}(\mathcal{P}(\mathbb{X}))$, and $\mathbb{Q} \in \mathcal{P}(\mathcal{P}(\mathbb{X}))$. We also write $\tilde{P}_n, \tilde{P}$ for random measures with $\tilde{P}_n \sim \mathbb{Q}_n$ and $\tilde{P} \sim \mathbb{Q}$.

We first assume that $d_{\mathcal{F}}(\mathbb{Q}_n, \mathbb{Q})$ converges to 0 as $n \to +\infty$. In particular for any $g \in \mathcal{F}$, we know that $\mathcal{W}(\tilde{P}_n(g), \tilde{P}(g))$ converges to zero. This easily extends to any g in $\mathcal{F}' = \{af + b \ : \ a, b \in \mathbb{R}, \ f \in \mathcal{F}\}$. We extend to $C_b(\mathbb{X})$ by our density assumption. If $f \in C_b(\mathbb{X})$ is any continuous and bounded function and $\varepsilon > 0$, by assumption we can find $g \in \mathcal{F}'$ with $\|f - g\|_\infty \leq \varepsilon$. With the triangle inequality we easily obtain

$$\mathcal{W}(\tilde{P}_n(f), \tilde{P}(f)) \leq \mathcal{W}(\tilde{P}_n(f), \tilde{P}_n(g)) + \mathcal{W}(\tilde{P}_n(g), \tilde{P}(g)) + \mathcal{W}(\tilde{P}(g), \tilde{P}(f)) \leq \mathcal{W}(\tilde{P}_n(g), \tilde{P}(g)) + 2\varepsilon.$$

As $n \to +\infty$, the first term in the right hand side converges to 0. As ε can then be chosen arbitrary we have proved that $\mathcal{W}(\tilde{P}_n(f), \tilde{P}(f))$ converges to 0 as $n \to +\infty$ for any continuous and bounded function f. Convergence in $\mathcal{W}$ implies converges in law, thus the random variable $\tilde{P}_n(f)$ converges in law to $\tilde{P}(f)$. This concludes the proof of weak convergence of $\mathbb{Q}_n$ to $\mathbb{Q}$.

Conversely assume the weak convergence of $\mathbb{Q}_n$ to $\mathbb{Q}$. Using Proposition 3.3, we know that $d_{\mathcal{F}}(\mathbb{Q}_n, \mathbb{Q}) \leq \mathcal{W}_{\mathcal{F}}(\mathbb{Q}_n, \mathbb{Q})$. The first part of the present theorem yields that the right hand side goes to zero, and thus so does the left hand side.

Proof of Lemma 3.2

The proof strategy relies on the Hahn-Banach theorem and is similar to the one used to prove density results in the context of MMD distances, see e.g. Proposition 2 in Sriperumbudur et al. (2011).

Step 1. Characterization as linear span. Using the absolute convexity of $\mathcal{F}$, we prove that $\mathcal{F}' = \{af+b \ : \ a, b \in \mathbb{R}, \ f \in \mathcal{F}\}$ coincides with the linear span of $\mathcal{F} \cup \{1\}$, where 1 denotes the constant function equal to 1. First we observe that $\mathcal{F}'$ is a vector space. Indeed, for any $a_1, a_2 \in \mathbb{R}$ and $f_1, f_2 \in \mathcal{F}$,

$$a_1 f_1 + a_2 f_2 = (|a_1| + |a_2|) \left(\frac{a_1}{|a_1| + |a_2|} f_1 + \frac{a_2}{|a_1| + |a_2|} f_2 \right)$$

belongs to $\mathcal{F}'$ since by absolute convexity $a_1/(|a_1| + |a_2|)f_1 + a_2/(|a_1| + |a_2|)f_2 \in \mathcal{F}$. It easily follows that for any $a_1, a_2 \in \mathbb{R}$ and $f_1, f_2 \in \mathcal{F}'$, $a_1 f_1 + a_2 f_2 \in \mathcal{F}'$ and thus $\mathcal{F}'$ is a vector space. As in addition $\mathcal{F} \cup \{1\} \subseteq \mathcal{F}'$, we see that $\mathcal{F}'$ also contains the linear span of $\mathcal{F} \cup \{1\}$. The other inclusion easily holds, and thus there is equality.

Step 2. Density of the linear span in $C_b(\mathbb{X})$. By a corollary of the Hahn-Banach theorem, see Theorem 5.19 in Rudin (1987), it is enough to prove that any continuous linear form σ on $(C_b(\mathbb{X}), \mathrm{L}_\infty)$ that vanishes on $\mathcal{F} \cup \{1\}$ is necessarily identically zero. As $\mathbb{X}$ is compact, the Riesz theorem (see e.g. Theorem 2.14 in Rudin (1987)) guarantees that any such σ is characterized by a finite signed measure, which we denote σ as well, i.e. $\sigma(f) = \int_{\mathbb{X}} f \mathrm{d}\sigma$. We write $\sigma = \sigma_+ - \sigma_-$ for its Jordan decomposition into a positive and a negative part. As $\sigma(1) = 0$, we deduce that $\sigma_+(\mathbb{X}) = \sigma_-(\mathbb{X})$. If $\sigma_+(\mathbb{X}) = 0$ then $\sigma_+ = \sigma_- = \sigma = 0$ and we are done. If not, up to dividing σ by $\sigma_+(\mathbb{X})$, we can assume that $\sigma_+(\mathbb{X}) = \sigma_-(\mathbb{X}) = 1$, that is, σ_+ and σ_- are probability distributions over $\mathbb{X}$. Since σ vanishes on $\mathcal{F}$, $\sigma_+(f) - \sigma_-(f) = 0$ for any $f \in \mathcal{F}$, which implies $\mathcal{I}_{\mathcal{F}}(\sigma_+, \sigma_-) = 0$. As $\mathcal{I}_{\mathcal{F}}$ is distance, we obtain $\sigma_+ = \sigma_-$. This implies $\sigma = 0$ and concludes the proof.

Proof of Proposition 3.3

Take $\tilde{P}_1, \tilde{P}_2$ distributed according to $\mathbb{Q}_1$ and $\mathbb{Q}_2$ respectively. Given the definition of the two distances and the linearity of the integral, the property we want to prove is

$$\sup_{f \in \mathcal{F}} |\mathbb{E}(\tilde{P}_1(f)) - \mathbb{E}(\tilde{P}_2(f))| \leq \sup_{h \in \mathfrak{F}} |\mathbb{Q}_1(h) - \mathbb{Q}_2(h)| \leq \sup_{h \in \mathrm{Lip}_1(\mathcal{I}_{\mathcal{F}})} |\mathbb{Q}_1(h) - \mathbb{Q}_2(h)|,$$

where we recall that

$$\mathfrak{F} = \{P \mapsto g(P(f)) \ : \ f \in \mathcal{F} \text{ and } g \in \mathrm{Lip}_1(\mathbb{R})\}.$$

Following the definition we see that $\mathfrak{F} \subset \mathrm{Lip}_1(\mathcal{I}_{\mathcal{F}})$ thus the second inequality holds. Then we notice that for any $f \in \mathcal{F}$ there holds $\mathbb{E}(\tilde{P}_i(f)) = \mathbb{Q}_i(h_f)$ for $i = 1, 2$ with $h_f : P \mapsto P(f) \in \mathfrak{F}$. The first inequality easily follows.

