A The Perturbed Variation - Supplementary Material

A.1 Hypothesis Testing Procedures

The statistical tests in this section are based on the convergence bounds in Section 4.

Notations Throughout this section the probabilities $\mathbb{P}_0$ and $\mathbb{P}_1$ represent the probability conditioned on the null hypothesis $\mathcal{H}_0$, and the alternative hypothesis $\mathcal{H}_1$.

The following procedure tests the hypothesis $\mathcal{H}_0^{(1)}$: $\text{PV}(P, Q, \epsilon) \leq \theta$ against the alternative $\mathcal{H}_1^{(1)}$: $\text{PV}(P, Q, \epsilon) > \theta$.

Procedure 1. *Similarity Testing Based on $\widehat{PV}$.*
Input*: ϵ, θ and significance level α.*

1. *Sample $S_1 = \{x_1, ..., x_n\} \sim P$ and $S_2 = \{y_1, ..., y_m\} \sim Q$ (define $N = \min(n, m)$).*
2. *Normalize the data to be in $[0, 1]^d$.*
3. *Compute $\widehat{PV}(S_1, S_2, \epsilon, \|\cdot\|_\infty)$ by Algorithm 1.*
4. *Compute $t = \sqrt{\frac{(2\log(2(2^{(1/\epsilon)^d} - 2)) + 2\log(1/\alpha)}{N}}$.*

Output*: Reject $\mathcal{H}_0^{(1)}$ if*

$$\widehat{PV}(S_1, S_2, \epsilon, \|\cdot\|_\infty) > t + \theta$$

.

The probability to reject $\mathcal{H}_0^{(1)}$ by applying Procedure 1 when in fact it holds – also known as the Type 1 error – is bounded in the following corollary.

Corollary 6. *Assume that for a given ϵ and θ values $\mathcal{H}_0^{(1)}$: $PV(P, Q, \epsilon, \mathbf{d}) \leq \theta$ holds. Then for the threshold t of Procedure 1 and any $\alpha \in (0, 1)$ we have that*

$$\mathbb{P}_0\left(\widehat{PV}(S_1, S_2, \epsilon, \|\cdot\|_\infty) \geq t + \theta\right) \leq \alpha. \tag{5}$$

Moreover, the procedure is consistent: when $n, m \to \infty$ we have that $t \to 0$ and $\mathbb{P}_1(\widehat{PV}(S_1, S_2, \epsilon, \|\cdot\|_\infty) > \theta) = 1$.

The corollary is a direct result of Theorem 3.

Next, we consider the probability that Procedure 1 fails to reject $\mathcal{H}_0^{(1)}$ when the alternative hypothesis $\mathcal{H}_1^{(1)}$ holds, also known as the Type 2 error. Unfortunately, it is not possible to bound this probability for a finite sample of *any* two distributions. To see this, consider the following example: let P, Q be two distributions with $PV(P, Q, \epsilon) > 0$, but differ only in an area of very low probability. Then, for any finite sample size, there is a high probability that the samples are identical, resulting in $\widehat{PV}(S_1, S_2, \epsilon) = 0$. As a result, the null hypothesis will not be rejected even though $\mathcal{H}_1^{(1)}$ holds.

However, if the PV is larger than some constant the Type 2 error is bounded.

Corollary 7. *For $PV(P, Q, \epsilon, \mathbf{d}) > \theta + t + b$, with t of Procedure 1, and $b = \sqrt{\frac{2(\log(2(2^{(1/\epsilon)^d} - 2)) + 2\log(1/\beta)}{N}}$ we have that*

$$\mathbb{P}\left(\widehat{PV}(S_1, S_2, \epsilon, \|\cdot\|_\infty) > t + \theta\right) \geq 1 - \beta.$$

Note that as N grows, the values of b and t get smaller, and the lower bound $PV(P, Q, \epsilon, \mathbf{d}) > \theta + t + b$ decreases.

Proof. We have that

$$\mathbb{P}_1\left(\widehat{PV}(S_1, S_2, \epsilon, \|\cdot\|_\infty) > t + \theta\right) = \mathbb{P}_1\left(\widehat{PV}(S_1, S_2, \epsilon, \|\cdot\|_\infty) > b + t + \theta - b\right) \geq$$
$$\mathbb{P}\left(\widehat{PV}(S_1, S_2, \epsilon, \|\cdot\|_\infty) > PV(P, Q, \epsilon, \|\cdot\|_\infty) - b\right) \geq 1 - \beta.$$

The first inequality holds by inserting the assumption on PV, and the second holds by applying the convergence bound of Theorem 3. □

To give an estimate of the sample size needed for the procedure, first define the effect size θ_0: the minimal value of PV that is significant. Given θ_0, set the sample size so that

$$N \geq \frac{4\log(2(2^{(1/\epsilon)^d} - 2)) + 2\log(1/\alpha) + 2\log(1/\beta)}{\theta_0^2}.$$

Using this size ensures a false positive rate bounded by α (Corollary 6), and a false negative rate bounded by β (Corollary 7).

