Optimal Estimation of Wasserstein Distance on A Tree with An Application to Microbiome Studies

Shulei Wang, T. Tony Cai and Hongzhe Li

Abstract

The weighted UniFrac distance, a plug-in estimator of the Wasserstein distance of read counts on a tree, has been widely used to measure the microbial community difference in microbiome studies. Our investigation however shows that such a plug-in estimator, although intuitive and commonly used in practice, suffers from potential bias. Motivated by this finding, we study the problem of optimal estimation of the Wasserstein distance between two distributions on a tree from the sampled data in the high-dimensional setting. The minimax rate of convergence is established. To overcome the bias problem, we introduce a new estimator, referred to as the moment-screening estimator on a tree (MET), by using implicit best polynomial approximation that incorporates the tree structure. The new estimator is computationally efficient and is shown to be minimax rate-optimal. Numerical studies using both simulated and real biological datasets demonstrate the practical merits of MET, including reduced biases and statistically more significant differences in microbiome between the inactive Crohn's disease patients and the normal controls.

Keywords: Estimation of non-smooth functional; Pylogenetic tree; Polynomial approximation

Shulei Wang is a Postdoctoral Fellow, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *Shulei.Wang@pennmedicine.upenn.edu*). T. Tony Cai is Daniel H. Silberberg Professor of Statistics, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail:*tcai@wharton.upenn.edu*). Hongzhe Li is Professor of Biostatistics and Statistics, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *hongzhe@upenn.edu*).

1 Introduction

High throughput sequencing technologies allow a high resolution characterization of the collection of all microbes in a sample, leading to a comprehensive understanding of microbial communities. The composition of microbes in a given microbial community can be represented by discrete distributions $P = \{p_v\}_{v \in V}$, where V is a finite set of microbe taxa (or operational taxonomic units (OTU) in some applications), and p_v is the relative abundance of the vth bacterial taxon. Phylogenetic tree provides an effective way of summarizing how bacterial species or OTUs are related through evolution based on the sequences of certain marker genes such as 16s rRNA gene. As an example, Figure 1 shows the phylogenetic tree of the 3991 bacterial OTUs identified in a Crohn's disease study detailed in Section 7, where the leaf notes represent the OTUs, branch lengths reflect the evolutionary distances and the internal notes represent the common ancestry of the nodes below.

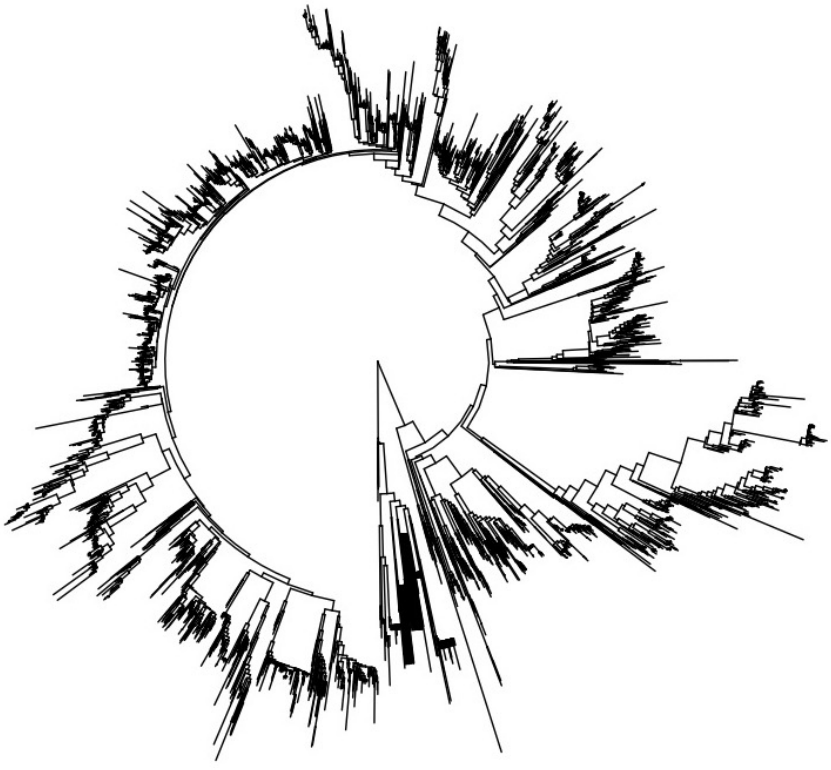


Figure 1: Phylogenetic tree used in the 16S rRNA sequecning data of the Crohn's disease study. There are a total 3991 leaves (tips) and 3990 internal nodes.

Measuring the distance between two communities is the first step towards understanding the microbial similarity and differences across samples. Various symmetrical distances between two distributions are used to quantify difference between the two communities, including the total variation distance, Kullback-Leibler divergence and Hellinger distance. Here, a symmetrical distance means that it is invariant with respect to the permutation of the microbe species. However, such symmetrical distances ignore the similarity among different microbe species as given by their phylogenetic relationships. In order to account for such a similarity between the microbe species, phylogenetic distances based on the empirical distributions of the read counts and the underlying phylogentic tree have been proposed as a more powerful and precise way to quantify difference between two microbial communities. In particular, as one of the most popular phylogenetic distances, the unweighted and weighted UniFrac distances are introduced by Lozupone and Knight (2005) and Lozupone et al. (2007) and have been used in a wide range of microbiome studies (see, e.g. Lozupone et al., 2007; Fierer et al., 2008; Charlson et al., 2010; Chang, Luan, and Sun, 2011; Wong, Wu, and Gloor, 2016). As shown in Evans and Matsen (2012), the weighted UniFrac distance can also be viewed as the plug-in estimator of the Wasserstein distance, also known as Kantorovich-Rubinstein distance or earth mover's distance (see, e.g. Monge, 1781; Kantorovitch, 1958; Villani, 2008), on a phylogenetic tree. Such a distance can be generalized to the L^α Zolotarev-type distance.

To be more specific, we consider two microbiome communities represented by discrete distributions $P = \{p_v\}_{v\in V}$ and $Q = \{q_v\}_{v\in V}$, where V is a finite microbe species (or operational taxonomic units (OTU) in some applications) set. Let $D(P, Q)$ denote the Wasserstein distance between P and Q. In practice, P and Q are unknown and one only has access to the empirical sequencing read distributions $\hat{P} = \{\hat{p}_v\}_{v\in V}$ and $\hat{Q} = \{\hat{q}_v\}_{v\in V}$. In the present paper, we consider optimal estimation of the distance between distributions $D(P, Q)$ when only the empirical distributions $\hat{P}$ and $\hat{Q}$ are available. Since the introduction of the UniFrac distance, the plug-in estimator $D(\hat{P}, \hat{Q})$ has been virtually the only estimator for $D(P, Q)$ because of its simplicity and ease of computation (see, e.g. Lozupone and Knight, 2005; Evans and Matsen, 2012). Despite the popu-

larity of this classical plug-in estimator in various microbiome studies, it is still largely unknown if there are more efficient estimators and to what extent such a distance $D(P, Q)$ can be estimated consistently. The main goal of this paper is to address these issues by answering the following two questions: 1) what sample size of the data can guarantee reliable estimation of $D(P, Q)$? 2) what is an optimal estimator for $D(P, Q)$?

To answer these questions, we first investigate the classical plug-in estimator. As the maximum likelihood estimator (MLE), the empirical distribution $\hat{P}$ has been shown to achieve the minimax optimality for estimating the distribution P itself under various loss functions (see, e.g. Trybula, 1958; Olkin and Sobel, 1979; Daskalakis, Diakonikolas, and Servedio, 2012; Kamath et al., 2015). Moreover, when the sample size goes to infinity and the number of microbiome species is fixed, the asymptotic theory for the MLE guarantees that the classical plug-in estimator performs optimally for estimating smooth functionals (see, e.g. Le Cam, 1986). However, our investigation shows that for estimating the Wasserstein distance the plug-in estimator $D(\hat{P}, \hat{Q})$ is sub-optimal due to its large bias resulting from the non-smooth nature of the functional $D(P, Q)$ and the high dimensionality of the parameter space. Similar phenomena have been observed in recent papers (see, e.g., Lepski, Nemirovski, and Spokoiny, 1999; Cai and Low, 2011; Valiant and Valiant, 2011, 2013; Jiao et al., 2015; Wu and Yang, 2016; Jiao, Han, and Weissman, 2018, among many others) in other settings.

Cai and Low (2011) introduced the best polynomial approximation approach to estimation of non-smooth functionals to reduce the bias. This approach proceeds by first constructing the best polynomial approximation to the target functional and the unbiased estimator for the best polynomial is then constructed as the final estimator of the original functional. This idea has since been widely used to estimate symmetrical non-smooth functionals of distributions, such as the Shannon entropy, Rényi entropy, support size, L_1 distance, χ^2 divergence, Kullback-Leibler divergence and Hellinger divergence (See Acharya et al. (2014), Jiao et al. (2015), Wu and Yang (2016), Han, Jiao, and Weissman (2016), Jiao, Han, and Weissman (2018), Bu et al. (2018) and references therein). One main difficulty of the approximation method is that it requires the construction of approxi-

mation specifically for each individual functional. To address this issue, two adaptive approaches have been proposed: local moment matching (see, e.g. Han, Jiao, and Weissman, 2018) and profile maximum likelihood (see, e.g. Acharya et al., 2017; Pavlichin, Jiao, and Weissman, 2017; Acharya, 2018). Both methods are designed to first estimate the sorted version of the distribution and then plugin the sorted distribution into arbitrary symmetrical functionals. Unfortunately, these methods and analyses are not directly applicable to estimating the Wasserstein distance due to the fact that $D(P, Q)$ is an asymmetrical distance.

Motivated by the best polynomial approximation method and moment matching method, we introduce a new Moment-screening Estimator on a Tree, called MET hereafter, to estimate the Wasserstein distance $D(P, Q)$. MET first conducts moment matching by taking advantage of the unique structure of the phylogenetic tree and then estimates the Wasserstein distance by an implicit approximation method. In doing so, MET requires no specific construction for the best polynomial approximation, but achieves the same bias reduction effect as the approximation method. We establish the minimax rate of convergence for estimating $D(P, Q)$ under the mean squared error as

$$\frac{s \log(2^{d+2}/s)}{n \log n}, \tag{1}$$

where n is the sample size, s is the number of nodes of the phylogenetic tree and d is the height of the phylogenetic tree. In this minimax rate (1), the term $\log(2^{d+2}/s)$ is mainly determined by the shape of underlying phylogenetic tree. When the tree is short enough ($2^d \asymp s$), the difficulty of estimating Wasserstein distance does not rely on the height of phylogenetic tree d any more. The minimax rate (1) increases along with d linearly if the tree is tall ($2^d \gg s$). If we compare (1) with results in Jiao, Han, and Weissman (2018), $s \log(2^{d+2}/s)$ can be seen as effective alphabet size after incorporating the phylogenetic tree structure. Moreover, $D(P, Q)$ can be estimated consistently if and only if

$$n \gtrsim \frac{s \log(2^{d+2}/s)}{\log s}.$$

We also show that MET is rate optimal, while the classical plug-in estimator such as the UniFrac distance is sub-optimal.

Furthermore, we consider estimation of the L^α Zolotarev-type generalization of the Wasserstein distance, denoted by $D_\alpha(P, Q)$. The Wasserstein distance is a special case with $\alpha = 1$. Our analysis shows that MET is still minimax optimal estimator for $D_\alpha(P, Q)$ when $0 < \alpha < 2$, while the simple plug-in estimator can achieve the minimax optimal rate when $\alpha \geq 2$. We also show that estimation of $D_\alpha(P, Q)$ becomes more difficult as α gets smaller.

The rest of the paper is organized as follows. We first introduce the setting and the Wasserstein distance as well as its plug-in UniFrac distance estimator in Section 2. Section 3 presents MET for estimating the Wasserstein distance $D(P, Q)$. In Section 4, we investigate the theoretical properties of MET and compare it with the classical plug-in estimator theoretically. Section 5 studies estimation of the L^α Zolotarev-type generalization of the Wasserstein distance and provides the corresponding theoretical analysis. Section 6 discusses the algorithmic details of implementing MET. We analyze both the simulated and real data sets in Section 7 to demonstrate the numerical performance of MET. Proofs and auxiliary results are relegated to the online Supplemental Materials.