Proof of Lemma 3.4

Let $P_i = \mathbb{E}(\tilde{P}_i)$. The fact that $\mathcal{W}(P_1, P_2) \leq \mathcal{W}_{\mathcal{W}}(\mathcal{L}(\tilde{P}_1), \mathcal{L}(\tilde{P}_2))$ follows from Proposition 3.3 and was also proved in Lemma 3.1 of Nguyen (2016). We prove that $\mathcal{W}(P_1, P_2) \geq \mathcal{W}_{\mathcal{W}}(\mathcal{L}(\tilde{P}_1), \mathcal{L}(\tilde{P}_2))$. Let T be the optimal transport map such that $T_{\#}\tilde{P}_1 \overset{\mathrm{d}}{=} \tilde{P}_2$. Then by considering the coupling $(\tilde{P}_1, T_{\#}\tilde{P}_1)$,

$$\mathcal{W}_{\mathcal{W}}(\mathcal{L}(\tilde{P}_1), \mathcal{L}(\tilde{P}_2)) \leq \mathbb{E}_{\tilde{P}_1}(\mathcal{W}(\tilde{P}_1, T_{\#}\tilde{P}_1)) = \mathbb{E}_{\tilde{P}_1}(\mathbb{E}_{X|\tilde{P}_1 \sim \tilde{P}_1}(|X - T(X)| \mid \tilde{P}_1),$$

since T is an optimal transport map. If $X|\tilde{P}_1 \sim \tilde{P}_1$ then $X \sim \mathbb{E}(\tilde{P}_1) = P_1$. Thus by the tower property, the right hand side is equal to $\mathbb{E}_{X \sim P_1}(|X - T(X)|)$. We observe that $T_{\#}\tilde{P}_1 = \tilde{P}_2$ a.s. implies that $T_{\#}P_1 = P_2$. Indeed, $P_2(A) = \mathbb{E}(\tilde{P}_2(A)) = \mathbb{E}(\tilde{P}_1(T^{-1}(A)) = P_1(T^{-1}(A)) = T_{\#}P_1(A)$. Since T is an optimal transport map, $\mathbb{E}_{X \sim P_1}(|X - T(X)|) = \mathcal{W}(P_1, P_2)$. Thus, $\mathcal{W}_{\mathcal{W}}(\mathcal{L}(\tilde{P}_1), \mathcal{L}(\tilde{P}_2)) \leq \mathcal{W}(P_1, P_2)$ and the conclusion follows.

Proof of Theorem 3.5

If there exists T an optimal transport map between P_1 and P_2, by assumptions (a), (b), and (c) it follows that

$$T_{\#}\tilde{P}_1 = \sum_{j \geq 1} J_j^{(1)} \delta_{T(X_j^{(1)})} \overset{\mathrm{d}}{=} \tilde{P}_2.$$

We can apply Lemma 3.4 and the conclusion follows.

Consider now the case when an optimal transport map T does not exist and let γ be the optimal coupling on $\mathbb{X} \times \mathbb{X}$ between P_1 and P_2. We build the following coupling between $\mathcal{L}(\tilde{P}_1)$ and $\mathcal{L}(\tilde{P}_2)$

$$(\tilde{P}_1, \tilde{P}_2) = \left(\sum_{j \geq 1} J_j \delta_{Y_j^{(1)}}, \sum_{j \geq 1} J_j \delta_{Y_j^{(2)}} \right),$$

where $\{J_j\}$ have the same distribution as $\{J_j^{(1)}\}$ and $\{J_j^{(2)}\}$, and they are independent from $(Y_j^{(1)}, Y_j^{(2)}) \overset{\mathrm{iid}}{\sim} \gamma$. Then

$$\mathcal{W}_{\mathcal{W}}(\tilde{P}_1, \tilde{P}_2) \leq \mathbb{E}\left(\mathcal{W}\left(\sum_{j \geq 1} J_j \delta_{Y_j^{(1)}}, \sum_{j \geq 1} J_j \delta_{Y_j^{(2)}} \right) \right) \leq \mathbb{E}\left(\sum_{j \geq 1} J_j d_{\mathbb{X}}(Y_j^{(1)}, Y_j^{(2)}) \right) = \sum_{j \geq 1} \mathbb{E}(J_j) \mathbb{E}(d_{\mathbb{X}}(Y_j^{(1)}, Y_j^{(2)})).$$

Since $(Y_j^{(1)}, Y_j^{(2)}) \overset{\mathrm{iid}}{\sim} \gamma$ is the optimal coupling between P_1 and P_2,

$$\sum_{j \geq 1} \mathbb{E}(J_j) \mathbb{E}(d(Y_j^{(1)}, Y_j^{(2)})) = \mathcal{W}(P_1, P_2) \mathbb{E}\left(\sum_{j \geq 1} J_j \right) = \mathcal{W}(P_1, P_2).$$

This proves that $\mathcal{W}_{\mathcal{W}}(\tilde{P}_1, \tilde{P}_2) \leq \mathcal{W}(P_1, P_2)$. The other inequality follows from Proposition 3.3.

Proof of Theorem 4.1: the upper bound

Step 1: reducing to an estimation of Rademacher complexity. The quantity of interest can be written, following the definition,

$$\mathbb{E}(\mathcal{W}_{\mathcal{W}}(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q})) = \mathbb{E}\left(\sup_{h \in \mathrm{Lip}_1^*(\mathcal{W})} \left| \tilde{\mathbb{Q}}_{(n)}(h) - \mathbb{Q}(h) \right| \right),$$

where $\mathrm{Lip}_1^*(\mathcal{W})$ is the class of functions defined over $\mathcal{P}(\mathbb{X})$ which are 1-Lipschitz with respect with the Wasserstein distance and vanish on a distinguished measure P_0, which will not play a role. The original definition is stated with $\mathrm{Lip}_1(\mathcal{W})$ but clearly replacing by $\mathrm{Lip}_1^*(\mathcal{W})$ does not change the value. The gain is that the class $\mathrm{Lip}_1^*(\mathcal{W})$ is uniformly bounded as $\mathbb{X}$ is bounded. The classical symmetrization argument (see e.g. Lemma 2.3.1 in Vaart & Wellner (1996) or Section 4.2 in Wainwright (2019)) yields

$$\mathbb{E}(\mathcal{W}_{\mathcal{W}}(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q})) \leq 2\mathcal{R}_n(\mathrm{Lip}_1^*(\mathcal{W})),$$

where the Rademacher complexity of the class $\mathrm{Lip}_1^*(\mathcal{W})$ is defined as

$$\mathcal{R}_n(\mathrm{Lip}_1^*(\mathcal{W})) = \mathbb{E}_{\tilde{\mathrm{P}}_{1:\mathrm{n}}, \epsilon_{1:\mathrm{n}}}\left(\sup_{\mathrm{h} \in \mathrm{Lip}_1^*(\mathcal{W})} \left| \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \epsilon_{\mathrm{i}} \mathrm{h}(\tilde{\mathrm{P}}_{\mathrm{i}}) \right| \right), \tag{5}$$

being $\epsilon_1, \dots, \epsilon_n$ i.i.d. Rademacher random variables independent from $P_1, \dots, P_n$, which are i.i.d. with law $\mathbb{Q}$. Note that we use $\mathbb{E}_{\tilde{P}_{1:n}, \epsilon_{1:n}}$ as a shortcut for $\mathbb{E}_{P_1, \dots, P_n, \epsilon_1, \dots, \epsilon_n}$.