The second test we consider is an equivalence type test [11]. Equivalence is achieved when $\text{PV}(P, Q, \epsilon) < \theta$, for some chosen θ, and may be obtained by switching the roles of the null and the alternative of Procedure 1. Namely, to claim similarity we need to reject $\mathcal{H}_0^{(2)} : \text{PV}(P, Q, \epsilon) \geq \theta$. To test this hypothesis, a similar procedure to Procedure 1 may be applied, with a principal difference in the rejection area, which is changed to $\widehat{PV}(S_1, S_2, \epsilon, \|\cdot\|_\infty) < \theta - t$.

A.2 1D Projections

We present a method to gain insight on the value of the PV by multiple random projections to one dimension. While the PV between two distributions is not retained after projection to a single dimension, as the projection is a non-expansive mapping of the samples, we show that multiple projections can still aid to distinguish between two situations: $\text{PV}(P, Q, \epsilon) = 0$ and $\text{PV}(P, Q, \epsilon) \neq 0$ [4].

We define a score that is based on the value of the PV after projections.

Definition 8. *Let $f_i : \mathbb{R}^d \to \mathbb{R}$ for $i = 1, ..., K$ define a set of random projection mappings, and let X and Y be random variables with distributions P and Q. The* projected perturbed variation *of two distributions P and Q is*

$$PPV(P, Q, \epsilon, K) = \max_{i=1,...,K} PV(f_i(X), f_i(Y), \epsilon).$$

For K i.i.d. samples $S_{i1} = \{x_{i1}, ..., x_{in}\} \sim P$ and $S_{i2} = \{y_{i1}, ..., y_{im}\} \sim Q$ the score is

$$\widehat{PPV}(S_1, S_2, \epsilon, K) = \max_{i=1,...,K} \widehat{PV}(f_i(S_{i1}), f_i(S_{i2}), \epsilon),$$

where S_1, S_2 denote the K samples.

We denote $\widehat{\text{PPV}}_i(\epsilon) = \widehat{\text{PV}}(f_i(S_{i1}), f_i(S_{i2}), \epsilon)$ and $\text{PPV}_i(\epsilon) = PV(f_i(X), f_i(Y), \epsilon)$ as the value of the sampled and distributional perturbed variation after the ith projection. The next theorem presents the convergence rate of $\widehat{\text{PPV}}(S_1, S_2, \epsilon, K)$ under the assumption that $\text{PV}(P, Q, \epsilon) = 0$. Under this assumption, the projected PV is also zero, and $\widehat{\text{PPV}}(S_1, S_2, \epsilon, K)$ converges to $\text{PV}(P, Q, \epsilon)$.

Theorem 9. *Let P and Q be two distributions on the space $([0, 1]^d, \mathbf{d})$, and $S_1 = \{x_1, ..., x_n\} \sim P$ and $S_2 = \{y_1, ..., y_m\} \sim Q$ two i.i.d. samples ($N = \min(n, m)$). Perform K i.i.d. random projections of samples S_1 and S_2 to one dimension. If $PV(P, Q, \epsilon) = 0$, then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$*

$$\widehat{PPV}(S_1, S_2, \epsilon, K) \leq \sqrt{\frac{2\log(2K(2^{1/\epsilon} - 2)/\delta)}{N}}.$$

In the following derivations we drop the sample specification of S_1, S_2 for brevity, and let $\mathbb{P}_0(a)$ denote the probability of event a under the assumption $\text{PV}(P, Q, \epsilon) = 0$.

Proof. Given $\text{PV}(P, Q, \epsilon) = 0$, we have that for all K projections $\text{PPV}_i(\epsilon) = 0$, as the projection to 1D is a non-expansion. To bound the probability of the event $\widehat{\text{PPV}}(S_1, S_2, \epsilon, K) \geq \eta$ we apply the

[4] Recall that PV=0 not only when the distributions are equal, but also when they are ϵ similar.

union bound, and then apply Theorem 3 on each of the K projections $\widehat{\text{PPV}}_i(\epsilon)$ for $i = 1, ..., K$.

$$\begin{aligned}\mathbb{P}_0\left(\widehat{\text{PPV}}(S_1, S_2, \epsilon, K) \geq \eta\right) &= \mathbb{P}_0\left(\max_{1\leq i\leq K} \widehat{\text{PPV}}_i(\epsilon) \geq \eta\right) = \mathbb{P}_0\left(\exists 1 \leq i \leq K : \widehat{\text{PPV}}_i(\epsilon) \geq \eta\right) \\ &\leq \sum_{i=1}^{K} \mathbb{P}_0\left(\widehat{\text{PPV}}_i(\epsilon) \geq \eta\right) \leq K \max_{1\leq i\leq K} \mathbb{P}_0\left(\widehat{\text{PPV}}_i(\epsilon) - \text{PPV}_i(\epsilon) \geq \eta\right) \\ &\leq 2K(2^{1/\epsilon} - 2)e^{-N\eta^2/2}.\end{aligned}$$

Setting $\delta = 2K(2^{1/\epsilon} - 2)e^{-N\eta^2/2}$ concludes the proof. □

For $\text{PV}(P, Q) > 0$, we provide a similar lower bound on the projected perturbed variation. We will need a further assumption for this bound.