2 Wasserstein Distance and the Plug-in UniFrac Distance

2.1 Wasserstein Distance and Poisson-Multinomial Model

Let $T = (V, E)$ be the phylogenetic tree of microbe species, where V is the collection of microbe species and their ancestors and E is the collection of edges/branches of the tree T. In particular, we always assume that the tree T is rooted at ρ. Denote by L_e the length of the branch $e \in E$. For any pairs of nodes $v_1, v_2 \in V$, the unique path between them is denoted by $[v_1, v_2]$ and the corresponding distance between them is defined as

$$d(v_1, v_2) := \sum_{e \in [v_1, v_2]} L_e.$$

The height/depth of the tree is thus defined as the maximum of distance between the root ρ and the other nodes of tree

$$d(T) = \max_{v \in V} d(\rho, v).$$

We write the Wasserstein distance between the distributions P and Q on the tree T as

$$D(P,Q) = \inf_{\{r_{v_1,v_2}\}_{v_1,v_2\in V}\in\Sigma(P,Q)} \sum_{v_1,v_2\in V} d(v_1,v_2) r_{v_1,v_2}.$$

where $\{r_{v_1,v_2}\}_{v_1,v_2\in V}$ is the joint probability distribution on $V \times V$ and $\Sigma(P,Q)$ is the collection of the joint probability distributions of which the marginal distributions are P and Q, respectively. If we define the descendants of a given branch $e \in E$ as

$$\tau(e) = \{v \in V : e \in [\rho, v]\},$$

then the above Wasserstein distance can be rewritten (see, e.g. Evans and Matsen, 2012) as

$$D(P,Q) = \sum_{e\in E} L_e \left| P_e - Q_e \right|, \tag{2}$$

where P_e and Q_e are the total proportion of subtree below edge e

$$P_e = \sum_{v\in\tau(e)} p_v \qquad \text{and} \qquad Q_e = \sum_{v\in\tau(e)} q_v.$$

Note that (2) is also the original form of the weighted UniFrac distance (see Lozupone and Knight, 2005; Lozupone et al., 2007).

In microbiome studies, the sequencing read data can be modeled by a Poisson-multinomial model. More concretely, denote by $X_1, ..., X_{n_X}$ and $Y_1, ..., Y_{n_Y}$ the reads from the two samples. We assume the total numbers of reads n_X and n_Y are independent random variables drawn from a Poisson distribution, i.e. $n_X, n_Y \overset{i.i.d.}{\sim} \text{Pois}(n)$. Conditioning on n_X and n_Y, the reads are modeled as a multinomial distribution: $X_1, ..., X_{n_X}|n_X \overset{i.i.d.}{\sim} \text{Multi}(1; \{p_v\}_{v\in V})$ and $Y_1, ..., Y_{n_Y}|n_Y \overset{i.i.d.}{\sim} \text{Multi}(1; \{q_v\}_{v\in V})$. The empirical distribution thus can be written as

$$\hat{p}_v = \frac{\sum_{i=1}^{n_X} \mathbf{I}(X_i = v)}{n_X} \qquad \text{and} \qquad \hat{q}_v = \frac{\sum_{i=1}^{n_Y} \mathbf{I}(Y_i = v)}{n_Y}.$$

Clearly, the Poisson-multinomial model suggests $\hat{p}_v$s and $\hat{q}_v$s are independent from each other and

$$n\hat{p}_v \sim \text{Pois}(np_v) \qquad \text{and} \qquad n\hat{q}_v \sim \text{Pois}(nq_v).$$

We also use the following notation in this paper

$$\hat{P}_e = \sum_{v \in \tau(e)} \hat{p}_v \qquad \text{and} \qquad \hat{Q}_e = \sum_{v \in \tau(e)} \hat{q}_v.$$

2.2 The Classical Plug-in Estimator and the UniFrac Distance

The most natural estimator for $D(P, Q)$ is perhaps the plug-in estimator:

$$D(\hat{P}, \hat{Q}) := \sum_{e \in E} L_e \left| \hat{P}_e - \hat{Q}_e \right|.$$

This plug-in estimator, known as the UniFrac distance, has been widely used in many applications, including community comparison (see, e.g. Lozupone and Knight, 2005; Lozupone et al., 2007; Chang, Luan, and Sun, 2011; Evans and Matsen, 2012), clustering based on pairwise distance (see, e.g. Lozupone et al., 2007) and two sample testing (see, e.g. Charlson et al., 2010; Wong, Wu, and Gloor, 2016). Despite its popularity, the performance of the plug-in estimator is still unclear.

We now examine the theoretical performance of $D(\hat{P}, \hat{Q})$. We use the mean squared error to evaluate the accuracy of an estimator $\hat{D}$,

$$\mathbb{E}\left(\hat{D} - D(P, Q)\right)^2.$$

The following proposition characterizes the performance of the plug-in estimator.

Proposition 1. *Suppose T is a tree of height d. There exists some constant C such that*

$$\mathbb{E}\left(D(\hat{P}, \hat{Q}) - D(P, Q)\right)^2 \le C R_{\text{plug-in}}(T; P, Q)$$

where

$$R_{\text{plug-in}}(T, P, Q) = \left(\sum_{e \in E} P_e \wedge \sqrt{\frac{P_e}{n}}\right)^2 + \left(\sum_{e \in E} Q_e \wedge \sqrt{\frac{Q_e}{n}}\right)^2 + \frac{d^2}{n}.$$

Furthermore, there exist two pairs of distributions (T, P_1, Q_1) and (T, P_2, Q_2) such that

$$\inf_{\hat{D}} \sup_{(T,P_1,Q_1),(T,P_2,Q_2)} \mathbb{E}\left(\hat{D} - D(P, Q)\right)^2 \ge c\frac{d^2}{n}.$$

where $c > 0$ is some constant and the infimum takes over all possible estimators.

In $R_{\text{plug}-\text{in}}(T, P, Q)$, the first two terms corresponding to the bias of the plug-in estimator and the last term is the variance of the plug-in estimator. The lower bound suggests that the last term in the upper bound cannot be improved in the minimax sense. This naturally brings about the question of whether it is possible to reduce the bias in order to construct a more efficient estimator for $D(P, Q)$.

3 Moment-screening Estimator on a Tree

3.1 Behavior of the Bias Term in the Plug-in Estimator

We first investigate behavior of the bias of the plug-in estimator. The conditional expectation of the plug-in estimator given n_X and n_Y can be written explicitly as

$$\mathbb{E}\left(D(\hat{P}, \hat{Q})\Big| n_X, n_Y\right) = \sum_{e\in E} L_e \left(\sum_{k_1,k_2=0}^{n_X,n_Y} f(k_1, k_2) P_e^{k_1}(1-P_e)^{n_X-k_1} Q_e^{k_2}(1-Q_e)^{n_Y-k_2}\right), \quad (3)$$

where

$$f(k_1, k_2) = \binom{n_X}{k_1}\binom{n_Y}{k_2}\frac{|k_1 - k_2|}{n}.$$

Equation (3) suggests that the expectation of the plug-in estimator is essentially a polynomial of $\{P_e\}_{e\in E}$ and $\{Q_e\}_{e\in E}$ and the bias of the plug-in estimator mainly results from the polynomial approximation error for absolute value function $|x - y|$ near the diagonal line $x = y$. Actually, the expectation of any estimator based on $\hat{P}$ and $\hat{Q}$ can always be expressed as a polynomial. Similar phenomena are observed in functional estimation of single distribution (see, e.g. Paninski, 2003; Jiao et al., 2015; Wu and Yang, 2016). It is clear from the above discussion that we can reduce the bias by redesigning the coefficient of polynomial $f(k_1, k_2)$ to better approximating the absolute value function near diagonal line.

A prerequisite step for the bias reduction is to identify the pairs (P_e, Q_e) that are near diagonal line. We consider the following uncertain set covering the diagonal line

$$\mathcal{P} = \left\{(p, q) : |p - q| \le \min\left(\sqrt{\frac{1.1c_1(p+q)\log n}{n}}, |p+q|\right)\right\}$$

for some constant c_1 that will be specified later. To identify if (P_e, Q_e) belongs to $\mathcal{P}$, we adopt sample splitting techniques on the Poisson distribution. To be specific, we draw an independent uniform Bernoulli variable $\mathbb{P}(B = 0) = \mathbb{P}(B = 1) = 0.5$ for each X_i and each Y_i, and then split the samples according to the value of B. The split empirical distributions thus can be written

$$\hat{p}_{v,j} = \frac{\sum_{i=1}^{n_X} \mathbf{I}(X_i = v)\mathbf{I}(B_i^X = j)}{n_X} \qquad \text{and} \qquad \hat{q}_{v,j} = \frac{\sum_{i=1}^{n_Y} \mathbf{I}(Y_i = v)\mathbf{I}(B_i^Y = j)}{n_Y}$$

for $j = 0, 1$. Here B_i^Xs and B_i^Ys are independent uniform Bernoulli random variables. Similarly, we write $\hat{P}_{e,j} = \sum_{v\in\tau(e)} \hat{p}_{v,j}$ and $\hat{Q}_{e,j} = \sum_{v\in\tau(e)} \hat{q}_{v,j}$ for $j = 0, 1$. The construction suggests that $\hat{p}_{v,0}$ and $\hat{p}_{v,1}$ are independent Poisson random variables with mean $np_v/2$, and $\hat{q}_{v,0}$ and $\hat{q}_{v,1}$ are independent Poisson random variables with mean $nq_v/2$. Hereafter, we redefine $n/2$ as n. This sample splitting strategy allows us to use $(\hat{P}_{e,0}, \hat{Q}_{e,0})$ to localize whether (P_e, Q_e) belong to $\mathcal{P}$ and estimate the functional by $(\hat{P}_{e,1}, \hat{Q}_{e,1})$. When $(\hat{P}_{e,0}, \hat{Q}_{e,0}) \notin \mathcal{P}$, it holds with high probability that (P_e, Q_e) satisfies either $Q_e < P_e$ or $P_e < Q_e$ only. This implies that $|P_e - Q_e|$ can be estimated by $|\hat{P}_{e,1} - \hat{Q}_{e,1}|$ in an unbiased way and we could simply adopt the classical plug-in estimator. On the other hand, if $(\hat{P}_{e,0}, \hat{Q}_{e,0}) \in \mathcal{P}$, it is necessary to design an estimator to carefully reduce the approximation bias. For brevity, we write hereafter

$$E_r = \left\{e \in E : (\hat{P}_{e,0}, \hat{Q}_{e,0}) \in \mathcal{P}\right\} \quad \text{and} \quad E_c = \left\{e \in E : (\hat{P}_{e,0}, \hat{Q}_{e,0}) \notin \mathcal{P}\right\}.$$

It is worth noting that the sample splitting technique is mainly used to simplify the analysis. It is not necessary to split the samples in practice and we do not split the samples in the numerical experiments in Section 7.

3.2 A Bias Reduction Strategy

A natural bias reduction strategy inspired by (3) is to construct an unbiased estimator of the best polynomial approximation for the target function $|x-y|$ or $e \in E_r$. As mentioned earlier, the use of the best polynomial approximation method was pioneered in Cai and Low (2011) for bias reduction in estimation of non-smooth functionals. Its popularity is justified as it leads to the construction of the rate-optimal estimators in different problems (see e.g., Cai and Low (2011), Jiao et al. (2015),

Wu and Yang (2016) and Jiao, Han, and Weissman (2018) and references therein). To be more specific, let $F_e^K(P_e, Q_e)$ be some polynomial of degree at most K designed for (P_e, Q_e) at edge $e \in E_r$

$$F_e^K(P_e, Q_e) = \sum_{k_1,k_2=0}^{K} f_e(k_1, k_2) P_e^{k_1} Q_e^{k_2}.$$

The choice of $F_e^K(P_e, Q_e)$ is usually determined by $(\hat{P}_{e,0}, \hat{Q}_{e,0})$ to approximate target functional locally (see, e.g. Jiao, Han, and Weissman, 2018). The corresponding unbiased estimator of $F_e^K(P_e, Q_e)$ can be written as

$$\hat{F}_e^K = \sum_{k_1,k_2=0}^{K} f_e(k_1, k_2) H_{k_1}(\hat{P}_{e,1}) H_{k_2}(\hat{Q}_{e,1}),$$

where $H_k(\hat{P}_{e,1})$ is an unbiased estimator for P_e^k, i.e. $H_k(x) = \prod_{m=0}^{k-1} \left(x - \frac{m}{n}\right)$ if $k \geq 1$ and $H_k(x) = 1$ when $k = 0$. Thus, the bias of estimator $\sum_{e \in E_r} L_e \hat{F}_e^K$ is mainly the approximation error of the carefully chosen polynomials

$$\left| \mathbb{E} \sum_{e \in E_r} L_e \left(\hat{F}_e^K - |P_e - Q_e| \right) \right| \leq \sum_{e \in E_r} L_e \sup_{x,y} \left| F_e^K(x, y) - |x - y| \right|.$$

The squared error of estimator $\sum_{e \in E_r} L_e \hat{F}_e^K$ can thus be decomposed into bias (Bias A) and variance (Variance A) as illustrated in left half of Figure 2.