Step 2: estimation of the Rademacher complexity via a covering number. As trying to get the sharpest rates would take us to expressions which are not analytically tractable, we will only use the one step bound rather than the integral bound (see Proposition 5.17 and Example 5.21 in Wainwright (2019)) which yields

$$\mathcal{R}_n(\mathrm{Lip}_1^*(\mathcal{W})) \leq \inf_{\epsilon \in [0, \epsilon_0)} \left\{ \epsilon + \frac{\mathrm{C}_1}{\sqrt{\mathrm{n}}} \sqrt{\log \mathrm{N}\left(\epsilon, \mathrm{Lip}_1^*(\mathcal{W}), \mathrm{L}_\infty\right)} \right\},$$

where ϵ_0, C_1 are two constants. Here $N\left(\epsilon, \mathrm{Lip}_1^*(\mathcal{W}), \mathrm{L}_\infty\right)$ is the covering number of the space of 1-Lipschitz functions defined on $\mathcal{P}(\mathbb{X})$ with respect to L_∞, the supremum norm on the space of functions defined on $\mathcal{P}(\mathbb{X})$. The original result is stated with L_∞ replaced with the L_2 norm with respect to the empirical process, but is always dominated by L_∞. To estimate the covering number we then use the bound of Kolmogorov & Tikhomirov (1961) which is recalled in (3.1) of Sriperumbudur et al. (2012): for a bounded metric space $\mathbb{Y}$ with metric $d_\mathbb{Y}$, the covering number of $\mathrm{Lip}_1^*(\mathbb{Y})$ the one-Lipschitz functions with respect to the uniform norm can be estimated as

$$\log N(\epsilon, \mathrm{Lip}_1^*(\mathbb{Y}), \mathrm{L}_\infty) \leq \mathrm{N}\left(\frac{\epsilon}{4}, \mathbb{Y}, \mathrm{d}_\mathbb{Y}\right) \log\left(2\left\lceil \frac{2\mathrm{diam}(\mathbb{Y})}{\epsilon} \right\rceil + 1\right). \tag{6}$$

Here and in the sequel, $\lceil a \rceil$ is the integral value greater than a. Considering $(\mathbb{Y}, d_\mathbb{Y}) = (\mathcal{P}(\mathbb{X}), \mathcal{W})$ we find

$$\log N\left(\epsilon, \mathrm{Lip}_1^*(\mathcal{W}), \mathrm{L}_\infty\right) \leq N\left(\frac{\epsilon}{4}, \mathcal{P}(\mathbb{X}), \mathcal{W}\right) \log\left(2\left\lceil \frac{2\mathrm{diam}(\mathcal{P}(\mathbb{X}))}{\epsilon} \right\rceil + 1\right).$$

As $\mathrm{diam}(\mathcal{P}(\mathbb{X})) = \mathrm{diam}(\mathbb{X})$, what is left to do is to estimate the metric entropy of $\mathcal{P}(\mathbb{X})$ when endowed with the Wasserstein distance.

Step 3: estimating the covering number of the Wasserstein space. We claim that for any $\epsilon > 0$

$$N\left(\epsilon, \mathcal{P}(\mathbb{X}), \mathcal{W}\right) \leq N\left(\frac{\epsilon}{2}, \mathbb{X}, d_\mathbb{X}\right)^{\lceil 2\mathrm{diam}(\mathbb{X})/\epsilon \rceil + 1}. \tag{7}$$

Indeed, let $x_1, x_2, \dots, x_A$ with $A = N(\epsilon/2, \mathbb{X}, d_\mathbb{X})$ an $\epsilon/2$-covering of the space $\mathbb{X}$. Let $B = \lceil 2\mathrm{diam}(\mathbb{X})/\epsilon \rceil$ and let us consider $\mathcal{Z}$ the subset of $\mathcal{P}(\mathbb{X})$ made of probability measures supported on $x_1, x_2, \dots, x_A$ and such that the mass of each x_j belongs to $\{0, 1/B, \dots, (B-1)/B, 1\}$ for $j = 1, \dots, A$: the cardinality of $\mathcal{Z}$ is bounded by A^{B+1} and we will prove $\mathcal{Z}$ is an ϵ-covering of $\mathcal{P}(\mathbb{X})$.

Take any $P \in \mathcal{P}(\mathbb{X})$. Then there is a probability measure P_1 supported on $x_1, x_2, \dots, x_A$ such that $\mathcal{W}(P, P_1) \leq \epsilon/2$: it is obtained by projecting the mass of P onto the set $x_1, x_2, \dots, x_A$. Then, by rounding up the mass of each atom x_j, we can find P_2 a measure in $\mathcal{Z}$ such that $\|P_1 - P_2\|_{\mathrm{TV}} \leq 1/B$, being $\|\cdot\|_{\mathrm{TV}}$ the total variation norm. Using the easy inequality $\mathcal{W}(P_1, P_2) \leq \mathrm{diam}(\mathbb{X}) \|\mathrm{P}_1 - \mathrm{P}_2\|_{\mathrm{TV}}$ (see Theorem 4 in Gibbs & Su (2002)) followed by the triangle inequality, we find that

$$\mathcal{W}(P, P_2) \leq \mathcal{W}(P, P_1) + \mathcal{W}(P_1, P_2) \leq \frac{\epsilon}{2} + \frac{\mathrm{diam}(\mathbb{X})}{B} \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Thus $\mathcal{Z}$ is an ϵ-covering of $\mathcal{P}(\mathbb{X})$, and given the expression of its cardinality we obtain the estimate (7).

Step 4: conclusion with a well chosen ϵ. Chaining our estimates, we see that we can estimate the Rademacher complexity by

$$\mathcal{R}_n(\mathrm{Lip}_1^*(\mathcal{W})) \leq \inf_{\epsilon \in [0, \epsilon_0)} \left\{ \epsilon + \frac{\mathrm{C}_1}{\sqrt{\mathrm{n}}} \mathrm{N}\left(\frac{\epsilon}{2}, \mathbb{X}, \mathrm{d}_\mathbb{X}\right)^{\lceil 2\mathrm{diam}(\mathbb{X})/\epsilon \rceil/2 + 1/2} \sqrt{\log\left(2\left\lceil \frac{2\mathrm{diam}(\mathcal{P}(\mathbb{X}))}{\epsilon} \right\rceil + 1\right)} \right\}.$$

At this point we will not keep track of the explicit constant anymore. As $\mathbb{X}$ is a bounded set of $\mathbb{R}^d$ using Theorem 2.7.1 in Vaart & Wellner (1996), we have $N\left(\frac{\epsilon}{2}, \mathbb{X}, d_\mathbb{X}\right) \leq C_2 \epsilon^{-d}$ for some constant C_2. Injecting this in our estimate, for some constant C_1, C_3, C_4 large enough and up to decreasing ϵ_0, we get

$$\mathcal{R}_n(\mathrm{Lip}_1^*(\mathcal{W})) \leq \inf_{\epsilon \in [0, \epsilon_0)} \left\{ \epsilon + \frac{\mathrm{C}_1}{\sqrt{\mathrm{n}}} \exp\left(\frac{\mathrm{C}_3}{\epsilon} \log\left(\frac{1}{\epsilon}\right)\right) \sqrt{\log\left(\frac{\mathrm{C}_4}{\epsilon}\right)} \right\}. \tag{8}$$

Exact optimization of this expression in ϵ seems tricky and leads to intractable analytical expressions. Note however that if $\epsilon \lesssim 1/\log(n)$, then the expression $C_3/\epsilon \log(1/\epsilon)$ in the argument of the exponential grows much faster than $\log(n)$, and so the exponential grows much faster than any positive power of n. Thus if $\epsilon \lesssim 1/\log(n)$ the second term in the right hand side of (8) blows up to $+\infty$. So we would look for $\epsilon = \epsilon_n = \beta_n/\log(n)$ with $1 \ll \beta_n \ll \log(n)$, and then we see that the argument of the exponential reads

$$\frac{C_3}{\epsilon} \log\left(\frac{1}{\epsilon}\right) = C_3 \frac{\log(n)}{\beta_n} \left(\log(\log(n)) - \log(\beta_n)\right) \sim C_3 \frac{\log(n)\log(\log(n))}{\beta_n}.$$