Definition 10. *Given distributions P and Q with PV$(P, Q, \epsilon) > 0$, they are* 1D distinguishable *if* $\lim_{K\to\infty} PPV_K(P, Q, \epsilon) > 0$ *almost surely.*

This assumption of 1D distinguishability ensures that the difference in the PV value exists in at least one projection.

Theorem 11. *Let P and Q be two distributions on the space $([0, 1]^d, \mathbf{d})$ that are 1D distinguishable. Given $i = 1, ..., K$ i.i.d. samples $S_{i1} = \{x_{i1}, ..., x_{in}\} \sim P$ and $S_{i2} = \{y_{i1}, ..., y_{im}\} \sim Q$, and K mappings f_i, there exists some $q \in (0, 1)$, for which for any $\delta \in (0, 1)$ with probability at least $1 - (q - q\delta + \delta)^K$*

$$\widehat{PPV}(S_1, S_2, \epsilon) \geq \sqrt{\frac{2\log(2K(2^{1/\epsilon} - 2)/\delta)}{N}}.$$

Notice that $q - q\delta + \delta < 1$, and therefore there is an exponential decay in the number of projections K.

Proof. Let $f : \mathbb{R}^d \to \mathbb{R}$ define a random projection mappings, and let X and Y be random variables generated by P and Q. Note that there are two sources of randomization, the sample's and the projection's, and therefore PPV_i is also a random variable. The samples are independent and therefore

$$\mathbb{P}(\widehat{\text{PPV}}(S_1, S_2, \epsilon, K) \leq \eta) = \mathbb{P}(\forall\, 1 \leq i \leq K\,, \widehat{\text{PPV}}_i(\epsilon) \leq \eta) = \prod_{i=1}^{K} \mathbb{P}(\widehat{\text{PPV}}_i(\epsilon) \leq \eta). \tag{6}$$

For each of the projections $i = 1, ..., K$, we define two complementary events
$a : PPV_i(\epsilon) \geq 2\eta$ and $a^c : PPV_i(\epsilon) < 2\eta$.

$$\begin{aligned}&\mathbb{P}(\widehat{\text{PPV}}_i(\epsilon) \leq \eta) = \mathbb{P}(a)\mathbb{P}(\widehat{\text{PPV}}_i(\epsilon) \leq \eta \,|\, a) + \mathbb{P}(a^c)\mathbb{P}(\widehat{\text{PPV}}_i(\epsilon) \leq \eta \,|\, a^c) \\ &\leq \mathbb{P}(a)\mathbb{P}\left(\widehat{\text{PPV}}_i(\epsilon) \leq \text{PPV}_i - \eta \,|\, a\right) + \mathbb{P}(a^c)\mathbb{P}\left(\widehat{\text{PPV}}_i(\epsilon) \leq \eta \,|\, a^c\right) \\ &\overset{(*)}{\leq} \mathbb{P}(a)2(2^{1/\epsilon} - 2)e^{-N\eta^2/2} + \mathbb{P}(a^c) \leq \mathbb{P}(a)2K(2^{1/\epsilon} - 2)e^{-N\eta^2/2} + 1 - \mathbb{P}(a).\end{aligned} \tag{7}$$

Inequality $^{(*)}$ is obtained by applying Theorem 3 for any $\eta \in (0, 1)$.
Setting $\delta \doteq 2K(2^{1/\epsilon} - 2)e^{-N\tilde{\eta}^2/2}$ yields $\tilde{\eta} \doteq \sqrt{\frac{2\log(2K(2^{1/\epsilon}-2)\delta)}{N}}$. Substituting $\tilde{\eta}$ into (7) yields

$$\mathbb{P}(\widehat{\text{PPV}}_i(\epsilon) \leq \tilde{\eta}) \leq 1 - (1 - \delta)\mathbb{P}(\text{PPV}_i(\epsilon) \geq 2\tilde{\eta}). \tag{8}$$