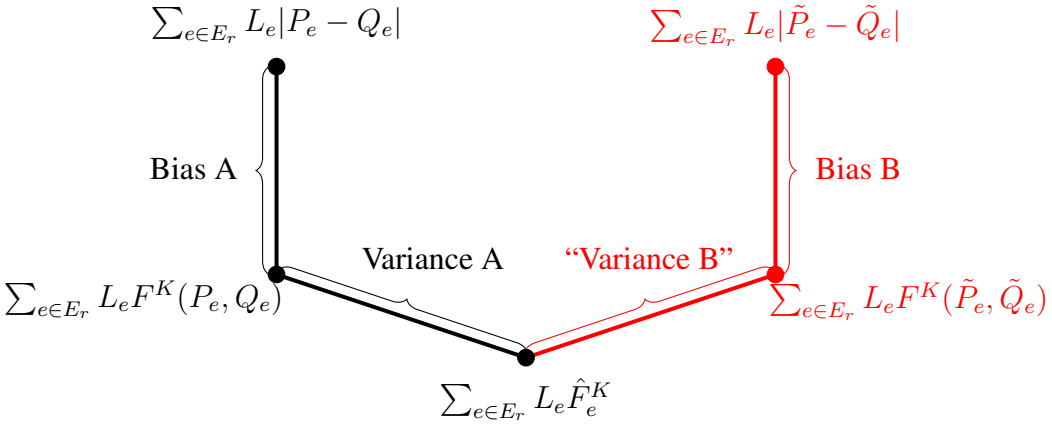


Figure 2: Bias and variance decomposition of polynomial approximation estimator and moment screening estimator.

However, one of the main difficulties of this strategy is that we need to construct the best polynomial approximation explicitly for each edge and each target functional. For instance, the best approximated polynomials for $|x-y|^\alpha$ are different for different $\alpha > 0$. To address this issue, we appeal to the observation that the unbiased estimator $\hat{F}_e^K$ can be further approximated by some plug-in estimator. To illustrate the intuition, we consider a simple example that $F_e^K(P_e, Q_e)$ are the same across different $e \in E_r$, i.e. there exists a polynomial $F^K(P_e, Q_e)$ such that $F_e^K(P_e, Q_e) = F^K(P_e, Q_e)$. Suppose $(\tilde{P}_e, \tilde{Q}_e)$ are chosen in a way such that

$$\left| \sum_{e \in E_r} L_e F^K(P_e, Q_e) - \sum_{e \in E_r} L_e \hat{F}_e^K \right| \approx \left| \sum_{e \in E_r} L_e F^K(\tilde{P}_e, \tilde{Q}_e) - \sum_{e \in E_r} L_e \hat{F}_e^K \right|. \tag{4}$$

In other words, variance A and "variance B" are almost equal in Figure 2. Since the approximation error of chosen polynomial is under control uniformly, bias A and bias B in Figure 2 have the same order. Combining bias and variance term implies

$$\left| \sum_{e \in E_r} L_e |P_e - Q_e| - \sum_{e \in E_r} L_e \hat{F}_e^K \right| \approx \left| \sum_{e \in E_r} L_e |\tilde{P}_e - \tilde{Q}_e| - \sum_{e \in E_r} L_e \hat{F}_e^K \right|.$$

This suggests that the plug-in estimator $\sum_{e \in E_r} L_e |\tilde{P}_e - \tilde{Q}_e|$ is as efficient as polynomial approximation estimator $\sum_{e \in E} L_e \hat{F}_{e,K}$, but has no need for explicit knowledge of the best approximated polynomial.

To search for $(\tilde{P}_e, \tilde{Q}_e)$ satisfying (4), it is sufficient to consider each monomial of $F^K(P_e, Q_e)$. More specifically, $(\tilde{P}_e, \tilde{Q}_e)$ can be chosen in a way such that

$$\left| \sum_{e \in E_r} L_e \left(\tilde{P}_e^{k_1} \tilde{Q}_e^{k_2} - H_{k_1}(\hat{P}_{e,1}) H_{k_2}(\hat{Q}_{e,1}) \right) \right| \approx \left| \sum_{e \in E_r} L_e \left(P_e^{k_1} Q_e^{k_2} - H_{k_1}(\hat{P}_{e,1}) H_{k_2}(\hat{Q}_{e,1}) \right) \right| \tag{5}$$

for $k_1, k_2 = 0, \ldots, K$. The way to choose $(\tilde{P}_e, \tilde{Q}_e)$ in (5) is referred as moment screening hereafter. Because of no need for explicit construction of best approximated polynomial, we adopt moment screening strategy to improve the classical plug-in estimator.

3.3 Moment-screening Estimator on Tree (MET)

Although the above approximation strategy could help reduce the bias, the main difficulty of moment screening is that we need to estimate the deviation of the unbiased estimator on the right

hand side of (5) in the presence of heteroskedastic variance and the complex dependence structure among $(\hat{P}_{e,1}, \hat{Q}_{e,1})$. To address this challenge, we adopt the following proposition to decouple the dependence structure among $(\hat{P}_{e,1}, \hat{Q}_{e,1})$.

Proposition 2. *Suppose $\{x_v\}_{v\in V}$ is a collection of non-negative number such that $\sum_{v\in V} x_v \leq W$. Let $\tilde{E}(w)$ be a subset of edges on tree T such that*

$$\tilde{E}(w) = \left\{ e \in E : w/2 < \sum_{v\in\tau(e)} x_v \leq w \right\}.$$

Then, $\tilde{E}(w)$ can be decomposed into a collection of disjoint paths

$$\tilde{E}(w) = \bigcup_{l=1}^{S} [v_l^L, v_l^U],$$

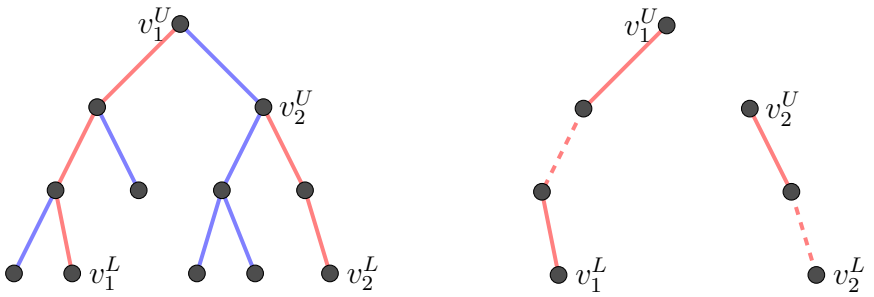
where v_l^L and v_l^U $l = 1, \ldots, S$, are the nodes of the tree, and the number of disjoint paths S satisfies $S \leq 2W/w$. In addition, any two edges from different paths in the above decomposition do not share any descendants.

This proposition suggests that if subset of E can be written in the form of $\tilde{E}(w)$ for some $\{x_v\}_{v\in V}$, then it can be decomposed into a collection of disjoint paths. A typical example of $\tilde{E}(w)$ is colored in red in Figure 3a, where $\tilde{E}(w)$ is decomposed into two paths $[v_1^L, v_1^U]$ and $[v_2^L, v_2^U]$. Because of this decomposition, the dependence structure of $(\hat{P}_{e,1}, \hat{Q}_{e,1})$ is clear on $\tilde{E}(w)$. To be specific, $(\hat{P}_{e,1}, \hat{Q}_{e,1})$ are highly dependent for any two edges on the same path, but $(\hat{P}_{e,1}, \hat{Q}_{e,1})$ are independent for any two edges from different path since they do not share any descendants.

Motivated by Proposition 2, we decompose E_r with respect to the value of $\hat{P}_{e,0} + \hat{Q}_{e,0} = \sum_{v\in\tau(e)} (\hat{p}_{v,0} + \hat{q}_{v,0})$. We consider the following stratification of E_r

$$E_j = \left\{ e \in E_r : \frac{1}{2^j} < \hat{P}_{e,0} + \hat{Q}_{e,0} \leq \frac{1}{2^{j-1}} \right\},$$

for $j = 1, \ldots, J := \lfloor \log_2 (n/c_1 \log n) \rfloor$. By definition, E_j is a subset of $\{e \in E : 2^{-j} < \hat{P}_{e,0} + \hat{Q}_{e,0} \leq 2^{-(j-1)}\}$, which satisfies the condition of Proposition 2. Therefore, each E_j can be decomposed into a collection of subsets of disjoint paths and has a clear dependence structure as



(a) An example of $\tilde{E}(w)$ (b) An example of E_j

Figure 3: Examples of $\tilde{E}(w)$ and E_j: the red edges in (a) shows an example of $\tilde{E}(w)$ with respect to $\hat{P}_{e,0} + \hat{Q}_{e,0}$, which can be decomposed two paths $[v_1^L, v_1^U]$ and $[v_2^L, v_2^U]$; solid red edges in (b) shows an example of E_j, which is a subset of $\tilde{E}(w)$ in (a); dashed red edges in (b) belong to E_c.

well. Figure 3b shows a typical example of E_j. Besides each E_j, $j = 1, \ldots, J$, we also define

$$E_0 = \left\{ e \in E_r : \hat{P}_{e,0} + \hat{Q}_{e,0} \leq 2^{-J} \right\}.$$

Since E_0 does not satisfy the conditions of Proposition 2, the bounded difference property can be used to estimate deviation in (5). We are now in a position to carry out moment screening in (5) for each E_j.

We define the following set for E_0

$$I_0 = \left\{ \{(x_e, y_e)\}_{e \in E_0} \in A_0^{|E_0|} : \left| \sum_{e \in E_0} L_e \left(x_e^{k_1} y_e^{k_2} - \hat{H}_{k_1,k_2} \right) \right| \leq R_{0,k_1,k_2}, 0 \leq k_1, k_2 \leq K \right\},$$

where $\hat{H}_{k_1,k_2} = H_{k_1}(\hat{P}_{e,1}) H_{k_2}(\hat{Q}_{e,1})$, $K = c_2 \log n$ for some constant c_2 and deviation in the right hand side of (5) is

$$R_{0,k_1,k_2} = C_{0,k_1,k_2} d \sqrt{n \log^2 n} \left(\frac{\log n}{n} \right)^{k_1+k_2}.$$

Here, d is the height of tree T and $C_{0,k_1,k_2} = 4dM^2(76c_1)^{k_1+k_2}$, where $M = \max_{e \in E} L_e$. A_0 is defined as small 2D region $A_0 = [0, 2c_1 \log n/n]^2$ so that each (x_e, y_e) is constrained in this region. We choose $\{(\tilde{P}_e, \tilde{Q}_e)\}_{e \in E_0}$ arbitrarily from I_0. When $P_e + Q_e$ is large than $c_1 \log n/n$, $|P_e - Q_e|$ can be well approximated by a polynomial of $P_e - Q_e$ (see Jiao, Han, and Weissman, 2018). Following this observation, we shall conduct the moment screening for $P_e - Q_e$ directly in order to simplify the moment screening procedure. In particular, $\{(\tilde{P}_e - \tilde{Q}_e)\}_{e \in E_j}$ is chosen

directly from the following set

$$I_j = \left\{ \{x_e\}_{e\in E_j} \in A_j^{|E_j|} : \left| \sum_{e\in E_j} L_e \left(x_e^k - G_k(\hat{P}_{e,1}, \hat{Q}_{e,1}) \right) \right| \le R_{j,k},\ 0 \le k \le K \right\},$$

where $G_k(\hat{P}_{e,1}, \hat{Q}_{e,1})$ is unbiased estimator of $(P_e - Q_e)^k$

$$G_k(\hat{P}_{e,1}, \hat{Q}_{e,1}) = \sum_{l=0}^{k} \binom{k}{l} (-1)^l H_l(\hat{P}_{e,1}) H_{k-l}(\hat{Q}_{e,1}),$$

and the deviation can be represented as

$$R_{j,k} = C_{j,k} d \sqrt{S_j \log n} \left(\frac{\log n}{2^j n} \right)^{k/2}.$$

Here, A_j is defined as an interval $[-\sqrt{4c_1 \log n/2^j n}, \sqrt{4c_1 \log n/2^j n}]$, S_j is the number of the disjoint paths in E_j in Proposition 2 and $C_{j,k} = 6dM^2(48c_1)^{k/2}$. By doing this, instead of $(K+1)^2$ constraints, only $K+1$ constraints are required to choose $(\tilde{P}_e, \tilde{Q}_e)$. After moment screening, we plugin $(\tilde{P}_e, \tilde{Q}_e)$ into estimator $D(P, Q)$ to obtain our new estimator

$$\tilde{D}_{\mathrm{MET}} = \sum_{j=0}^{J} \sum_{e\in E_j} L_e |\tilde{P}_e - \tilde{Q}_e| + \sum_{e\in E_c} L_e |\hat{P}_{e,1} - \hat{Q}_{e,1}|. \tag{6}$$

In light of the fact that $0 \le D(P, Q) \le dM$, our final estimator is defined as

$$\hat{D}_{\mathrm{MET}} = \min(\tilde{D}_{\mathrm{MET}}, dM).$$

We call this estimator the moment-screening estimator on tree (MET). The corresponding algorithm is summarized in Algorithm 1.