For this to grow slower than $\log(n)$ (so that the exponential grows polynomially) we need β_n to grow at least as fast as $\log(\log(n))$. Following this computation we choose

$$\frac{1}{\epsilon} = \frac{1}{3C_3} \frac{\log(n)}{\log(\log(n))},$$

which we can do for n large enough, and it will be the best we can do to bound the right hand side of (8). We claim that the second term in the right hand side of (8) is negligible compare to the first one $\epsilon \asymp \log(\log(n))/\log(n)$. Indeed looking at the argument of the exponential, as $1/\epsilon \leq \log(n)$ for n large enough, we have

$$\frac{C_3}{\epsilon} \log\left(\frac{1}{\epsilon}\right) \leq \frac{1}{3} \log(n),$$

at least for n large enough. Thus the exponential is bounded by $n^{1/3}$ (actually any exponent strictly smaller than $1/2$ would be enough). Moreover $\sqrt{\log(C_4/\epsilon)} \leq n^{\eta}$ for any $\eta > 0$ if n is large enough, and $\eta = 1/12$ will be sufficient for our purposes. So for n large enough

$$\frac{C_1}{\sqrt{n}} \exp\left(\frac{C_3}{\epsilon} \log\left(\frac{1}{\epsilon}\right)\right) \sqrt{\log\left(\frac{C_4}{\epsilon}\right)} \leq \frac{C_1}{\sqrt{n}} \cdot n^{1/3} \cdot n^{1/12} = C_1 n^{-1/12},$$

which is negligible compared to the first term which scales like $\log(\log(n))/\log(n)$. Thus

$$\mathcal{R}_n(\mathrm{Lip}_1^*(\mathcal{W})) \leq \frac{\mathrm{C}_5 \log(\log(\mathrm{n}))}{\log(\mathrm{n})}$$

for some constant C_5, at least if n is large enough. Plugging this into the first step yields the conclusion.

Proof of Theorem 4.1: the lower bound

Let P_0 a measure on $\mathbb{R}^d$ whose support has non-empty interior. Let $\mathbb{Q} = \mathrm{DP}(\alpha, P_0)$ be a Dirichlet process with parameters $\alpha > 0$ and P_0. For any exponent $\gamma > 0$ we define $D = \lceil 1/\gamma \rceil$.

The support of P_0 contains a ball in $\mathbb{R}^d$. In particular we can find D disjoint and closed sets $A_1, \ldots, A_D$ in the support of P_0 each of them with non-empty interior. Let $f_1, \ldots, f_D : \mathbb{X} \to \mathbb{R}$ be 1-Lipschitz functions with support in $A_1, \ldots, A_D$ respectively and which are not identically zero. E.g., we can set $f_i(\cdot) = d(\cdot, \mathbb{X} \setminus A_i)$ the distance to the complement of A_i. We consider the embedding ι of $\mathcal{P}(\mathbb{X})$ into $[0,1]^D$ given by

$$\iota(P) = (P(f_1), \ldots, P(f_D)).$$

Our conclusion will follow by combining the two following claims: (1) as ι is a Lipschitz map, the rate at which $\mathbb{E}(\mathcal{W}_{\mathcal{W}}(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q}))$ goes to zero is worse than the statistical rate of convergence for the embedded measures $\iota_{\#}\tilde{\mathbb{Q}}_{(n)}$, $\iota_{\#}\mathbb{Q}$; (2) the embedded measures live in a Euclidean space of dimension D, for which the Wasserstein distance suffers the curse of dimensionality, thus giving a lower bound on the rate.

Step 1: analysis of the embedding ι. The embedding $\iota : \mathcal{P}(\mathbb{X}) \to [0,1]^D$ is a Lipschitz map with respect to the Wasserstein-1 metric and the standard Euclidean distance: this is because the f_i are 1-Lipschitz so that

$$\|\iota(P_1) - \iota(P_2)\| = \sqrt{\sum_{i=1}^{D} |P_1(f_i) - P_2(f_i)|^2} \leq \sqrt{D} \sup_{f \in \mathrm{Lip}_1(\mathbb{X})} |P_1(f) - P_2(f)| = \sqrt{D}\mathcal{W}(P_1, P_2).$$

Integrating this inequality with $(\tilde{P}_1, \tilde{P}_2)$ any coupling between $(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q})$, and minimizing over $\mathbb{Q}$, we find

$$\mathcal{W}_{\mathcal{W}}(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q}) = \inf_{(\tilde{P}_n, \tilde{P}) \in \Gamma(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q})} \mathbb{E}(\mathcal{W}(\tilde{P}_n, \tilde{P})) \\ \geq D^{-1/2} \inf_{(\tilde{P}_n, \tilde{P}) \in \Gamma(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q})} \mathbb{E}(\|i(\tilde{P}_n) - i(\tilde{P})\|) \geq D^{-1/2} \mathcal{W}_{\mathrm{Euc}}(\iota_{\#}\tilde{\mathbb{Q}}_{(n)}, \iota_{\#}\mathbb{Q}),$$

where the $\mathcal{W}_{\mathrm{Euc}}$ denotes the standard 1-Wasserstein distance on $[0,1]^D$ with the Euclidean distance. Note that $\iota_{\#}\tilde{\mathbb{Q}}_{(n)} = (\iota_{\#}\tilde{\mathbb{Q}})_{(n)}$ is the empirical estimator of $\iota_{\#}\mathbb{Q}$.

Step 2: analysis of the law of $\iota(\tilde{P})$. Recall that $\tilde{P} \sim \mathbb{Q}$ follows a Dirichlet process of parameters α and P_0. The law of $\iota(\tilde{P}) = (\tilde{P}(f_1), \dots, \tilde{P}(f_D))$ as a distribution on $[0,1]^D$ is usually referred as the law of the *vector of random means*. We use Theorem 10 in Lijoi & Regazzini (2004) to prove that $\iota_{\#}\mathbb{Q}$, the distribution of $\iota(\tilde{P})$, is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^D$. This theorem requires $f = (f_1, \dots f_D)$ to not be affinely αP_0-degenerated, that is, for every $(v_1, \dots, v_D) \in \mathbb{R}^D \setminus \{0\}$, the function $x \mapsto \sum_i v_i f_i(x)$ is not P_0-a.s. a constant. This assumption is clearly satisfied as the functions $f_1, \dots, f_D$ are continuous, not identically zero, with disjoint support, each included in the support of P_0.

Step 3: conclusion with the curse of dimensionality of the Wasserstein distance. It is well understood that for an absolutely continuous measure with respect to the Lebesgue measure on $\mathbb{R}^D$ the rate of convergence cannot be better than $n^{-1/D}$ (see, e.g., Proposition 2.1 in Dudley (1969)). Thus, given Step 2, there exists a constant $c_D > 0$ such that

$$\mathbb{E}\left(\mathcal{W}_{\mathrm{Euc}}((\iota_{\#}\tilde{\mathbb{Q}})_{(n)}, \iota_{\#}\mathbb{Q})\right) \geq c_D n^{-1/D}.$$

Combining this with the result of Step 1,

$$\mathbb{E}\left(\mathcal{W}_{\mathcal{W}}(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q})\right) \geq c_D D^{-1/2}\, n^{-1/D}.$$

The conclusion follows as D was chosen with $1/D \leq \gamma$.