The probability $\mathbb{P}(\text{PPV}_i(P, Q, \epsilon) \leq 2\tilde{\eta})$ depends on the generating distributions P and Q. Its support is $[0, \sup_i(\text{PPV}_i(P, Q, \epsilon))]$. We assume that $\sup_i(\text{PPV}_i(P, Q, \epsilon)) > 0$, and therefore there must be some $q \in (0, 1)$ for which for all $i = 1, ..., K$

$$\mathbb{P}(\text{PPV}_i(P, Q, \epsilon) < 2\tilde{\eta}) \leq q. \tag{9}$$

Combining the results of Equations (6)-(9), we have that for any $0 < \delta < 1$

$$\mathbb{P}(\widehat{\text{PPV}}(S_1, S_2, \epsilon) \leq \tilde{\eta}) \leq \prod_{i=1}^{K} (1 - (1 - \delta)\mathbb{P}(\text{PPV}_i(\epsilon) \geq 2\tilde{\eta})) \leq (q - q\delta + \delta)^K,$$

which concludes the proof. Note that $q - q\alpha + \alpha < 1$ always holds, and therefore we get exponential decay as the number of projections K grows. For example, if $q = 1/2$, in which case $2\tilde{\eta}$ is smaller then the median, we have $(q - q\delta + \delta)^K = \left(\frac{1+\delta}{2}\right)^K$. □

Theorems 9 and 11 are complementary, and may be used together to infer whether or not $\mathrm{PV}(P, Q) = 0$. Next, we describe the suitable hypothesis testing procedure for this goal. Procedure 2 provides statistical tests based on the score $\widehat{\mathrm{PPV}}$ (Definition 8). The procedure tests an hypothesis of the first type with $\theta = 0$: $\mathcal{H}_0^{(1)} : PV(P, Q, \epsilon) = 0$ against the alternative $\mathcal{H}_1^{(1)} : PV(P, Q, \epsilon) > 0$.

Procedure 2. *Similarity testing based on $\widehat{PPV}$.*
Input*: ϵ level, number of projections K, and significance level α.*
For $i = 1, ..., K$ **do**

1. *Sample $S_{i1} = \{x_1, ..., x_n\} \sim P$ and $S_{i2} = \{y_1, ..., y_m\} \sim Q$ i.i.d. examples on $[0, 1]^d$.*
2. *Sample a unit random vector $r_i \in \mathbb{S}^{d-1}$.*
3. *Project to 1D: $s_{i1} = \{r_i^T x_1, ..., r_i^T x_n\}$ and $s_{i2} = \{r_i^T y_1, ..., r_i^T y_m\}$.*
4. *Compute $\widehat{PV}(s_{i1}, s_{i2}, \epsilon)$.*

end for
Compute $\widehat{PPV}(\epsilon, K) = \max_{i=1,...,K} \widehat{PV}(s_{i1}, s_{i2}, \epsilon)$.
Compute $t = \sqrt{\frac{\log(K) + 2\log(2(2^{1/\epsilon} - 2)) + 2\log(1/\alpha)}{N}}$, where $N = \min(n, m)$.
Output*: Reject $\mathcal{H}_0$ if $\widehat{PPV}(\epsilon, K) > t$.*

This procedure is more limited than Procedure 1 as it holds only for $\theta = 0$. However, it may provide better results for high dimensional distributions. Theorems 9 and 11 bound the Type 1 error and Type 2 error of Procedure 2 respectively. The Type 2 error is dependent on the number of projections K, and the fraction q that is distribution dependent. The bound exponentially decays as K grows, and therefore, to gain statistical power, a larger number of projections can be used.

A.3 Proof of Theorem 3

We restate the theorem for clarity:

Theorem 3. *Suppose we are given two i.i.d. samples $S_1 = \{x_1, ..., x_n\} \in \mathbb{R}^d$ and $S_2 = \{y_1, ..., y_m\} \in \mathbb{R}^d$ generated by distributions P and Q, respectively. Let the ground distance be $\mathbf{d} = \|\cdot\|_\infty$ and let $\mathcal{N}(\epsilon)$ be the cardinality of a disjoint cover of the distributions' support. Then, for any $\delta \in (0,1)$, $N = \min(n,m)$, and $\eta = \sqrt{\frac{2(\log(2(2^{\mathcal{N}(\epsilon)}-2))+\log(1/\delta))}{N}}$ we have that*

$$\mathbb{P}\left(\left|\widehat{PV}(S_1, S_2, \epsilon) - PV(P, Q, \epsilon)\right| \leq \eta\right) \geq 1 - \delta.$$

The proof of the theorem is carried out in two steps. First, the relations between the continuous and discrete versions of the PV is formulated. Then, turning to the discrete versions, we bound the difference between the PV (Problem (2)) of the discretized samples (i.e. the histograms defined on the discretization) and the discretized distributions. This part of the proof exploits the special form of the optimization in Problem 2.

Before providing the proof we present the required definitions and lemmas. To aid the reading flow the proofs of some lammas are presented immediately after the proof of the main theorem. We assume the domain is totally bounded, and, for simplicity of presentation, we assume the metric space is $([0,1]^d, \mathbf{d}_\infty = \|\cdot\|_\infty)$.

We define a discretization on the support of the distributions.