We close this section by comparing MET with local moment matching (LMM) proposed by Han, Jiao, and Weissman (2018). Although both methods adopt implicit approximation method through comparing monomials after "localization" of (P_e, Q_e), there are several key differences between two methods. LMM is designed for estimating symmetric functionals of a single discrete distribution adaptively, while MET aims to estimate a collection of asymmetrical distances between a pair of discrete distributions on a tree. Because of different purposes, MET incorporates branch length L_e in moment screening (5) and adopts a new scheme of partition for edges of tree in order

Algorithm 1 Moment-screening Estimator on Tree (MET)

Input: Empirical distribution $\{\hat{P}_e\}_{e\in E}$, $\{\hat{Q}_e\}_{e\in E}$ and Tree T.
Output: Estimation of distance $D(P, Q)$.
Split samples into $\{\hat{P}_{e,0}, \hat{Q}_{e,0}\}_{e\in E}$ and $\{\hat{P}_{e,1}, \hat{Q}_{e,1}\}_{e\in E}$.
Use $\{\hat{P}_{e,0}, \hat{Q}_{e,0}\}_{e\in E}$ to group edges into E_j, $0 \le j \le J$ and E_c.
if $I_0 = \emptyset$ **then**
$\{(\tilde{P}_e, \tilde{Q}_e)\}_{e\in E_0} = \{(\hat{P}_{e,1}, \hat{Q}_{e,1})\}_{e\in E_0}$
else
Choose $\{(\tilde{P}_e, \tilde{Q}_e)\}_{e\in E_0}$ from I_0 arbitrarily.
end if
for $j \in 1 : J$ **do**
if $I_j = \emptyset$ **then**
$\{(\tilde{P}_e - \tilde{Q}_e)\}_{e\in E_j} = \{(\hat{P}_{e,1} - \hat{Q}_{e,1})\}_{e\in E_j}$
else
Choose $\{(\tilde{P}_e - \tilde{Q}_e)\}_{e\in E_j}$ from I_j arbitrarily.
end if
end for
Evaluate $\hat{D}_{\mathrm{MET}} = \min(\sum_{j=0}^{J} \sum_{e\in E_j} L_e|\tilde{P}_e - \tilde{Q}_e| + \sum_{e\in E_c} L_e|\hat{P}_{e,1} - \hat{Q}_{e,1}|, dM)$.
return $\hat{D}_{\mathrm{MET}}$

to account for complex dependence structure among $(\hat{P}_e, \hat{Q}_e)$ and 2 dimensional nature of the problem.

4 Theoretical Properties

We now turn to analyzing the theoretical properties of proposed estimator $\hat{D}_{\mathrm{MET}}$, and compare it with the classical plug-in estimator. We evaluate the performance of an estimator $\hat{D}$ based on the samples $\boldsymbol{X}$ and $\boldsymbol{Y}$ by the maximum mean squared error

$$R(\hat{D}; \Theta) := \sup_{(T,P,Q)\in\Theta} \mathbb{E}(\hat{D}(\boldsymbol{X}, \boldsymbol{Y}) - D(P, Q))^2,$$

where parameter set Θ is a collection of combinations of tree T and probability distributions P and Q. The minimax risk in estimating $D(P, Q)$ is defined as

$$R^*(\Theta) = \inf_{\hat{D}} R(\hat{D}; \Theta),$$

where the infimum is taken with respect to all measurable estimators based on the samples $\boldsymbol{X}$ and $\boldsymbol{Y}$. In particular, we are interested in the following parameter set

$$\Theta(s,d) := \left\{\theta = (T,P,Q) : T \in \mathcal{T}(s,d), P, Q \in \mathcal{M}_{|V|}\right\},$$

where $\mathcal{M}_s$ is collection of all discrete distribution with alpha-beta size s and

$$\mathcal{T}(s,d) = \left\{T : 1 \le L_e \le M, \forall e \in E; d(T) \le d; |V| \le s\right\}.$$

The choices of $M \ge 1$ in above definition is arbitrarily, but need to be a fixed constant. For simplicity of analysis, we shall focus on the case when T is a binary tree, i.e. each node has at most two children, although all the analysis can also be applied to the more general cases. The parameter space $\Theta(s,d)$ requires that the number of the nodes of the tree T is less than or equal to s and the depth of the tree T is less than or equal to d. Clearly, an implicit constraint for d and s is $\log_2 s \le d \le Ms$ due to the facts that the shortest tree T is a complete binary tree and highest tree is a chain. For brevity, we also write $R^*(\hat{D}; \Theta(s,d))$ as $R^*(\hat{D}; s, d)$ and $R^*(\Theta(s,d))$ as $R^*(s,d)$.

The performance of MET on $\Theta(s,d)$ is characterized by the following theorem.

Theorem 1. *Consider estimating $D(P,Q)$ by $\hat{D}_{\mathrm{MET}}$ on $\Theta(s,d)$. Let $K = c_2 \log n$ for some constant $c_2 < c_1$ and $c_1 > 40$. If $n \log n \gg s \log(2^{d+2}/s)$ and $\log n \le C_1 \log(s/d)$ for arbitrary constant C_1, then there exists a constant C such that*

$$R(\hat{D}_{\mathrm{MET}}; s, d) \le C \frac{s \log(2^{d+2}/s)}{n \log n},$$

when c_2 is chosen small enough.

One main challenge in the proof of Theorem 1 is that $(\tilde{P}_e, \tilde{Q}_e)$s can be highly dependent. To decouple the high dependence, Proposition 2 helps to segment the collections of nodes into relatively independent ones by taking advantage of the tree structure. The analysis itself may be of independent interest and can be applied to other problems on trees such as deriving the asymptotic distribution for the test statistics of UniFrac distance. Theorem 1 assumes an upper bound condition on sample size ($\log n \le C_1 \log(s/d)$). The similar upper bound of sample size also appears in previous papers (see, e.g. Jiao et al., 2015; Wu and Yang, 2016; Jiao, Han, and

Weissman, 2018). Under this kind of condition, the bias of estimator dominates its variance so the plugin estimator can be improved by bias reduction. Another implication of this condition is $\log n \asymp \log(s/d)$ when $\log n \leq C_1 \log(s/d)$, thus the result in Theorem 1 is still valid if we replace $\log n$ by $\log(s/d)$.

The following lower bound shows that MET as shown in Theorem 1 is indeed rate-optimal.

Theorem 2. *Consider estimating $D(P,Q)$ on $\Theta(s,d)$. We have*

$$\inf_{\hat{D}} R(\hat{D}; s, d) \geq c \frac{s \log(2^{d+2}/s)}{n \log n}$$

for some constant $c > 0$. Moreover, if $n \log n \ll s \log(2^{d+2}/s)$, there is no consistent estimator for $D(P,Q)$.

Theorems 1 and 2 together show that the optimal rate of convergence under condition $\log n \leq C_1 \log(s/d)$ is

$$R^*(s,d) \asymp \frac{s \log(2^{d+2}/s)}{n \log n}.$$

Depending on d and s, there are two different regimes for the minimax optimal rate.

- For the short trees where $s \asymp 2^d$, the optimal rate for estimating $D(P,Q)$ is $s/n \log n$. Recall that estimating the L_1 distance has the same minimax optimal rate (see Jiao, Han, and Weissman, 2018). Putting differently, estimating $D(P,Q)$ is as difficult as estimating the L_1 distance when the tree is short enough, i.e. almost like a complete binary tree.

- For the trees of tall heights, i.e. $s \ll 2^d$, the optimal rate becomes $sd/n \log n$. We can see that the distance on taller tree is more difficult to estimate.

We now compare the performance of $\hat{D}_{\mathrm{MET}}$ and the classical plug-in estimator $D(\hat{P}, \hat{Q})$. The performance of the classical plug-in estimator $D(\hat{P}, \hat{Q})$ on $\Theta(s,d)$ is characterized in the following theorem.

Theorem 3. *Let $D(\hat{P}, \hat{Q})$ be the classical plug-in estimator for $D(P, Q)$. When $n \ll s\log(2^{d+2}/s)$, $D(\hat{P}, \hat{Q})$ is inconsistent. And when $n \gg s\log(2^{d+2}/s)$, there exist constants c and C such that*

$$c\frac{s\log(2^{d+2}/s)}{n} \leq R(D(\hat{P}, \hat{Q}); s, d) \leq C\frac{s\log(2^{d+2}/s)}{n}.$$

Comparison between Theorems 1 and 3 shows that the accuracy of MET is better than the classical plug-in estimator when $\log n \leq C_1 \log(s/d)$. In particular, the loss of $D(\hat{P}, \hat{Q})$ is inflated by $\log n$ times, although it is much simpler to implement plug-in estimator. If the tree is fixed, i.e. s and d are determined, the minimax risk is of the order n^{-1}, which is consistent with asymptotic results in Sommerfeld and Munk (2018). We provide a more accurate characterization when s and d increase along with n.

5 Estimation of the L^α Zolotarev-type Distance

Evans and Matsen (2012) generalize $D(P, Q)$ into L^α Zolotarev-type distance

$$\left(\sum_{e\in E} L_e \left|P_e - Q_e\right|^\alpha\right)^{(1/\alpha)\wedge 1}$$

where $0 < \alpha < \infty$. To fix the idea, we focus on estimating its equivalent form

$$D_\alpha(P, Q) = \sum_{e\in E} L_e \left|P_e - Q_e\right|^\alpha. \tag{7}$$

It is clear that, as a special case, $D_1(P, Q)$ is just the Wasserstein distance we discussed in the previous sections. As pointed by Fukuyama et al. (2012), $D_2(P, Q)$ is the distance used in DP-CoA (see, Pavoine, Dufour, and Chessel, 2004). To assess the performance of an estimator $\hat{D}$ of $D_\alpha(P, Q)$, we still adopt the mean squared error on $\Theta(s, d)$

$$R_\alpha(\hat{D}; s, d) := \sup_{(T,P,Q)\in\Theta(s,d)} \mathbb{E}(\hat{D} - D_\alpha(P, Q))^2.$$

The corresponding minimax risk can then be defined as $R^*_\alpha(s, d) := \inf_{\hat{D}} R_\alpha(\hat{D}; s, d)$.

Through the discussion in Section 2, the main reason for the inflated bias in the classical plug-in estimator for $D(P, Q)$ is that the approximation error for $|x - y|$ when (x, y) lies around the diagonal line $x = y$. This naturally brings about the question of whether the classical plug-

in estimator for $D_\alpha(P,Q)$ also suffers from the same bias problem. We show that this actually depends on the choice of α, determining the smoothness of $|x-y|^\alpha$ at $x=y$. More specifically, we first show that, when $\alpha \geq 2$, the bias inflation is negligible so that the classical plug-in estimator achieves the minimax optimal rate.

Theorem 4. *Consider estimating $D_\alpha(P,Q)$ by the classical plug-in estimator $D_\alpha(\hat{P},\hat{Q})$ on $\Theta(s,d)$. If we assume $\alpha \geq 2$, then there exist constants C and c such that*

$$R_\alpha(D_\alpha(\hat{P},\hat{Q});s,d)) \leq C\frac{d^2}{n} \qquad \text{and} \qquad \inf_{\hat{D}} R_\alpha(\hat{D};s,d) \geq c\frac{d^2}{n}.$$

Furthermore, there is no consistent estimator for $D_\alpha(P,Q)$ when $n \ll d^2$.

The reason that the classical plug-in estimator attains the optimal rate when $\alpha \geq 2$ is that the function $|x-y|^\alpha$ is smooth in this case. Obviously, the smaller α is, the sharper $|x-y|^\alpha$ at $x=y$ becomes. This leads to significant bias for the classical plug-in estimator when $0<\alpha<2$. One might expect to only need to adopt the same approximation strategy in Section 2 to reduce bias when (P_e,Q_e)s are around diagonal line. It turns out that it is also necessary to reduce the bias of $|\hat{P}_e-\hat{Q}_e|^\alpha$ even when it lies outside of the adjacent region of the diagonal line.