Proof of Lemma 4.2

The starting point follows closely the proof of the upper bound of Theorem 4.1, as we reduce to the estimation of a Rademacher complexity and then a covering number. By combining the definition of $d_{\mathcal{F}}$ with the dual formulation of the Wasserstein distance on $\mathbb{R}$ it holds

$$\mathbb{E}(d_{\mathcal{F}}(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q})) = \mathbb{E}\left(\sup_{h \in \mathfrak{F}} \left|\tilde{\mathbb{Q}}_{(n)}(h) - \mathbb{Q}(h)\right|\right),$$

where the class $\mathfrak{F}$ is define by

$$\mathfrak{F} = \{h : \mathcal{P}(\mathbb{X}) \to \mathbb{R} \text{ s.t. } \exists f \in \mathcal{F}, \exists g \in \mathrm{Lip}_1(\mathbb{R}, \mathbb{R}) \text{ s.t. } \mathrm{h(P) = g(P(f))}\}.$$

Step 1: getting back to a bounded class. To go back to a bounded class we shift the functions by a constant. We define

$$\mathfrak{F}^* = \{h : \mathcal{P}(\mathbb{X}) \to \mathbb{R} \text{ s.t. } \exists f \in \mathcal{F}^*, \exists g \in \mathrm{Lip}_1^*([-\mathrm{K}, \mathrm{K}], \mathbb{R}) \text{ s.t. } \mathrm{h(P) = g(P(f))}\},$$

where $\mathcal{F}^* = \{f^* = f - f(x_0) \text{ s.t. } f \in \mathcal{F}\}$, with x_0 a fixed point in $\mathbb{X}$, K is its uniform bound, i.e., $f(x) \leq K$ for every $f \in \mathcal{F}^*$, and $\mathrm{Lip}_1^*([-\mathrm{K}, \mathrm{K}], \mathbb{R}) \subseteq \mathrm{Lip}_1(\mathbb{R}, \mathbb{R})$ denotes the 1-Lipschitz functions $g : [-K, K] \to \mathbb{R}$ such that $g(0) = 0$. As probability distributions all have the same mass, it is easy to check that the quantity of interest can be expressed as

$$\mathbb{E}(d_{\mathcal{F}}(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q})) = \mathbb{E}\left(\sup_{h \in \mathfrak{F}^*} \left|\tilde{\mathbb{Q}}_{(n)}(h) - \mathbb{Q}(h)\right|\right).$$

Moreover, as the function $g \in \mathrm{Lip}_1^*([-\mathrm{K}, \mathrm{K}], \mathbb{R})$ are obviously bounded by K, we see that $\mathcal{F}^*$ uniformly bounded by K over $\mathbb{X}$ translates in $\mathfrak{F}^*$ uniformly bounded by K over $\mathcal{P}(\mathbb{X})$.

Step 2: reducing to an estimation of a Rademacher complexity. After these preliminary remarks we start similarly to the proof of Theorem 4.1: using the classical symmetrization argument (see e.g. Lemma 2.3.1 in (Vaart & Wellner, 1996) or Section 4.2 in (Wainwright, 2019)) we obtain

$$\mathbb{E}(d_{\mathcal{F}}(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q})) \leq 2\mathcal{R}_n(\mathfrak{F}^*),$$

where the Rademacher complexity $\mathcal{R}_n(\mathfrak{F}^*)$ is defined in (5).

Step 3: estimating the Rademacher complexity from above with a covering number By standard arguments using Dudley's entropy integral and the sub-Gaussianity of the Rademacher process (see, e.g., Theorem 5.22 and equation (5.48) in Wainwright (2019)), as the functions in $\mathfrak{F}^*$ are uniformly bounded by K then the Rademacher complexity may be bounded through an integral of the covering number of $\mathfrak{F}^*$ with respect to the empirical L_2 metric, which is itself controlled by the uniform norm L_∞. Let us define $N(\delta; \mathfrak{F}^*, \mathrm{L}_\infty)$ denote the δ-covering number of $\mathfrak{F}^*$ with respect to the uniform norm L_∞. Then the aforementioned results read

$$\mathcal{R}_n(\mathfrak{F}^*) \leq \inf_{\epsilon>0} \left\{ 2\epsilon + \frac{32}{\sqrt{n}} \int_{\epsilon/4}^{K} \sqrt{\log N(\delta; \mathfrak{F}^*, \mathrm{L}_\infty)}\, \mathrm{d}\delta \right\}.$$

We now bound $N(\delta; \mathfrak{F}^*, \mathrm{L}_\infty)$ through the covering number of $\mathrm{Lip}_1^*([-\mathrm{K}, \mathrm{K}], \mathbb{R})$ and $\mathcal{F}^*$ with respect to the supremum norm. Let $g_1, \ldots, g_A$ be a $\delta/2$-covering for $\mathrm{Lip}_1^*([-\mathrm{K}, \mathrm{K}], \mathbb{R})$ and let $f_1, \ldots, f_B$ be a $\delta/2$-covering for $\mathcal{F}^*$. In particular $A = N(\delta/2, \mathrm{Lip}_1^*([-\mathrm{K}, \mathrm{K}], \mathbb{R}), \mathrm{L}_\infty)$ and $B = N(\delta/2, \mathcal{F}^*, \mathrm{L}_\infty)$. We claim that $\{h_{i,j}(P) = g_i(P(f_j))\}$ is a δ-covering for $\mathfrak{F}^*$. Indeed, for any $h \in \mathfrak{F}^*$ such that $h(P) = g(P(f))$, let g_i and f_j such that respectively $\|g_i - g\|_{\mathrm{L}_\infty} \leq \delta/2$ and $\|f_j - f\|_{\mathrm{L}_\infty} \leq \delta/2$. Then

$$\begin{aligned}
\|h - h_{i,j}\|_{\mathrm{L}_\infty} &\leq \sup_P |h(P) - h_{i,j}(P)| \\
&\leq \sup_P |g(P(f)) - g(P(f_j))| + |g(P(f_j)) - g_i(P(f_j))| \\
&\leq \sup_P |P(f) - P(f_j)| + \sup_t |g(t) - g_i(t)| \\
&\leq \|f_j - f\|_{\mathrm{L}_\infty} + \|g_i - g\|_{\mathrm{L}_\infty} = \frac{\delta}{2} + \frac{\delta}{2} = \delta.
\end{aligned}$$

This shows that $N(\delta; \mathfrak{F}^*, \mathrm{L}_\infty) \leq \mathrm{AB}$. The quantity A, which is the $\delta/2$ covering number of $\mathrm{Lip}_1^*([-\mathrm{K}, \mathrm{K}], \mathbb{R})$, can be estimated as $\log(A) \leq \log(2)\lceil 4K/\delta \rceil$ (see e.g. p.93 in Kolmogorov & Tikhomirov (1961)). Thus

$$\log N(\delta; \mathfrak{F}^*, \mathrm{L}_\infty) \leq \log(2) \left\lceil \frac{4\mathrm{K}}{\delta} \right\rceil + \log \mathrm{N}\left(\frac{\delta}{2}; \mathcal{F}^*, \mathrm{L}_\infty\right).$$

Plugging this back we obtain

$$2\mathcal{R}_n(\mathfrak{F}^*) \leq \inf_{\epsilon>0} \left\{ 4\epsilon + \frac{64}{\sqrt{n}} \int_{\epsilon/4}^{K} \sqrt{\log(2) \left\lceil \frac{4K}{\delta} \right\rceil + \log N\left(\frac{\delta}{2}; \mathcal{F}^*, \mathrm{L}_\infty\right)}\, \mathrm{d}\delta \right\}.$$

By using the subadditivity of the square root, i.e. $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $a, b \geq 0$, the bounds of Step 2 and Step 3 yield

$$\mathbb{E}(d_{\mathcal{F}}(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q})) \leq \frac{64\sqrt{\log(2)}}{\sqrt{n}} \int_0^K \sqrt{\left\lceil \frac{4K}{\delta} \right\rceil}\, \mathrm{d}\delta + \inf_{\epsilon>0} \left\{ 4\epsilon + \frac{64}{\sqrt{n}} \int_{\epsilon/4}^{K} \sqrt{\log N\left(\frac{\delta}{2}; \mathcal{F}^*, \mathrm{L}_\infty\right)}\, \mathrm{d}\delta \right\}$$

To finish the proof we need only to estimate the prefactor of the parametric part, which we obtain by computing the first integral on the right hand side. As $\lceil 4K/\delta \rceil \leq 4K/\delta + 1$ and again by subadditivity of the square root, it can be estimated from above as

$$\int_0^K \sqrt{\left\lceil \frac{4K}{\delta} \right\rceil}\, \mathrm{d}\delta \leq \int_0^K \sqrt{\frac{4K}{\delta} + 1} \leq 2\sqrt{K} \int_0^K \delta^{-1/2}\, \mathrm{d}\delta + K = 5K.$$

The conclusion follows by chaining the estimates.