Definition 12 (Discretization)**.** *The* ϵ-discretization *over the space $([0,1]^d, \mathbf{d}_\infty = \|\cdot\|_\infty)$ is a partition on the set $C(\epsilon) = \{a_1, ..., a_N\}$, with cardinality $N = (1/\epsilon)^d$, which covers $[0,1]^d$. Each element in $a_i \in C(\epsilon)$ is the center of a box of volume ϵ^d, with density equal to the distribution's mass in its neighborhood: $B(a_i, \mathbf{d}_\infty, \epsilon) = \{z : \mathbf{d}_\infty(a_i, z) \leq \epsilon/2\}$.*

We use the following structure of two discretizations:

Definition 13 (Refinement of a discretization)**.** *Define an initial ϵ-discretization $C_1(\epsilon) = \{b_1, ..., b_{N(\epsilon)}\}$ on $([0,1]^d, \|\cdot\|_\infty)$. The* refined discretization*, for any ϵ and $T > 1$, is defined as a ν-discretization on $C_2(\nu) = \{a_1, ..., a_{N(\nu)}\}$, where $\nu = \epsilon/T$, such that each element of the refinement is a result of equally splitting an element of the initial cover to $(\epsilon/T)^d$.*

We refer to the discretized versions of the distributions P and Q as $\mu_1(\epsilon)$, $\mu_2(\epsilon)$ respectively, where ϵ is the size of the partition. Also, we refer to the histograms of the samples S_1 and S_2 defined on the same discretization as $\hat{\mu}_1(\epsilon)$, $\hat{\mu}_2(\epsilon)$.

The relation between the different versions of the PV, continuous, discrete and sampled, is provided in the next lemma.

Lemma 14. *Let $S_1 = \{x_1, ..., x_n\} \sim P$ and $S_2 = \{y_1, ..., y_m\} \sim Q$ be two samples. Let $\mu_1(\nu)$ and $\mu_2(\nu)$ be the ν-discretizations of P and Q for any integer $T > 1$ and $\nu = \frac{\epsilon}{T}$. Let $\hat{\mu}_1(\nu)$ and $\hat{\mu}_2(\nu)$ be their empirical distributions. The following relations hold for any ϵ, $\epsilon' = \frac{\epsilon(T-1)}{T}, \epsilon'' = \frac{\epsilon(T+1)}{T}$ and $\mathbf{d} = \|\cdot\|_\infty$:*

$$PV(\hat{\mu}_1, \hat{\mu}_2, \epsilon'') \leq \widehat{PV}(S_1, S_2, \epsilon) \leq PV(\hat{\mu}_1, \hat{\mu}_2, \epsilon') \tag{10}$$

$$PV(\mu_1, \mu_2, \epsilon'') \leq PV(P, Q, \epsilon) \leq PV(\mu_1, \mu_2, \epsilon'). \tag{11}$$

The following representation of Problem (2) will be useful for our derivations.

Lemma 15. *The solution of Problem (2) may be obtained by solving the following problem*

$$\begin{aligned}
&\min_{w_i, v_i, Z_{ij}} \frac{1}{2}\sum_{i=1}^{N}|w_i| + \frac{1}{2}\sum_{j=1}^{N}|v_j| \\
&\sum_{a_j \in ng(a_i,\epsilon)} Z_{ij} + w_i = \mu_1(a_i),\ i = 1, ..., N \\
&\sum_{a_i \in ng(a_j,\epsilon)} Z_{ij} + v_j = \mu_2(a_j),\ j = 1, ..., N \\
&Z_{ij} \geq 0,\ \forall i, j,
\end{aligned} \tag{12}$$

which we call PV$_{eq}(\mu_1(\nu), \mu_2(\nu), \epsilon)$.

The lemma states that the constraints $w_i \geq 0,\ v_j \geq 0$ may be removed, and instead the sum in the objective is taken over the absolute values.

The next lemma bounds the difference between the PV of the distributions $\hat{\mu}_1(\nu), \hat{\mu}_2(\nu)$ and the distributions $\mu_1(\nu)$ and $\mu_2(\nu)$.

Lemma 16. *Let $C_1(\epsilon)$ be an ϵ-discretization on $[0,1]^d$, and $C_2(\nu)$ its refined discretization (Definition 13). Let $\hat{\mu}_i(\epsilon)$ and $\mu_i(\epsilon)$ be distributions on $C_1(\epsilon)$, and $\hat{\mu}_i(\nu)$ and $\mu_i(\nu)$ distributions on the refinement $C_2(\nu)$. For any $\epsilon \in (0,1)$ and $\mathbf{d} = \|\cdot\|_\infty$ we have that*

$$|PV(\hat{\mu}_1(\nu), \hat{\mu}_2(\nu), \epsilon) - PV(\mu_1(\nu), \mu_2(\nu), \epsilon)| \leq \frac{1}{2}\left(\|\mu_1(\epsilon) - \hat{\mu}_1(\epsilon)\|_1 + \|\mu_2(\epsilon) - \hat{\mu}_2(\epsilon)\|_1\right).$$

Observe that the L_1-norm is computed over the elements of $C_1(\epsilon)$, the original discretization rather than the refinement.