To address this issue, we reduce the bias of $|\hat{P}_e-\hat{Q}_e|^\alpha$s in two steps. Specifically, the same moment screening strategy is used to reduces the bias, when $(\hat{P}_{e,0},\hat{Q}_{e,0}) \in \cup_{j=0}^{J} E_j$. Since no explicit polynomial construction is required by MET, we can simply plugin $(\tilde{P}_e,\tilde{Q}_e)$ into $|P_e-Q_e|^\alpha$ to achieve the bias reduction. On the other hand, another step based on the Taylor expansion is adopted when $(\hat{P}_{e,0},\hat{Q}_{e,0}) \in E_c$. In particular, the Taylor expansion suggests

$$\mathbb{E}|\hat{P}_{e,1}-\hat{Q}_{e,1}|^\alpha - |P_e-Q_e|^\alpha \approx \frac{\alpha(\alpha-1)}{2}|P_e-Q_e|^{\alpha-2}\mathrm{Var}(\hat{P}_{e,1}-\hat{Q}_{e,1}).$$

Thus, we consider the following first order bias-corrected estimator for $|P_e-Q_e|^\alpha$

$$U_\alpha(\hat{P}_{e,1},\hat{Q}_{e,1}) = |\hat{P}_{e,1}-\hat{Q}_{e,1}|^\alpha + \frac{\alpha(1-\alpha)}{2n}I_n(\hat{P}_{e,1},\hat{Q}_{e,1})|\hat{P}_{e,1}-\hat{Q}_{e,1}|^{\alpha-2}(\hat{P}_{e,1}+\hat{Q}_{e,1}).$$

Here, $I_n(P_e,Q_e) = \mathbf{I}_{(|P_e-Q_e|>\sqrt{c_1(P_e+Q_e)\log n/4n})}\mathbf{I}_{(P_e+Q_e>c_1\log n/4n)}$ is a truncation function designed to make $U_\alpha(\hat{P}_{e,1},\hat{Q}_{e,1})$ a bounded function. This is inspired by interpolation function in Jiao et al. (2015). Putting the two bias reduction steps together yields our MET estimator for

$D_\alpha(P,Q)$

$$\hat{D}_{\mathrm{MET},\alpha} := \sum_{j=0}^{J} \sum_{e \in E_j} L_e |\tilde{P}_e - \tilde{Q}_e|^\alpha + \sum_{e \in E_c} L_e U_\alpha(\hat{P}_{e,1}, \hat{Q}_{e,1}). \tag{8}$$

It is worth noting that $\hat{D}_{\mathrm{MET},1}$ coincides with the estimator in (6). When $0 < \alpha < 1$, the performance of $\hat{D}_{\mathrm{MET},\alpha}$ can be characterized in the following theorem.

Theorem 5. *Let $\hat{D}_{\mathrm{MET},\alpha}$ be the estimator defined in* (8)*. Let $K = c_2 \log n$ for some small enough constant c_2 and $c_1 > 40$. For $0 < \alpha < 1$, there exist constants $C > 0$ and $c > 0$ such that*

$$R_\alpha(\hat{D}_{\mathrm{MET},\alpha}; s, d) \le C \frac{s^{2-\alpha} \log^\alpha(2^{d+2}/s)}{(n \log n)^\alpha},$$

when $n \log n \gg s^{(2-\alpha)/\alpha} \log(2^{d+2}/s)$ and

$$\inf_{\hat{D}} R_\alpha(\hat{D}; s, d) \ge c \frac{s^{2-\alpha} \log^\alpha(2^{d+2}/s)}{(n \log n)^\alpha}.$$

Moreover, no consistent estimator for $D_\alpha(P,Q)$ exists when $n \log n \ll s^{(2-\alpha)/\alpha} \log(2^{d+2}/s)$.

Theorem 5 shows that the minimax rate is dominated by bias and $\hat{D}_{\mathrm{MET},\alpha}$ is a minimax rate-optimal estimator for $D_\alpha(P,Q)$ when $0 < \alpha < 1$. We now show that the bias and variance dominate in different regimes when $1 < \alpha < 2$. In the following theorem, we write $r(s,d,n) = \log n / \log(s/d)$ and $T(\alpha) = (2-\alpha)/(\alpha-1)$.

Theorem 6. *Consider estimating $D_\alpha(P,Q)$ when $1 < \alpha < 2$. Let $\hat{D}_{\mathrm{MET},\alpha}$ be the estimator defined in* (8) *and $D(\hat{P},\hat{Q})$ be the plugin estimator. For the MET, we assume $K = c_2 \log n$ for some small enough constant c_2 and $c_1 > 40$. If $r(s,d,n) \le C_1 < T(\alpha)$ and $n \log n \gg s^{(2-\alpha)/\alpha} \log(2^{d+2}/s)$, then there exist constants $C > 0$ such that*

$$R_\alpha(\hat{D}_{\mathrm{MET},\alpha}; s, d) \le C \frac{s^{2-\alpha} \log^\alpha(2^{d+2}/s)}{(n \log n)^\alpha}.$$

On the other hand, if $r(s,d,n) \ge T(\alpha)$ and $n \gg d^2$, then

$$R_\alpha(D_\alpha(\hat{P},\hat{Q}); s, d) \le C \frac{d^2}{n}.$$

Furthermore, there exists a small constant c such that

$$\inf_{\hat{D}} R_\alpha(\hat{D}; s, d) \geq c \left(\frac{s^{2-\alpha} \log^\alpha (2^{d+2}/s)}{(n \log n)^\alpha} + \frac{d^2}{n} \right).$$

Moreover, no consistent estimator for $D_\alpha(P, Q)$ exists when $n \log n \ll \max(s^{(2-\alpha)/\alpha} \log(2^{d+2}/s), d^2)$.

This theorem suggests that a composite estimator shall be adopted to achieve the minimax optimal rate of estimating $D_\alpha(P, Q)$ when $1 < \alpha < 2$. More specifically, a new composite estimator of MET and plugin estimator can be defined as

$$\hat{D}_{\mathrm{COM},\alpha} = \begin{cases} \hat{D}_{\mathrm{MET},\alpha} & r(s, d, n) < T(\alpha) \\ D_\alpha(\hat{P}, \hat{Q}) & r(s, d, n) \geq T(\alpha) \end{cases}.$$

Theorem 6 shows that $\hat{D}_{\mathrm{COM},\alpha}$ is minimax rate-optimal either $r(s, d, n) < T(\alpha)$ or $r(s, d, n) \geq T(\alpha)$. Theorems 4, 5 and 6 together characterize the minimax rate $R^*_\alpha(s, d)$ for different value of α. We summarize the minimax rate and optimal sample complexity in Table 1. The minimax rates suggest that the estimation of $D_\alpha(P, Q)$ becomes more difficult as α decreases. Putting it differently, given s, d and n, $D_\alpha(P, Q)$ is only estimable for some value of α.

Table 1: Summary of the minimax risk rate and optimal sample complexity.

	$0 < \alpha < 1$	$1 < \alpha < 2$	$\alpha \geq 2$
Minimax Rate	$\frac{s^{2-\alpha} \log^\alpha (2^{d+2}/s)}{(n \log n)^\alpha}$	$\frac{s^{2-\alpha} \log^\alpha (2^{d+2}/s)}{(n \log n)^\alpha} + \frac{d^2}{n}$	$\frac{d^2}{n}$
Sample Complexity	$\frac{s^{(2-\alpha)/\alpha} \log(2^{d+2}/s)}{\log s}$	$\max\left(\frac{s^{(2-\alpha)/\alpha} \log(2^{d+2}/s)}{\log s}, d^2 \right)$	d^2

6 Implementation of MET

We now address several practical issues in implementation of MET in this section. As mentioned earlier, the purpose of sample splitting is mainly for simplifying the theoretical analysis. We do not split the samples in the numerical experiments. In other words, we replace the role of both $(\hat{P}_{e,0}, \hat{Q}_{e,0})$ and $(\hat{P}_{e,1}, \hat{Q}_{e,1})$ by $(\hat{P}_e, \hat{Q}_e)$. To implement MET, we need to assign appropriate values

for two turning parameters: c_1 and c_2. The choices of c_1 and c_2 are crucial to tradeoff between bias and variance, because c_1 affects the size of uncertain set $\mathcal{P}$ and the deviation of moment screening R_{0,k_1,k_2} and $R_{j,k}$, and c_2 determines the degree of moment screening K. A combination of larger c_1 and c_2 implies applying bias reduction to more pairs of (P_e, Q_e) and higher degree of approximated polynomial in MET, thus leading to smaller approximation error along with increasing variance. On the other hand, the variance is well controlled when both c_1 and c_2 are chosen to be small. The choice of $c_1 = 1.8$ and $c_2 = 1$ is supported by our experience and will be used in all numerical experiments in Section 7.

The core component of MET is the moment screening: choosing $(\tilde{P}_e, \tilde{Q}_e)$ or $(\tilde{P}_e - \tilde{Q}_e)$ from I_j for $j = 0, \ldots, J$. The main difficulty is to check the feasibility of each I_j as the constraints of I_j are highly non-linear. To overcome this issue, we formulate the moment screening as a linear programming. A similar formulation has been used in Han, Jiao, and Weissman (2018). More specifically, for E_0, the square $[0, c_1 \log n/n] \times [0, c_1 \log n/n]$ is divided into a collection of small bins with the width w_0: $H_{h_1,h_2} := [(h_1 - 1)w_0, h_1 w_0) \times [(h_2 - 1)w_0, h_2 w_0)$, $1 \le h_1, h_2 \le \lceil c_1 \log n / w_0 n \rceil$. The width of w_0 can be chosen as $1/2n$. We assign a weight W_{h_1,h_2} for each H_{h_1,h_2} and write

$$W^o_{h_1,h_2} = \sum_{e \in E_0} L_e \mathbf{I}((\hat{P}_e, \hat{Q}_e) \in H_{h_1,h_2}).$$

Thus, I_0 can be approximated by

$$\tilde{I}_0 = \left\{ \{W_{h_1,h_2}\} : \left| \sum_{h_1,h_2} W_{h_1,h_2} (h_1 w_0)^{k_1} (h_2 w_0)^{k_2} - \sum_{e \in E_0} L_e H_{k_1}(\hat{P}_e) H_{k_2}(\hat{Q}_e) \right| \le R_{0,k_1,k_2}, 0 \le k_1, k_2 \le K \right\},$$

and the feasibility of $\tilde{I}_0$ can then be checked by linear programming. To make the choice of W_{h_1,h_2} more stable, we consider the following optimization problem

$$\begin{aligned} &\min_{W_{h_1,h_2}} \sum_{h_1,h_2} |W_{h_1,h_2} - W^o_{h_1,h_2}| \\ &\text{s.t.} \left| \sum_{h_1,h_2} W_{h_1,h_2} (h_1 w_0)^{k_1} (h_2 w_0)^{k_2} - \sum_{e \in E_0} L_e H_{k_1}(\hat{P}_e) H_{k_2}(\hat{Q}_e) \right| \le R_{0,k_1,k_2} \\ &\qquad \text{for all } 0 \le k_1, k_2 \le K. \end{aligned}$$

Algorithm 2 Algorithm for Implementing the Moment-screening Estimator on Tree (MET)

Input: Empirical distributions $\{\hat{P}_e\}_{e\in E}$, $\{\hat{Q}_e\}_{e\in E}$ and Tree T.
Output: Estimate of the distance $D(P,Q)$.

Use $\{\hat{P}_e, \hat{Q}_e\}_{e\in E}$ to group edges into E_j, $0 \le j \le J$ and E_c.

Solve

$$\min_{W_{h_1,h_2}} \sum_{h_1,h_2} |W_{h_1,h_2} - W^o_{h_1,h_2}|$$
$$\text{s.t.} \left| \sum_{h_1,h_2} W_{h_1,h_2}(h_1 w_0)^{k_1}(h_2 w_0)^{k_2} - \sum_{e\in E_0} L_e H_{k_1}(\hat{P}_e) H_{k_2}(\hat{Q}_e) \right| \le R_{0,k_1,k_2}$$
$$\text{for all } 0 \le k_1, k_2 \le K$$

if above problem is infeasible **then**
$W_{h_1,h_2} = W^o_{h_1,h_2}$.
end if
$D_0 = \sum_{h_1,h_2} W_{h_1,h_2} |(h_2 - h_1)w_0|^\alpha$.
for $j \in 1 : J$ **do**
Solve

$$\min_{W_h} \sum_h |W_h - W^o_h|$$
$$\text{s.t.} \left| \sum_h W_h (h w_j)^k - \sum_{e\in E_j} L_e G_k(\hat{P}_e, \hat{Q}_e) \right| \le R_{j,k}, \qquad \text{for all } 0 \le k \le K.$$

if above problem is infeasible **then**
$W_h = W^o_h$.
end if
$D_j = \sum_h W_h |h w_j|^\alpha$.
end for
Evaluate $\hat{D}_{\mathrm{MET},\alpha} = \sum_{j=0}^{J} D_j + \sum_{e\in E_c} L_e U_\alpha(\hat{P}_e, \hat{Q}_e)$.
return $\hat{D}_{\mathrm{MET},\alpha}$

Because of the L_1 minimization, the sparse structure of $W^o_{h_1,h_2}$ is kept by the above optimization form. This optimization problem is equivalent to a linear programming that can be solved efficiently. We use the optimization software MOSEK (`https://www.mosek.com/`) to solve the corresponding linear program. After calculating W_{h_1,h_2}, $\sum_{e\in E_0} L_e |\tilde{P}_e - \tilde{Q}_e|^\alpha$ thus can be approx-

imated by

$$\sum_{h_1,h_2} W_{h_1,h_2} \left|(h_2 - h_1)w_0\right|^\alpha .$$

We use the same strategy to find a solution for $\sum_{e\in E_j} L_e|\tilde{P}_e - \tilde{Q}_e|^\alpha$. Specifically, we divided the interval $[-\sqrt{4c_1 \log n/2^j n}, \sqrt{4c_1 \log n/2^j n}]$ as a collection of small bins $H_h := [(h-1)w_j, hw_j)$ with the width $w_j = 1/\sqrt{2^j n}$. For each H_h, we define the weights W_h and $W_h^o = \sum_{e\in E_j} L_e \mathbf{I}(\hat{P}_e - \hat{Q}_e \in H_h)$. The optimization problem is then

$$\begin{aligned} &\min_{W_h} \sum_h |W_h - W_h^o| \\ &\text{s.t.} \left| \sum_h W_h (hw_j)^k - \sum_{e\in E_j} L_e G_k(\hat{P}_e, \hat{Q}_e) \right| \le R_{j,k}, \qquad \text{for all } 0 \le k \le K. \end{aligned}$$

Thus, $\sum_{e\in E_j} L_e|\tilde{P}_e - \tilde{Q}_e|^\alpha$ can be approximated by $\sum_h W_h \left|hw_j\right|^\alpha$. After incorporating this optimization formulation, the algorithmic version of MET is summarized in Algorithm 2.