Proof of Theorem 4.3

Thanks to Lemma 4.2 the proof now amounts to an evaluation of the covering number of the class $\mathcal{F}^*$, which is standard when studying statistical rates for IPM (Sriperumbudur et al., 2010; 2012). We report the reasoning for completeness.

When $\mathbb{X}$ is a bounded domain of $\mathbb{R}^d$ and $\mathcal{F} = \mathrm{Lip}(\mathbb{X}, \mathbb{R})$, Theorem 2.7.1 in Vaart & Wellner (1996) yields the existence of a constant C_1 depending on $\mathrm{diam}(\mathbb{X})$ such that $\log N(\epsilon, \mathrm{Lip}_1(\mathbb{X}), \mathrm{L}_\infty) \leq \mathrm{C}_1 \epsilon^{-\mathrm{d}}$. Substituting in the bound of Lemma 4.2 we obtain that there exists a constant C_2, C_3 such that

$$\mathbb{E}(d_{\mathrm{Lip}}(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q})) \leq \frac{C_2}{\sqrt{n}} + \inf_{\epsilon > 0} \left\{ 4\epsilon + \frac{C_3}{\sqrt{n}} \int_{\epsilon/4}^{2K} \delta^{-d/2} \, \mathrm{d}\delta \right\}.$$

To get the announced rate it is enough to optimize the expression in ϵ. For $d = 1$ then $\epsilon = 0$ works as the function $f(\delta) = \delta^{-d/2}$ is integrable in zero; and for $d \geq 2$ we can take $\epsilon = n^{-1/d}$.

Proof of Theorem 4.4

By using the triangle inequality,

$$\mathbb{E}(d_{\mathrm{Lip}}(\tilde{\mathbb{Q}}_{(n,m))}, \mathbb{Q})) \leq \mathbb{E}(d_{\mathrm{Lip}}(\tilde{\mathbb{Q}}_{(n,m)}, \tilde{\mathbb{Q}}_{(n)})) + \mathbb{E}(d_{\mathrm{Lip}}(\tilde{\mathbb{Q}}_{(n)}, \mathbb{Q})).$$

Theorem 4.3 bounds the second term in the right hand side and so we only have to show that for every n,

$$\mathbb{E}(d_{\mathcal{F}}(\tilde{\mathbb{Q}}_{(n,m)}, \tilde{\mathbb{Q}}_{(n)})) \leq \begin{cases} C_1 m^{-1/2} & \text{if } d = 1, \\ C_2 m^{-1/2} \log(m) & \text{if } d = 2, \\ C_d m^{-1/d} & \text{if } d > 2, \end{cases}$$

where C_d does not depend on n and m. By definition of HIPM,

$$d_{\mathrm{Lip}}(\tilde{\mathbb{Q}}_{(n,m)}, \tilde{\mathbb{Q}}_{(n)}) = \sup_{f \in \mathcal{F}} \mathcal{W}\left(\frac{1}{n} \sum_{i=1}^{n} \delta_{\tilde{P}_{i,(m)}(f)}, \frac{1}{n} \sum_{i=1}^{n} \delta_{\tilde{P}_i(f)} \right).$$

By coupling $\tilde{P}_{i,(m)}(f)$ with $\tilde{P}_i(f)$ we obtain a natural upper bound

$$\mathbb{E}(d_{\mathrm{Lip}}(\tilde{\mathbb{Q}}_{(n,m)}, \tilde{\mathbb{Q}}_{(n)})) \leq \mathbb{E}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} |\tilde{P}_{i,(m)}(f) - \tilde{P}_i(f)| \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(\sup_{f \in \mathcal{F}} |\tilde{P}_{i,(m)}(f) - \tilde{P}_i(f)| \right).$$

By Corollary 8 in Sriperumbudur et al. (2010) on the estimation of the Wasserstein distance of order 1 through the empirical distribution,

$$\mathbb{E}\left(\sup_{f \in \mathcal{F}} |\tilde{P}_{i,(m)}(f) - \tilde{P}_i(f)| \middle| \tilde{P}_i \right) \leq \begin{cases} C_1 m^{-1/2} & \text{if } d = 1, \\ C_2 m^{-1/2} \log(m) & \text{if } d = 2, \\ C_d m^{-1/d} & \text{if } d > 2, \end{cases}$$

whenever $\mathbb{X} \subset \mathbb{R}^d$ is a bounded convex set with non-empty interior. The same inequality holds for any bounded set in $\mathbb{R}^d$ as it can be embedded in a bounded convex set with non-empty interior. Since the upper bound does not depend on $\tilde{P}_i$, by the towering property we obtain the desired upper bound.

Proof of Proposition 6.1

Upper bounds for the Dirichlet multinomial $\tilde{P}_1$. Note that we can write the law of $\tilde{P}_1$ as

$$\mathcal{L}(\tilde{P}_1) = \mathbb{E}_{X_{1:N}} \left(\mathrm{DP}\left(\alpha, \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \delta_{\mathrm{X_i}} \right) \right),$$

where we use $X_{1:N}$ as shortcut for $X_1, \ldots, X_N \overset{\text{iid}}{\sim} P_0$. By convexity of the Wasserstein distance (Theorem 4.8 in Villani (2008)),

$$\mathcal{W}_{\mathcal{W}}(\tilde{P}, \tilde{P}_1) \leq \mathbb{E}_{X_{1:N}} \left(\mathcal{W}_{\mathcal{W}} \left(\mathrm{DP}(\alpha, \mathrm{P}_0), \mathrm{DP}\left(\alpha, \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \delta_{\mathrm{X_i}} \right) \right) \right) = \mathbb{E}_{\mathrm{X}_{1:\mathrm{N}}} \left(\mathcal{W}\left(\mathrm{P}_0, \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \delta_{\mathrm{X_i}} \right) \right),$$

where the last equality holds by Corollary 3.7. The convergence rate of this quantity is well-studied in the literature when $\mathbb{X} = \mathbb{R}^d$ (see, e.g., Dudley (1969); Boissard & Gouic (2014); Fournier & Guillin (2015); Weed & Bach (2019); Bobkov & Ledoux (2019) and references therein). In particular, if $\mathbb{X} = \mathbb{R}$ Theorem 3.2 in Bobkov & Ledoux (2019) guarantees that

$$\mathbb{E}\left(\mathcal{W}\left(P_0, \frac{1}{N}\sum_{i=1}^{N}\delta_{X_i}\right)\right) \leq \frac{1}{\sqrt{N}}\int_{-\infty}^{+\infty}\sqrt{F_0(x)(1-F_0(x))}\mathrm{d}x.$$