Proof. We bound the difference between PV(μ_1, μ_2, ϵ) and PV$_{eq}(\hat{\mu}_1, \hat{\mu}_2, \epsilon)$ instead of the difference between PV(μ_1, μ_2, ϵ) and PV$(\hat{\mu}_1, \hat{\mu}_2, \epsilon)$, as by Lemma 15 the two are equivalent. To bound this difference we start at the optimal solution for distributions μ_1 and μ_2 and make the needed changes to obtain a feasible solution for distributions $\hat{\mu}_1$ and $\hat{\mu}_2$. This solution may be suboptimal and therefore upper bounds the value of PV$_{eq}(\hat{\mu}_1, \hat{\mu}_2, \epsilon)$.

Let $opt(\mu_1, \mu_2) = \{Z^*_{ij}, w^*_i, v^*_j \text{ for } i, j = 1, ..., N\}$ be the optimal arguments of Problem (2) for distributions μ_1 and μ_2; namely,

$$\text{PV}(\mu_1, \mu_2, \epsilon) = \frac{1}{2}\sum_{i=1}^{N} w^*_i + \frac{1}{2}\sum_{j=1}^{N} v^*_j.$$

We substitute the variables $opt(\mu_1, \mu_2)$ into Problem (12) for distributions $\hat{\mu}_1, \hat{\mu}_2$. To transform this solution to a feasible solution we must fix the violations that are made to the constraints. The constraints are fixed in two manners. Some are fixed by optimizing the transportation plan, described by matrix Z, within the refinement of the discretization. Additional violations are fixed by changing the variables w_j and v_j.

We consider the first type of constraint violations. Define $s_k = \{a_i : a_i \in B(b_k, \|\cdot\|_\infty, \epsilon)\}$; i.e., the set of bins $a_i \in C_2(\nu)$ that are a refinement of element $b_k \in C_1(\epsilon)$ (Definition 13). Let $|s_k|$ be the cardinality of this set. By definition, all the bins in s_k are ϵ-neighbors: $\forall a_i \in s_k,\ s_k \in \text{ng}(a_i, \epsilon)$. For any $a_i, a_j \in s_k$, consider the following feasibility problem:

$$\begin{aligned} &\text{Find } C_{ij} \\ &s.t. \sum_{a_j \in s_k} C_{ij} = c_i, \quad \forall a_i \in s_k, \\ &\sum_{a_i \in s_k} C_{ij} = b_j, \quad \forall a_j \in s_k, \\ &Z^*_{ij} + C_{ij} \geq 0, \quad \forall a_i, a_j \in s_k, \end{aligned} \tag{13}$$

where

$$\begin{aligned} c_i &\doteq (\hat{\mu}_1(a_i) - \mu_1(a_i)) - \frac{1}{|s_k|}(\hat{\mu}_1(b_k) - \mu_1(b_k)), \\ b_j &\doteq (\hat{\mu}_2(a_j) - \mu_2(a_j)) - \frac{1}{|s_k|}(\hat{\mu}_2(b_k) - \mu_2(b_k)). \end{aligned}$$

Note that c_i and b_i may be positive or negative, and that $\sum_{a_i \in s_k} c_i = 0$ and $\sum_{a_j \in s_k} b_j = 0$.

In the following, we show that Problem (13) is indeed feasible. First, we rewrite the problem in vector form. Define $v = \text{Vec}(\{C_{ij}\}_{a_i, a_j \in s_k}) \in \mathbb{R}^{|s_k|^2 \times 1}$, the vector form of the sub-matrix $\{C_{ij}\}_{a_i, a_j \in s_k}$. Similarly, let $z^* = \text{Vec}(\{Z^*_{ij}\}_{a_i, a_j \in s_k}) \in \mathbb{R}^{|s_k|^2 \times 1}$. Let $A \in \mathbb{R}^{2|s_k| \times |s_k|^2}$ be the zero-one matrix defined by the left-hand sides of the equality constraints in (13), and

$d = [c_1, ..., c_{|s_k|}, b_1, ..., b_{|s_k|}]^T \in \mathbb{R}^{2|s_k|\times 1}$, the vector defined by the right-hand sides of these constraints. Using these notations, Problem (13) is equivalent to

$$\begin{aligned} &\text{Find } v \\ &Av = d\,, \quad -v - z^* \leq 0, \end{aligned}$$

Consider its dual representation: the existence of $\lambda \in \mathbb{R}_+^{|s_k|^2\times 1}, \eta \in \mathbb{R}^{2|s_k|\times 1}$ for which

$$g(\lambda, \eta) = \inf_v \lambda^T(-v - z^*) + \eta^T(Av - d) > 0. \tag{14}$$

The value of $g(\lambda, \eta)$ in (14) is not $-\infty$ only when $A^T\eta - \lambda = 0$, in which case

$$g(\lambda, \eta) = \inf_v v^T(-\lambda + A^T\eta) - \lambda^T z^* - \eta^T d = -\lambda^T z^* - \eta^T d.$$

Since $z^* \geq 0$ and $\lambda \geq 0$, we have that $-\lambda^T z^* \leq 0$. Also, since $\mathbf{1}^T d = \sum_{a_i\in s_k} c_i + \sum_{a_j\in s_k} b_j = 0$, we have that $-\eta^T d \leq -\min \eta_\ell \cdot \mathbf{1}^T d = 0$. We conclude that $g(\lambda, \eta) \leq 0$, and therefore Problem (14) is infeasible. By the theorem of alternatives Problem (13) is feasible [14].