7 Numerical Studies

In this section, we study the numerical performance of the proposed MET. We carry out simulation studies in Section 7.1 and real data analysis in 7.2 to investigate the numerical properties of MET in various settings.

7.1 Simulation Studies

We first demonstrate the merit of MET through simulation studies. In particular, the tree T we use here is phylogenetic tree of bacteria within the class Gammaproteobacteria, which is extracted from Greengenes 16S rRNA database version 13.8 clustered at 85% similarity (see, DeSantis et al., 2006, `http://greengenes.secondgenome.com`) by the package metagenomeFeatures. There is a total of 247 leaves (tips), which are denoted by V_L, and 246 internal nodes, which are denoted by V_I, on tree T and the length of edges/branches ranges from 0.00015 to 0.23597. The structure of the phylogenetic tree is shown in Figure 4.

We consider three distributions on T in the next two sets of simulation experiments. Specifically, the first distribution is a uniform distribution on all nodes, i.e. $p_v = 1/493,\ \forall v \in V$, and is denoted by $P^{(1)}$. The second distribution $P^{(2)}$ we consider here is a uniform distribution on all leaves, i.e. $p_v = 1/247$ if $v \in V_L$ and $p_v = 0$ if $v \in V_I$. To define the third distribution $P^{(3)}$, we rank the leaves according to its labeled number (which is shown in Figure 4) in increasing order and write them as $v_{(1)}, \ldots, v_{(247)}$. Then, we let $p_{v_{(i)}} \propto i$ such that $\sum_{v \in V_L} p_v = 0.75$ and $p_v = 0.25/246$ if $v \in V_I$.

The first set of simulation experiments is to assess the performance of MET when two target distributions are equal, i.e. $D_\alpha(P, Q) = 0$. To this end, we simulate reads data of both samples from the same multinomial distribution. In particular, the true distributions of P and Q are $P^{(1)}$, $P^{(2)}$ and $P^{(3)}$, respectively. To investigate the effect of the sample size n and α, we chose $n = 2000, 4000, 6000, 8000$ and 10000, and $\alpha = 1, 1.5$ and 2 in the simulation experiments. The experiment is repeated 100 times for each combination of the sample size n and different distributions. For comparison purpose, both MET and the plug-in estimator are calculated for each simulation run. The average squared error $(\hat{D} - D_\alpha(P, Q))^2$ with error bar at 10% and 90% quantile are summarized in Figure 5. These results clearly demonstrate the improved accuracy of the proposed estimator MET. The observed effect of n and α is consistent with the theoretical results given in the previous sections: the Wasserstein distance can be estimated more accurately when n and α are larger.

Next, we compare the performance of MET with that of the plug-in estimator on the simulated data when $D_\alpha(P, Q) \neq 0$. We consider distance estimation between three pair of distributions: $P^{(1)}$ v.s. $P^{(2)}$, $P^{(1)}$ v.s. $P^{(3)}$ and $P^{(2)}$ v.s. $P^{(3)}$. As in the previous simulation experiments, we still vary n and α. Instead of squared error, we use the ratio of absolute error to the true distance $|\hat{D} - D_\alpha(P, Q)|/D_\alpha(P, Q)$ to assess the estimation accuracy. The simulation results in Figure 6 are based on 100 runs for each combination of the sample size n and the distribution pairs. These results again demonstrate the advantage of the proposed method MET over the simple plug-in estimator.

The last set of simulation experiments aims to further compare MET and the plug-in estimator on estimation of different distances. More specifically, we consider a mixture of uniform distributions on V_L and V_I, i.e. $p_v = 0.75/247$ if $v \in V_L$ and $p_v = 0.25/246$ if $v \in V_I$, denoted by $P^{(4)}$ hereafter. We focus on estimation of these two distances $D(P^{(2)}, P^{(2)}) = 0$ and $D(P^{(2)}, P^{(4)}) = 0.026$ when the sample size $n = 3000$. The histograms of the estimated distances by both methods are reported in Figure 7, which are based on 200 runs for each distance. The naive plug-in estimator resulted in a larger bias than MET. It is clear from Figure 7 that MET is able to better distinguish these two distances from each other than the plug-in estimator due to the bias reduction strategy in MET. This also suggests that the new estimator might be used as a more powerful test statistic to detect the difference between the two communities.

7.2 A Real Data Example

We apply MET to a 16S rRNA microbiome dataset of 16 patients with inactive Crohn's disease and 18 normal controls in order to test the intestinal microbiome difference between these two groups of individual samples. These data were collected as part of a larger microbiome study of Crohn diseases conducted at the University of Pennsylvania. For each sample, the raw sequence reads data were placed into a reference phylogenic tree from Greengenes 16S rRNA database version 13.8 with a 99% similarity by using SEPP (see Mirarab, Nguyen, and Warnow, 2012; Janssen et al., 2018). All the processing steps were performed using QIIME 2 (see, `https://qiime2.org`). After the phylogenic placement of the reads, the reference phylogenic tree is trimmed by keeping all nodes related to the operational taxonomic units (OTUs) observed in the samples. The final phylogenetic tree is shown in Figure 1. On this phylogenetic tree, there are a total of 3991 leaves (tips) and 3990 internal nodes. Before applying MET, the OTU table is normalized by rarefaction so that all the samples have the same number of reads (see Weiss et al., 2017).

The newly proposed MET is applied to calculate the microbiome distance for every pair of the 34 samples. As a comparison, we also applied the plugin estimator to calculate the UniFrac distance between samples. To compare these two methods, Figure 8 shows the difference between

the estimated distances by these two methods, $\hat{D}_{\mathrm{MET}} - D(\hat{P}, \hat{Q})$, versus the estimated distance by MET, $\hat{D}_{\mathrm{MET}}$. As shown in Figure 8, $\hat{D}_{\mathrm{MET}}$ tends to be smaller than the plug-in estimator $D(\hat{P}, \hat{Q})$ and tends to be shrunk towards 0. This is mainly due to the bias reduction technique in MET. This was observed in the third simulation experiment as well. Furthermore, the difference between the two methods tends to increase as the estimated distance decreases. This is reasonable as more pairs of (P_e, Q_e) get closer and thus result in more bias inflation of the plug-in estimator when the distance becomes smaller.

To further compare these two distance estimation methods, we conduct graph-based two sample testing by using the estimated distance matrix. Graph-based two-sample testing method is introduced by Friedman and Rafsky (1979) and further developed by Schilling (1986); Callahan et al. (2016); Chen and Friedman (2017). We first build a graph using the distance thresholding and then use the number of edges between samples from different groups as the test statistic. The graphs obtained by thresholding at 0.3 are presented in Figure 9. The statistical significance is evaluated by permuting the sample labels randomly 1000 times. The p-values calculated from the distance matrix estimated by MET and the plug-in estimator are 0.0099 and 0.0249, respectively, indicating more significant difference in overall microbiome compositions between the inactive Crohn's disease patients and the controls. If we choose the critical value at 0.01, p-value calculated from MET indicates an overall difference in gut microbiome composition between the two groups.

8 Concluding Remarks

In this paper, we considered the problem of optimal estimation of the distance between two microbial communities based on the sequencing reads that are mapped to a phylogenetic tree, including the Wasserstein distance and its L^α Zolotarev-type generalization. Although the classical plug-in estimator implemented as the UniFrac distance has been widely used in the microbiome applications, our results show that it can be sub-optimal and the accuracy can be improved by a bias reduction technique. In particular, we proposed a novel and adaptive distance estimation procedure,

MET, by adopting a polynomial approximation approach on trees. Due to the incorporation of the moment screening method, MET does not require any explicit construction of the best polynomial approximation, thus allowing estimation of $D_\alpha(P, Q)$ for multiple α simultaneously. Through this implicit approximation strategy, MET is able to reduce the bias in distance estimation effectively and hence results in minimax rate optimal estimator.

Although the main focus of this paper is estimation of the Wasserstein distance and its L^α Zolotarev-type generalization, the techniques are readily applicable to more generalized cases. For instance, the results for the L_1 distance estimation in Jiao, Han, and Weissman (2018) can be generalized to estimating the L_α distance and leads to the following optimal rate of convergence,

$$\inf_{\hat{D}} \sup_{P,Q\in\mathcal{M}_s} \mathbb{E}(\hat{D} - \|P - Q\|_\alpha^\alpha)^2 \asymp \frac{s^{2-\alpha}}{(n\log n)^\alpha} + \frac{1}{n},$$

where $\mathcal{M}_s$ is the collection of the discrete distributions with alpha-beta size s. It is also interesting to compare estimation of $D_\alpha(P, Q)$ and $\|P - Q\|_\alpha^\alpha$. The optimal rates becomes the same when $s \asymp 2^d$. In other words, when the tree is very short (almost the same with complete binary tree), the behaviors of estimation of the two distances are very similar. Another potential generalization of our result is the estimation of Wasserstein distance on a tree when the tree has at most $\lambda > 2$ children. The MET itself is still rate optimal but $\log(2^{d+2}/s)$ needs to be replaced by $\log(\lambda^{d+2}/s)$ in the optimal rate.

We focused on estimation of the Wasserstein distance between two distributions on a tree in this paper, but the Wasserstein distance between distributions on other spaces, such as $\mathbb{R}^d$ space, has been also used in many applications, including computer vision (see, e.g. Ni et al., 2009; Solomon et al., 2015) and machine learning(see, e.g. Arjovsky, Chintala, and Bottou, 2017; Gulrajani et al., 2017). The results in this paper are difficult to be generalized directly in this case as the general Wasserstein distance cannot be written in an explicit way like (2). The empirical Wasserstein distance (plug-in estimator) on finite spaces has been studied when the finite space is fixed and sample size goes to infinity (see, e.g. Do Ba et al., 2011; Weed and Bach, 2017; Sommerfeld and Munk, 2018; Tameling, Sommerfeld, and Munk, 2017; Klatt, Tameling, and Munk, 2018;

Singh and Póczos, 2018). However, our paper provides a high-dimension results to the problem of distance estimation, allowing that all d, s and n can go to infinity.

Although our discussion is mainly in the context of microbial community comparisons, it is worth noting that the Wasserstein distance may also be used in other applications. For instance, one may be interested in comparing the protein expression levels measured by the flow/mass cytometry across different cell populations (see, e.g. Orlova et al., 2016; Chen et al., 2018) when the differentiation tree of the cells is available. In practice, the differentiation tree structure across the cells can be built by several techniques such as minimum spanning tree construction or hierarchical clustering (see, e.g. Anchang et al., 2016; Mao et al., 2017; Liu et al., 2018). In these situations, the Wasserstein distance on a tree reflects the difference between the cell populations in a more accurate fashion as the similarity between the cells along the differentiation tree is taken into account. Therefore, the methodology and theory developed in this paper can then be employed in these applications as well.

FUNDING

This research was supported by NIH grants R01GM123056 and R01GM129781.

SUPPLEMENTARY MATERIALS

In the online Supplemental Materials, we prove all the theorems and relevant lemmas in this online Supplementary Materials.

References

Acharya, J. (2018), "Profile Maximum Likelihood is Optimal for Estimating KL Divergence," in *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 1400–1404.

Acharya, J., Das, H., Orlitsky, A., and Suresh, A. T. (2017), "A unified maximum likelihood approach for estimating symmetric properties of discrete distributions," in *International Conference on Machine Learning*, pp. 11–21.