Upper bounds for the truncated stick breaking $\tilde{P}_2$. Given a random sequence $(J_i)_{i\geq 1}$ of stick-breaking weights, and a sequence $(X_i)_{i\geq 1}$ of i.i.d. atoms, we can naturally build a coupling between $\mathcal{L}(\tilde{P})$ and $\mathcal{L}(\tilde{P}_2)$ via

$$\left(\sum_{i\geq 1} J_i\delta_{X_i}, \sum_{i=1}^{N-1} J_i\delta_{X_i} + \left(1 - \sum_{i=1}^{N-1} J_i\right)\delta_{X_N}\right).$$

This provides a natural upper bound for the Wasserstein over Wasserstein distance. Using $\sum_i J_i = 1$ a.s., we rewrite it as

$$\mathcal{W}_{\mathcal{W}}(\tilde{P}, \tilde{P}_2) \leq \mathbb{E}\left(\mathcal{W}\left(\sum_{i=1}^{N-1} J_i\delta_{X_i} + \sum_{i=N}^{+\infty} J_i\delta_{X_i}, \sum_{i=1}^{N-1} J_i\delta_{X_i} + \left(\sum_{i=N}^{+\infty} J_i\right)\delta_{X_N}\right)\right).$$

By using the representation of the Wasserstein distance as an IPM,

$$\mathcal{W}\left(\sum_{i=1}^{N-1} J_i\delta_{X_i} + \sum_{i=N}^{+\infty} J_i\delta_{X_i}, \sum_{i=1}^{N-1} J_i\delta_{X_i} + \left(\sum_{i=N}^{+\infty} J_i\right)\delta_{X_N}\right) = \sup_{f\in \mathrm{Lip}_1(\mathbb{X})}\left|\sum_{i=N+1}^{+\infty} J_i(f(X_i) - f(X_N))\right|.$$

We then use the natural bound $|f(X_i) - f(X_N)| \leq d_{\mathbb{X}}(X_i, X_N)$ independent of f. Note that (X_i, X_N) has the same law as (X_1, X_2) by the i.i.d. assumption. Thus when we take the expectation:

$$\mathcal{W}_{\mathcal{W}}(\tilde{P}, \tilde{P}_2) \leq \mathbb{E}\left(\sum_{i=N+1}^{+\infty} J_i d_{\mathbb{X}}(X_i, X_N)\right) = \mathbb{E}(d_{\mathbb{X}}(X_1, X_2))\mathbb{E}\left(\sum_{i=N+1}^{+\infty} J_i\right) = \mathbb{E}(d_{\mathbb{X}}(X_1, X_2))\left(\frac{\alpha}{\alpha+1}\right)^N,$$

where the last equality follows from Section 3.2 of Ishwaran & James (2001). Moreover, when $\mathbb{X} = \mathbb{R}$ the following identity holds (cfr., e.g., Bobkov & Ledoux (2019))

$$\mathbb{E}(|X_1 - X_2|) = 2\int_{-\infty}^{+\infty} F_0(x)(1-F_0(x))\mathrm{d}x.$$

Upper bounds for the hierarchical empirical measure $\tilde{P}_3$. The coupling that sends $\tilde{P}$ to the Dirichlet process that defines the law of the atoms of $\tilde{P}_3$ ensures that

$$\mathcal{W}_{\mathcal{W}}(\tilde{P}, \tilde{P}_3) \leq \mathbb{E}_{\tilde{P}}\left(\mathcal{W}\left(\tilde{P}, \frac{1}{N}\sum_{i=1}^{N}\delta_{X_i}\right)\right) = \mathbb{E}_{\tilde{P}}\left(\mathbb{E}\left(\mathcal{W}\left(\tilde{P}, \frac{1}{N}\sum_{i=1}^{N}\delta_{X_i}\right)\middle|\tilde{P}\right)\right),$$

thanks to the tower property. By Theorem 3.2 in Bobkov & Ledoux (2019) it follows that

$$\mathcal{W}_{\mathcal{W}}(\tilde{P}, \tilde{P}_3) \leq \mathbb{E}_{\tilde{P}}\left(\frac{1}{\sqrt{N}}\int_{-\infty}^{+\infty}\sqrt{\tilde{F}(x)(1-\tilde{F}(x))}\mathrm{d}x\right) \leq \frac{1}{\sqrt{N}}\int_{-\infty}^{+\infty}\sqrt{\mathbb{E}(\tilde{F}(x)) - \mathbb{E}(\tilde{F}(x)^2)}\mathrm{d}x,$$

where the last inequality holds by linearity of the expectation, Fubini's Theorem, and Jensen's inequality. Standard properties of the Dirichlet process ensure that $\mathbb{E}(\tilde{F}(x)) = F_0(x)$ and $\mathbb{E}(\tilde{F}(x)^2) = \mathrm{Var}(\tilde{F}(x)) + \mathbb{E}(F(x))^2 = (F_0(x)(1-F_0(x)))/(1+\alpha) + F_0(x)^2$. Thus the last term is equal to

$$\sqrt{\frac{\alpha}{N(\alpha+1)}}\int_{-\infty}^{+\infty}\sqrt{F_0(x)(1-F_0(x))}\mathrm{d}x.$$

Algorithm 1 Computation of $\mathcal{W}_{\mathcal{W}}$

Routine needed: OT(A,a,b) which outputs the value of the transport problem with cost matrix A of size $q \times q$ and weights a, b for the marginals.
Input: sizes n, m, data $X^1_{i,j}$, $X^2_{i,j}$, $\omega^1_{i,j}$, $\omega^2_{i,j}$
Initialize $C = 0$ of size $n \times n$ // Compute the cost matrix for the $n \times n$ transport problem
for $i_1, i_2 = 1$ **to** n **do**
  $B \leftarrow$ PairwiseDistance($X^1_{i_1,\cdot}$, $X^2_{i_2,\cdot}$)
  $C_{i_1,i_2} \leftarrow$ OT(B, $\omega^1_{i_1,\cdot}$, $\omega^2_{i_2,\cdot}$)
end for
Return: OT(C, $1/n\,\mathbf{1}$, $1/n\,\mathbf{1}$) // Solve the $n \times n$ transport problem

B. Additional information on the algorithms

B.1. Wasserstein over Wasserstein

In Algorithm 1 we present the pseudocode for computing the Wasserstein over Wasserstein distance as described in Section 5.

B.2. Lipschitz HIPM in dimension one

We expand on the gradient ascent algorithm for our new distance d_{Lip}. Recall that we are solving the optimization problem:

$$\begin{aligned} &\sup_{\mathbf{f}\in\mathbb{R}^M} \mathcal{G}(\mathbf{f}) \\ &\text{such that } |\mathbf{f}_{q+1} - \mathbf{f}_q| \le \Delta x \ \forall q \in \{1,\dots,M-1\}, \\ &\text{where } \mathcal{G}(\mathbf{f}) := \inf_{\sigma\in\mathcal{S}(n)} \frac{1}{n}\sum_{i=1}^{n} \left| \sum_{q=1}^{M} (\omega^1_{i,q} - \omega^2_{\sigma(i),q})\mathbf{f}_q \right|. \end{aligned} \tag{9}$$

The function $\mathcal{G}$ is piece-wise linear. On each "facet", that is, when there exists a unique permutation σ^* which realizes the infimum, and when each term $(\omega^1_{i,q} - \omega^2_{\sigma(i),q})\mathbf{f}_q$ is either strictly positive or strictly negative, then we can easily compute the gradient. It reads: for any coordinate q,

$$\nabla_q \mathcal{G}(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n} (\omega^1_{i,q} - \omega^2_{\sigma^*(i),q}) \operatorname{sign}\left(\sum_{q'=1}^{N} (\omega^1_{i,q'} - \omega^2_{\sigma^*(i),q'})\mathbf{f}_q \right).$$