We claim that the following values $feas(\hat{\mu}_1, \hat{\mu}_2) = \{\bar{w}_i, \bar{v}_j, \bar{Z}_{ij} \text{ for } i, j = 1, ..., N(\nu)\}$ are a feasible solution to Problem (12) for distributions $\hat{\mu}_1, \hat{\mu}_2$:

$$\begin{aligned} \bar{w}_i &= w_i^* + \frac{1}{|s_k|}(\hat{\mu}_1(b_k) - \mu_1(b_k)) \\ \bar{v}_j &= v_j^* + \frac{1}{|s_k|}(\hat{\mu}_2(b_k) - \mu_2(b_k)) \\ \bar{Z}_{ij} &= \begin{cases} Z_{ij}^* & \text{if } a_j \in s_k^c,\ a_i \in s_k, \\ Z_{ij}^* + C_{ij} & \text{if } a_j \in s_k,\ a_i \in s_k, \end{cases} \end{aligned} \tag{15}$$

where C_{ij} is the solution to the (13).

First note that the constraints $\bar{Z}_{ij} \geq 0$ hold by the feasibility of (13). The equality constraints also hold, since

$$\begin{aligned} &\sum_{a_j\in \text{ng}(a_i,\epsilon)} \bar{Z}_{ij} + \bar{w}_i = \sum_{a_j\in \text{ng}(a_i,\epsilon)} Z_{ij}^* + \sum_{a_j\in s_k} C_{ij} + \bar{w}_i = \sum_{a_j\in \text{ng}(a_i,\epsilon)} Z_{ij}^* + c_i + \bar{w}_i = \\ &\sum_{a_j\in \text{ng}(a_i,\epsilon)} Z_{ij}^* + \hat{\mu}_1(a_i) - \mu_1(a_i) - \frac{1}{|s_k|}(\hat{\mu}_1(b_k) - \mu_1(b_k)) + w_i^* + \frac{1}{|s_k|}(\hat{\mu}_1(b_k) - \mu_1(b_k)) \\ &= \mu_1(a_i) + (\hat{\mu}_1(a_i) - \mu_1(a_i)) = \hat{\mu}_1(a_i), \end{aligned}$$

and in a similar manner $\sum_{a_i\in \text{ng}(a_j,\epsilon)} \bar{Z}_{ij} + \bar{v}_j = \hat{\mu}_2(a_j)$.

To conclude the proof, we bound the difference of the objective of Problem (2), obtained with the values $opt(\mu_1, \mu_2)$, and the objective of Problem (12), obtained with the values $feas(\hat{\mu}_1, \hat{\mu}_2)$. We have that

$$\begin{aligned} &PV(\hat{\mu}_1(\nu), \hat{\mu}_2(\nu), \epsilon) - PV(\mu_1(\nu), \mu_2(\nu), \epsilon) = \\ &= PV_{eq}(\hat{\mu}_1(\nu), \hat{\mu}_2(\nu), \epsilon) - PV(\mu_1(\nu), \mu_2(\nu), \epsilon) \overset{(a)}{\leq} \frac{1}{2}\sum_{i=1}^{N(\nu)} (|\bar{w}_i| + |\bar{v}_i|) - \frac{1}{2}\sum_{i=1}^{N(\nu)} (w_i^* + v_i^*) \\ &= \frac{1}{2}\sum_{k=1}^{N(\epsilon)} \sum_{a_i\in s_k} (|\bar{w}_i| + |\bar{v}_i|) - \frac{1}{2}\sum_{i=1}^{N(\nu)} (w_i^* + v_i^*) \overset{(b)}{\leq} \frac{1}{2}\sum_{k=1}^{N(\epsilon)} \sum_{a_i\in s_k} |w_i^* + \frac{1}{|s_k|}(\hat{\mu}_1(b_k) - \mu_1(b_k))| \\ &+ \frac{1}{2}\sum_{k=1}^{N(\epsilon)} \sum_{a_i\in s_k} |v_i^* + \frac{1}{|s_k|}(\hat{\mu}_2(b_k) - \mu_2(b_k))| - \frac{1}{2}\sum_{i=1}^{N(\nu)} (w_i^* + v_i^*) \\ &\overset{(c)}{\leq} \frac{1}{2}\sum_{i=1}^{N(\nu)} w_i^* + \frac{1}{2}\sum_{i=1}^{N(\nu)} |\hat{\mu}_1(a_i) - \mu_1(a_i)| + \frac{1}{2}\sum_{i=1}^{N(\nu)} v_i^* + \frac{1}{2}\sum_{i=1}^{N(\nu)} |\hat{\mu}_2(a_i) - \mu_2(a_i)| - \frac{1}{2}\sum_{i=1}^{N(\nu)} (w_i^* + v_i^*) \\ &= \frac{1}{2}\|\hat{\mu}_1(\epsilon) - \mu_1(\epsilon)\|_1 + \frac{1}{2}\|\hat{\mu}_2(\epsilon) - \mu_2(\epsilon)\|_1. \end{aligned}$$