Acharya, J., Orlitsky, A., Suresh, A. T., and Tyagi, H. (2014), "The complexity of estimating Rényi entropy," in *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1855–1869.

Anchang, B., Hart, T. D., Bendall, S. C., Qiu, P., Bjornson, Z., Linderman, M., Nolan, G. P., and Plevritis, S. K. (2016), "Visualization and cellular hierarchy inference of single-cell data using SPADE," *nature protocols*, 11, 1264.

Arjovsky, M., Chintala, S., and Bottou, L. (2017), "Wasserstein gan," *arXiv preprint arXiv:1701.07875*.

Bu, Y., Zou, S., Liang, Y., and Veeravalli, V. V. (2018), "Estimation of KL divergence: optimal minimax rate," *IEEE Transactions on Information Theory*, 64, 2648–2674.

Cai, T. T., and Low, M. (2011), "Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional," *The Annals of Statistics*, 39, 1012–1041.

Callahan, B. J., Sankaran, K., Fukuyama, J. A., McMurdie, P. J., and Holmes, S. P. (2016), "Bioconductor workflow for microbiome data analysis: from raw reads to community analyses," *F1000Research*, 5.

Chang, Q., Luan, Y., and Sun, F. (2011), "Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny," *BMC bioinformatics*, 12, 118.

Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., Hwang, J., Bushman, F. D., and Collman, R. G. (2010), "Disordered microbial communities in the upper respiratory tract of cigarette smokers," *PloS one*, 5, e15216.

Chen, H., and Friedman, J. H. (2017), "A new graph-based two-sample test for multivariate and object data," *Journal of the American statistical association*, 112, 397–409.

Chen, W. S., Zivanovic, N., van Dijk, D., Wolf, G., Bodenmiller, B., and Krishnaswamy, S. (2018), "Embedding the single-cell experimental variable state space to reveal manifold structure of drug perturbation effects in breast cancer," *bioRxiv*.

Daskalakis, C., Diakonikolas, I., and Servedio, R. A. (2012), "Learning k-modal distributions via testing," in *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pp. 1371–1385.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006), "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB," *Applied and environmental microbiology*, 72, 5069–5072.

Do Ba, K., Nguyen, H. L., Nguyen, H. N., and Rubinfeld, R. (2011), "Sublinear time algorithms for earth mover's distance," *Theory of Computing Systems*, 48, 428–442.

Evans, S., and Matsen, F. (2012), "The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 569–592.

Fierer, N., Hamady, M., Lauber, C. L., and Knight, R. (2008), "The influence of sex, handedness, and washing on the diversity of hand surface bacteria," *Proceedings of the National Academy of Sciences*, 105, 17994–17999.

Friedman, J. H., and Rafsky, L. C. (1979), "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests," *The Annals of Statistics*, 697–717.

Fukuyama, J., McMurdie, P. J., Dethlefsen, L., Relman, D. A., and Holmes, S. (2012), "Comparisons of distance methods for combining covariates and abundances in microbiome studies," in *Biocomputing 2012*, pp. 213–224.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017), "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, pp. 5767–5777.

Han, Y., Jiao, J., and Weissman, T. (2016), "Minimax rate-optimal estimation of divergences between discrete distributions," *arXiv preprint arXiv:1605.09124*.

——— (2018), "Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under Wasserstein distance," *arXiv preprint arXiv:1802.08405*.

Janssen, S., McDonald, D., Gonzalez, A., Navas-Molina, J. A., Jiang, L., Xu, Z., Winker, K., Kado, D. M., Orwoll, E., Manary, M., Mirarab, S., and Knight, R. (2018), "Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information," *MSystems*, 3, e00021–18.

Jiao, J., Han, Y., and Weissman, T. (2018), "Minimax estimation of the L1 distance," *IEEE Transactions on Information Theory*.

Jiao, J., Venkat, K., Han, Y., and Weissman, T. (2015), "Minimax estimation of functionals of discrete distributions," *IEEE Transactions on Information Theory*, 61, 2835–2885.

Kamath, S., Orlitsky, A., Pichapati, D., and Suresh, A. T. (2015), "On learning distributions from their samples," in *Conference on Learning Theory*, pp. 1066–1100.

Kantorovitch, L. (1958), "On the translocation of masses," *Management Science*, 5, 1–4.

Klatt, M., Tameling, C., and Munk, A. (2018), "Empirical regularized optimal transport: Statistical theory and applications," *arXiv preprint arXiv:1810.09880*.

Le Cam, L. (1986), *Asymptotic methods in statistical decision theory*, Springer.

Lepski, O., Nemirovski, A., and Spokoiny, V. (1999), "On estimation of the L r norm of a regression function," *Probability theory and related fields*, 113, 221–253.

Liu, Q., Herring, C. A., Sheng, Q., Ping, J., Simmons, A. J., Chen, B., Banerjee, A., Li, W., Gu, G., Coffey, R. J., Yu, S., and Ken, L. S. (2018), "Quantitative assessment of cell population diversity in single-cell landscapes," *PLoS biology*, 16, e2006687.

Lozupone, C., Hamady, M., Kelley, S., and Knight, R. (2007), "Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities," *Applied and environmental microbiology*, 73, 1576–1585.

Lozupone, C., and Knight, R. (2005), "UniFrac: a new phylogenetic method for comparing microbial communities," *Applied and environmental microbiology*, 71, 8228–8235.

Mao, Q., Wang, L., Tsang, I. W., and Sun, Y. (2017), "Principal graph and structure learning based on reversed graph embedding," *IEEE transactions on pattern analysis and machine intelligence*, 39, 2227–2241.

Mirarab, S., Nguyen, N., and Warnow, T. (2012), "SEPP: SATé-enabled phylogenetic placement," in *Biocomputing 2012*, World Scientific, pp. 247–258.

Monge, G. (1781), "Mémoire sur la théorie des déblais et des remblais," *Histoire de l'Académie Royale des Sciences de Paris*.

Ni, K., Bresson, X., Chan, T., and Esedoglu, S. (2009), "Local histogram based segmentation using the Wasserstein distance," *International journal of computer vision*, 84, 97–111.

Olkin, I., and Sobel, M. (1979), "Admissible and minimax estimation for the multinomial distribution and for k independent binomial distributions," *The Annals of Statistics*, 284–290.

Orlova, D. Y., Zimmerman, N., Meehan, S., Meehan, C., Waters, J., Ghosn, E. E., Filatenkov, A., Kolyagin, G. A., Gernez, Y., Tsuda, S., Moore, W., Moss, R. B., Herzenberg, L. A., and Walther, G. (2016), "Earth mover?s distance (EMD): a true metric for comparing biomarker expression levels in cell populations," *PloS one*, 11, e0151859.

Paninski, L. (2003), "Estimation of entropy and mutual information," *Neural computation*, 15, 1191–1253.

Pavlichin, D. S., Jiao, J., and Weissman, T. (2017), "Approximate profile maximum likelihood," *arXiv preprint arXiv:1712.07177*.

Pavoine, S., Dufour, A., and Chessel, D. (2004), "From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis," *Journal of theoretical biology*, 228, 523–537.

Schilling, M. F. (1986), "Multivariate two-sample tests based on nearest neighbors," *Journal of the American Statistical Association*, 81, 799–806.

Singh, S., and Póczos, B. (2018), "Minimax distribution estimation in Wasserstein distance," *arXiv preprint arXiv:1802.08855*.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. (2015), "Convolutional wasserstein distances: Efficient optimal transportation on geometric domains," *ACM Transactions on Graphics (TOG)*, 34, 66.

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

Tameling, C., Sommerfeld, M., and Munk, A. (2017), "Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications," *arXiv preprint arXiv:1707.00973*.

Trybula, S. (1958), "Some problems of simultaneous minimax estimation," *The Annals of Mathematical Statistics*, 29, 245–253.

Valiant, G., and Valiant, P. (2011), "Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new CLTs," in *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 685–694.

Valiant, P., and Valiant, G. (2013), "Estimating the unseen: improved estimators for entropy and other properties," in *Advances in Neural Information Processing Systems*, pp. 2157–2165.

Villani, C. (2008), *Optimal transport: old and new*, vol. 338, Springer Science & Business Media.

Weed, J., and Bach, F. (2017), "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance," *arXiv preprint arXiv:1707.00087*.

Weiss, S., Xu, Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., and R., K. (2017), "Normalization and microbial differential abundance strategies depend upon data characteristics," *Microbiome*, 5, 27.

Wong, R. G., Wu, J. R., and Gloor, G. B. (2016), "Expanding the UniFrac toolbox," *PloS one*, 11, e0161196.

Wu, Y., and Yang, P. (2016), "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Transactions on Information Theory*, 62, 3702–3720.

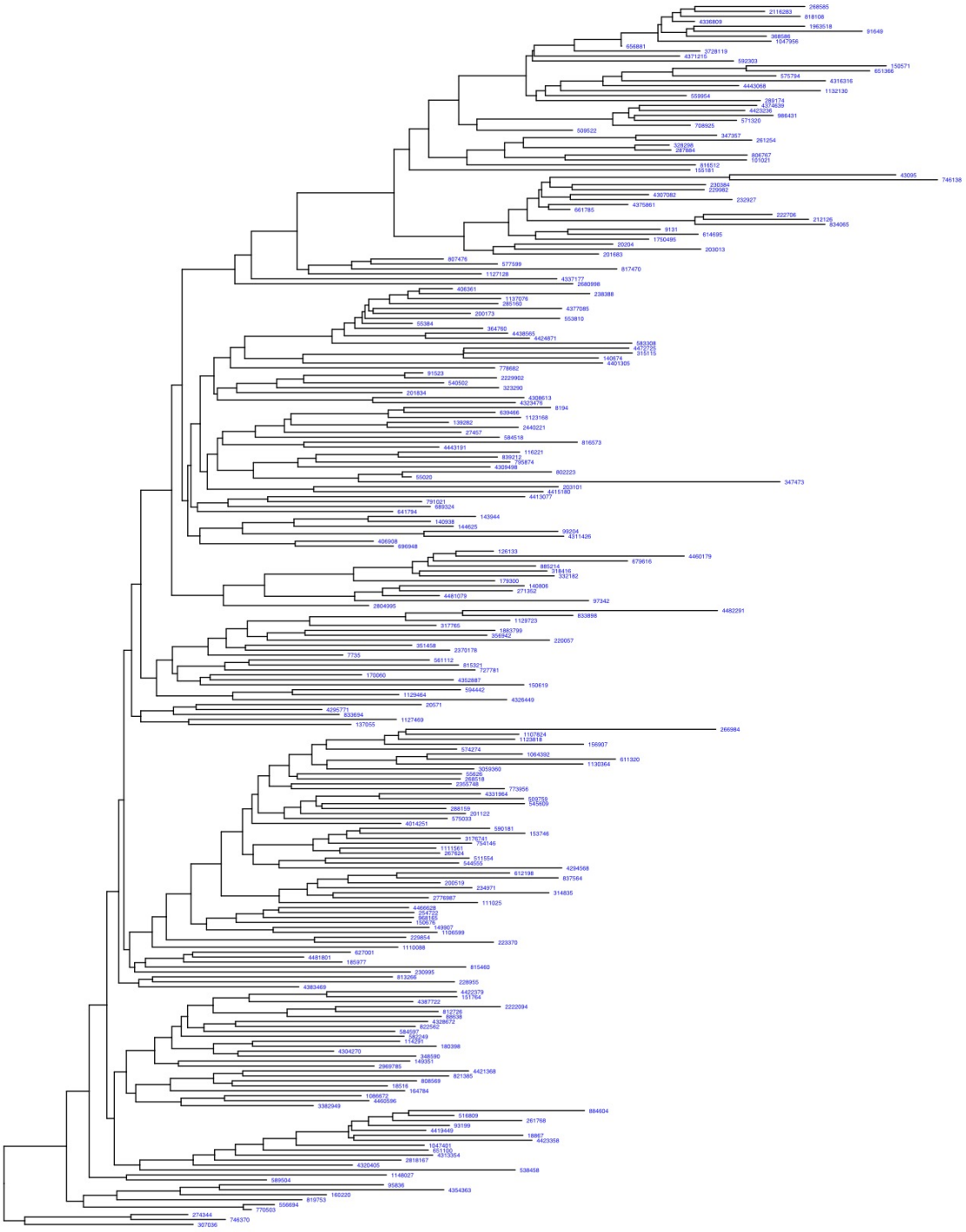


Figure 4: Phylogenetic tree of bacteria within the class *Gamma proteobacteria* used in simulation studies. There is a total of 247 leaves(tips) and 246 internal nodes. The leave number labels the bacterial species.