If σ^* is not unique or $(\omega^1_{i,q} - \omega^2_{\sigma(i),q})$ vanishes the gradient may not exist, but in practice we ignore these degenerate cases. We further rewrite the problem by parametrizing the function $\mathbf{f}$ rather by its derivative to simplify the Lipschitz constraint. Let $\mathbf{g} \in \mathbb{R}^{M-1}$ and consider the $M \times (M-1)$ matrix

$$A = \Delta x \begin{pmatrix} 0 & 0 & \dots & \dots & 0 \\ 1 & 0 & \dots & \dots & 0 \\ 1 & 1 & 0 & \ddots & \vdots \\ \vdots & & \ddots & & 0 \\ 1 & 1 & \dots & 1 & 1 \end{pmatrix}.$$

We consider $\mathbf{f} = A\mathbf{g}$, that is, $\mathbf{f}_q = \sum_{q'<q} \Delta x\, \mathbf{g}_{q'}$ for any q. Specifically we define $\hat{\mathcal{G}}(\mathbf{g}) = \mathcal{G}(A\mathbf{f})$, so that problem (9) can be rewritten as $\sup \hat{\mathcal{G}}(\mathbf{g})$ for $\mathbf{g} \in [-1,1]^{M-1}$. The gain is that the constraint on $\mathbf{g}$ is a very simple box constraint. Moreover the gradient can be easily computed via $\nabla\hat{\mathcal{G}}(\mathbf{g}) = A^\top \nabla\mathcal{G}(A\mathbf{f})$.

We implemented the following gradient ascent. Given $\mathbf{g}$ admissible we find an ascent direction $\mathbf{a}$ by orthogonally projecting $\nabla\hat{\mathcal{G}}(\mathbf{g})$ on the set of vectors that preserves the box constraint. This can be easily done: starting with $\mathbf{a} = \nabla\hat{\mathcal{G}}(\mathbf{g})$, we set $\mathbf{a}_q = 0$ if $\mathbf{g}_q = 1$ and $\nabla_q\hat{\mathcal{G}}(\mathbf{g}) > 0$ (respectively if $\mathbf{g}_q = -1$ and $\nabla_q\hat{\mathcal{G}}(\mathbf{g}) < 0$). We then find $t_{\max}$ the largest $t \ge 0$ such that $\mathbf{g} + t\mathbf{a} \in [-1,1]^M$, and we use a backtracking linesearch to find $t \in [0, t_{\max}]$ such that $\mathcal{G}(\mathbf{g} + t\mathbf{a})$ increases enough (Armijo, 1966). We stop the loop if the expected increase, that is, $\mathbf{a}^\top\nabla\hat{\mathcal{G}}(\mathbf{g})$, is too small.

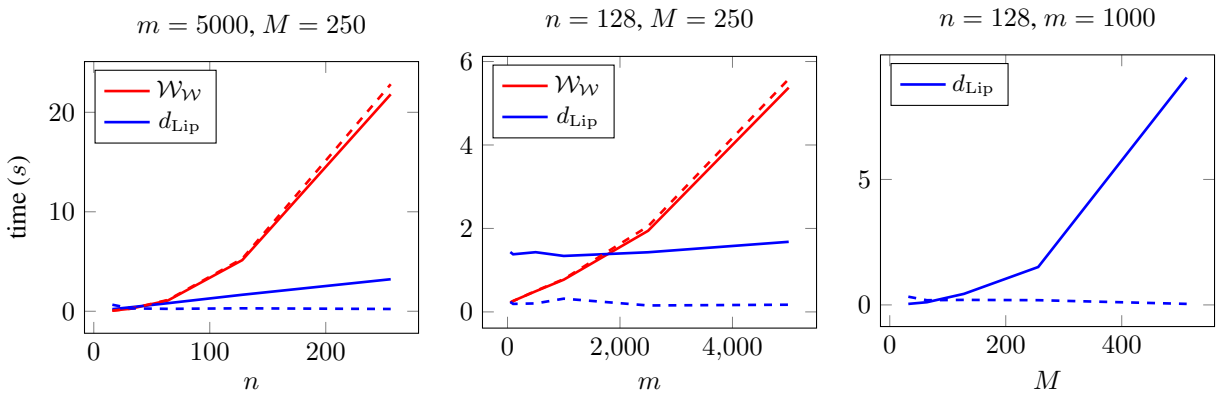

Figure 3. Execution time (in seconds) for several configurations of the parameters n, m and M. The setting corresponds to the one of Figure 1 top (for the solid lines) and Figure 1 bottom (for the dashed lines). Computations are repeated 12 times and the averaged execution time is reported.

As we have no guarantee of finding a global maximizer we run the gradient ascents with several different initializations. We usually include the initialization $f = \mathrm{Id}$, that is, $\mathbf{g} = (1, 1, \ldots, 1)$ in at least one of the runs, which is related to the bound (1). This is summarized in the Algorithm 2.

Algorithm 2 Computation of an approximation of d_{Lip}

Input: sizes n, M, data $\omega^1_{i,q}$, $\omega^2_{i,q}$, stepsize Δx,
Parameters: number of initializations n_{init}, number of steps n_{step}, tolerance ε
for $s = 1$ **to** n_{init} **do**
  $\mathbf{g} \leftarrow$ `Random Initialization` or $\mathbf{g} \leftarrow (1, 1, \ldots, 1)$
  repeat
    $\mathbf{a} \leftarrow \nabla\hat{\mathcal{G}}(\mathbf{g})$ // Ascent direction
    for $q = 1$ **to** $M - 1$ **do**
      if $\mathbf{g}_q = \pm 1$ **then** $\pm\mathbf{a}_q = \min(0, \pm\mathbf{a}_q)$ **end if** // Projection of the ascent direction
    end for
    $t_{\max} \leftarrow \sup\{t \geq 0 \,:\, \mathbf{g} + t\mathbf{a}\} \in [-1, 1]^{M-1}$
    $t \leftarrow t_{\max}$ // Backtracking line search
    while $\mathcal{G}(\mathbf{g} + t\mathbf{a}) < \mathcal{G}(\mathbf{g}) + t\mathbf{a}^\top\nabla\hat{\mathcal{G}}(\mathbf{g})/2$ **do**
      $t \leftarrow t/2$
    end while
    $\mathbf{g} \leftarrow \mathbf{g} + t\mathbf{a}$ // Update of the gradient ascent
  until $\mathbf{a}^\top\nabla\hat{\mathcal{G}}(\mathbf{g}) \leq \varepsilon$ **or** loop done n_{step} times
  Store $\hat{\mathcal{G}}(\mathbf{g})$
end for
Return: Maximum $\hat{\mathcal{G}}(\mathbf{g})$ among the n_{init} runs

B.3. Execution time

Computations were performed on the CPU of a standard laptop with a 6-core 2.10GHz AMD Ryzen 5 5500U processor with 8Go of RAM. We report in Figure 3 the execution time of the computation of $\mathcal{W}_{\mathcal{W}}$ and (the approximation of) d_{Lip}. We do so for different values of n, m and also M the number of grid points in the setting of Figure 1. Whereas the time for computing $\mathcal{W}_{\mathcal{W}}$ grows quickly with n and is quite insensible to the input measures, we see on the other hand that M and the input measures are the sensible parameters for d_{Lip}. The input measures determine the optimal $\mathbf{f}$, and this is likely to affect the convergence of the gradient ascent. Though d_{Lip} is unsurprisingly slow to compute when M is large, recall from that from our error analysis it is natural to take $M \asymp m^{1/2}$, thus moderately large.