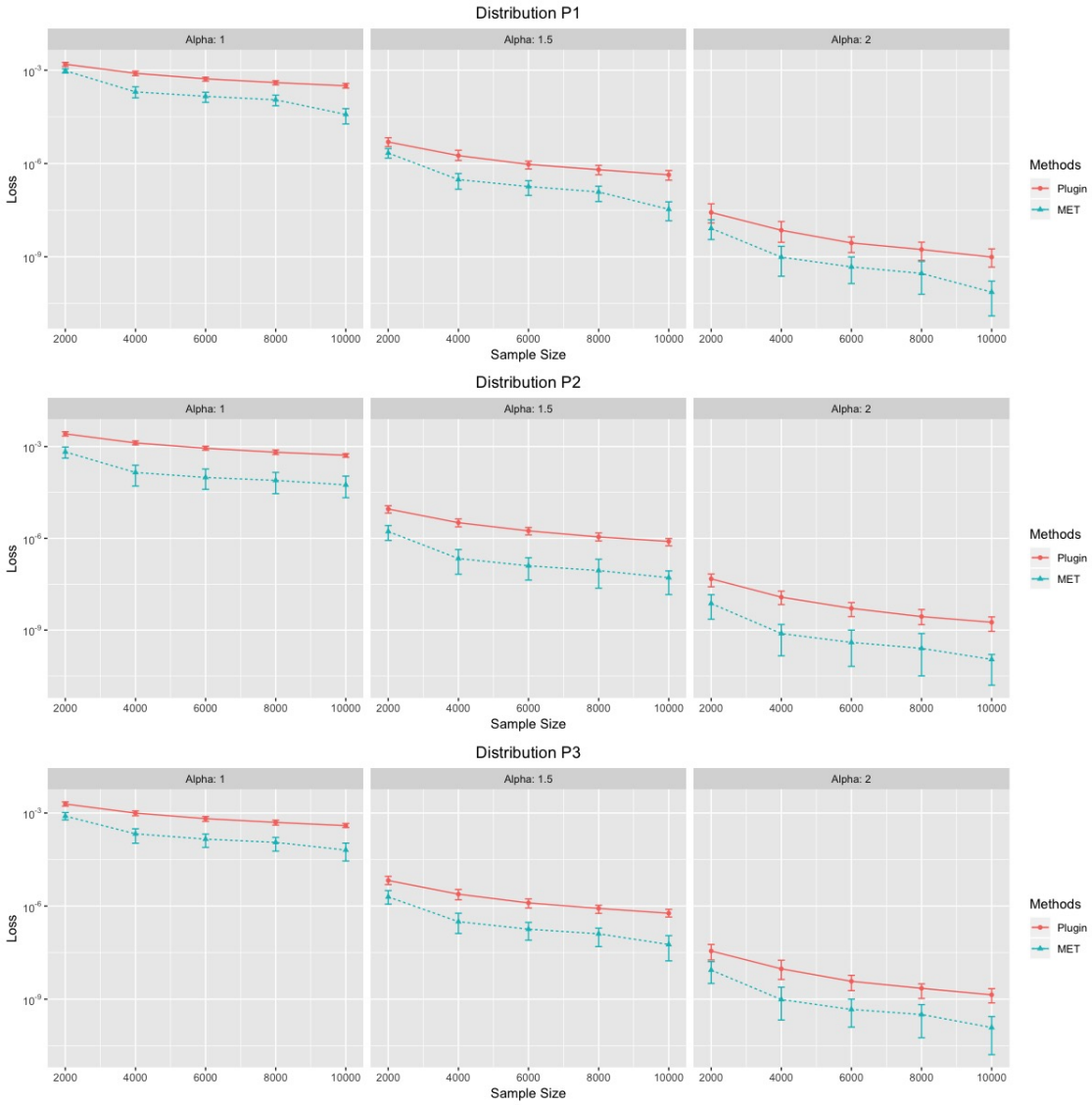


Figure 5: Comparison of the quadratic estimation losses between MET and the plug-in estimator when $D_\alpha(P, Q) = 0$ for three different read count distributions, $P1$, $P2$ and $P3$.

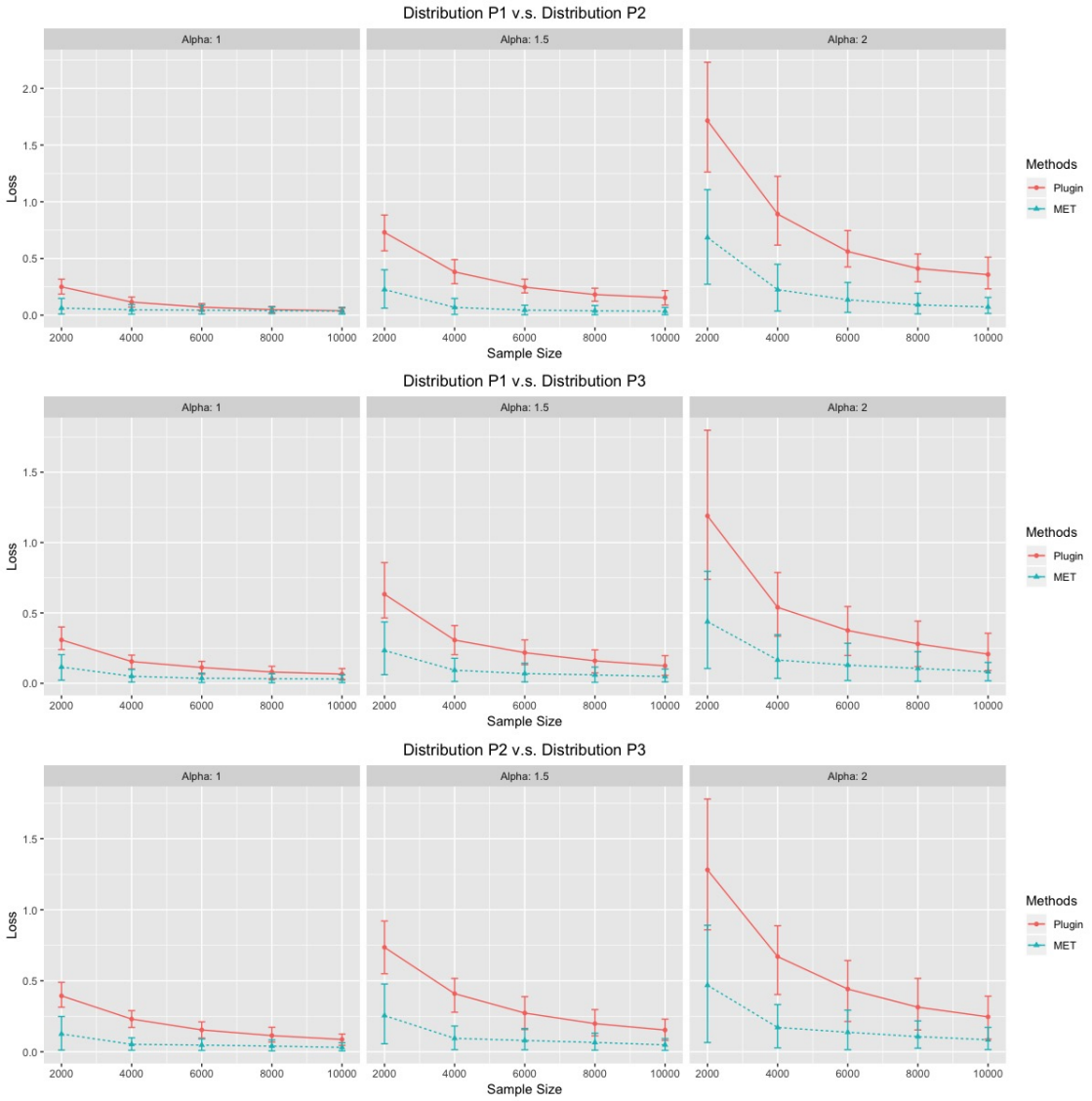


Figure 6: Comparison of the ratio of absolute error to the true distance between MET and the plug-in estimator when $D_\alpha(P, Q) \neq 0$ for three different P and Q distributions.

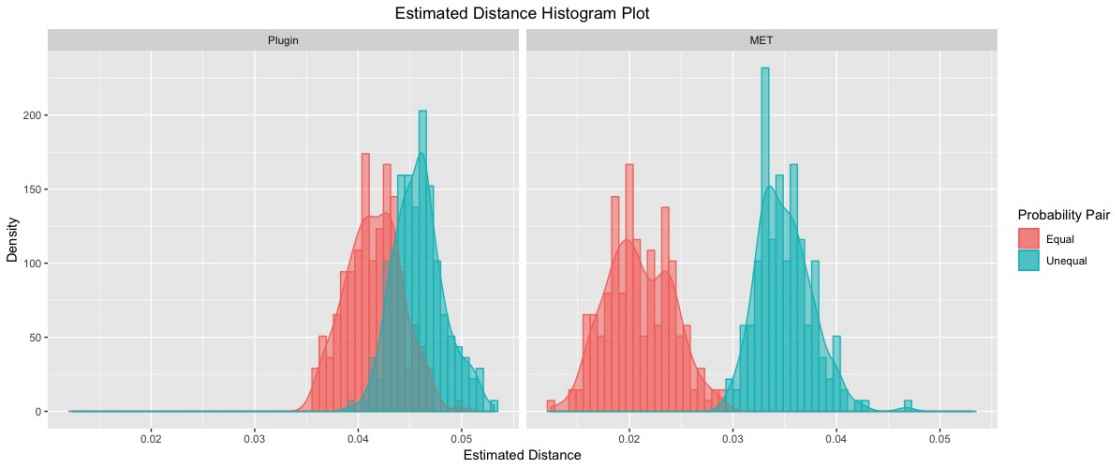


Figure 7: Histograms of the estimated distances by the plug-in estimator and MET when the true distance is 0.00 or 0.026.

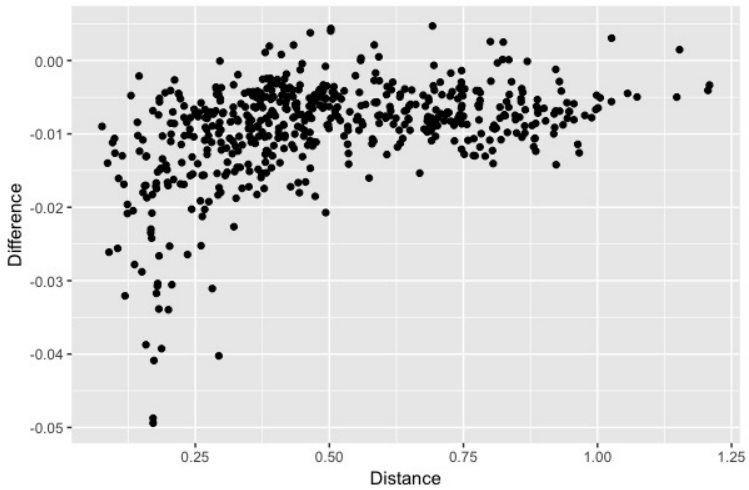
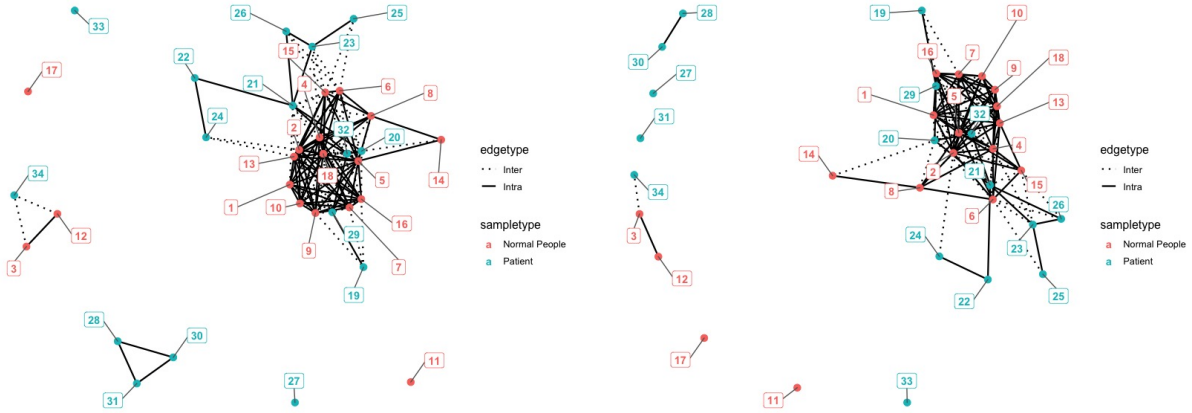


Figure 8: Comparison of the estimated distances of the Crohn's disease data sets. The difference between estimated distances by two methods, $\hat{D}_{\mathrm{MET}} - D(\hat{P}, \hat{Q})$, versus the estimated distance by MET, $\hat{D}_{\mathrm{MET}}$, are ploted for each pair of the samples.



(a) MET (b) Plugin

Figure 9: Estimated sample connectivity graphs by thresholding the distance matrix estimated by MET and the plug-in estimator for the inactive Crohn's disease samples and the control samples.