Inequality (a) holds since the solution $feas(\hat{\mu}_1, \hat{\mu}_2)$ is a feasible solution of Problem (12), and may be suboptimal. Equality (b) is obtained by substituting $feas(\hat{\mu}_1, \hat{\mu}_2)$ and by noting that $C_1(\nu)$ is a refinement $C_2(\epsilon)$ (Definition 13). Inequality (c) is obtained by applying the triangle inequality on each element in the sum and noting that by definition $w_i^*, v_j^* \geq 0$.

Using an analogous procedure starting at the optimal solution for $\hat{\mu}_1(\nu), \hat{\mu}_2(\nu)$ and finding a feasible solution for distributions $\mu_1(\nu), \mu_2(\nu)$ we obtain

$$PV(\mu_1(\nu), \mu_2(\nu), \epsilon) - PV(\hat{\mu}_1(\nu), \hat{\mu}_2(\nu), \epsilon) \leq \frac{1}{2}\|\mu_1(\epsilon) - \hat{\mu}_1(\epsilon)\|_1 + \frac{1}{2}\|\mu_2(\epsilon) - \hat{\mu}_2(\epsilon)\|_1.$$

Combining the last two inequalities concludes the proof of Lemma 16. □

For the convergence rates of the discrete distributions, we use the following result provided by [15] (Theorem 2.1).

Lemma 17. *Let μ be a probability distribution on the set $\mathcal{A} = 1, ..., a$. Let $X = x_1, x_2, ..., x_N$ be i.i.d. random variables distributed according to μ, and $\hat{\mu}_N$ the resulting empirical distribution. Then, for $\eta > 0$*

$$\mathbb{P}(\|\mu - \hat{\mu}_N\|_1 \geq \eta) \leq (2^a - 2)e^{-N\eta^2/2}.$$

We are now ready to provide the proof of the main theorem.

Proof. Theorem 3

Set $\epsilon' = \frac{\epsilon(T-1)}{T}$ and $\epsilon'' = \frac{\epsilon(T+1)}{T}$, and define

$$m(T) = PV(\mu_1(\nu), \mu_2(\nu), \epsilon') - PV(\mu_1(\nu), \mu_2(\nu), \epsilon'').$$

By Lemma 14, the value of $m(T)$ is positive. Combining Lemma 14 with Lemma 16 yields

$$\begin{aligned}
&\widehat{\text{PV}}(S_1, S_2, \epsilon) \leq \text{PV}(\hat{\mu}_1(\nu), \hat{\mu}_2(\nu), \epsilon') \qquad (16)\\
&\leq \text{PV}(\mu_1(\nu), \mu_2(\nu), \epsilon') + \frac{1}{2}\|\mu_1(\epsilon') - \hat{\mu}_1(\epsilon')\|_1 + \frac{1}{2}\|\mu_2(\epsilon') - \hat{\mu}_2(\epsilon')\|_1\\
&= \text{PV}(\mu_1(\nu), \mu_2(\nu), \epsilon'') + m(T) + \frac{1}{2}\|\mu_1(\epsilon'') - \hat{\mu}_1(\epsilon'')\|_1 + \frac{1}{2}\|\mu_2(\epsilon'') - \hat{\mu}_2(\epsilon'')\|_1\\
&\leq \text{PV}(P, Q, \epsilon) + m(T) + \frac{1}{2}\|\mu_1(\epsilon') - \hat{\mu}_1(\epsilon')\|_1 + \frac{1}{2}\|\mu_2(\epsilon') - \hat{\mu}_2(\epsilon')\|_1.
\end{aligned}$$

Recall that the number of elements for an ϵ-discretization on $C_1(\epsilon)$ is $\mathcal{N}(\epsilon) = (1/\epsilon)^d$. By applying Lemma 17 to $\|\mu_1(\epsilon') - \hat{\mu}_1(\epsilon')\|_1 \leq \eta$ and $\|\mu_2(\epsilon') - \hat{\mu}_2(\epsilon')\|_1 \leq \eta$ and inserting the result to (16) using the union bound, we have that with probability at least $1 - 2(2^{(1/\epsilon')^d} - 2)e^{-N\eta^2/2}$

$$\widehat{\text{PV}}(S_1, S_2, \epsilon) - \text{PV}(P, Q, \epsilon) \leq m(T) + \eta. \qquad (17)$$

In a similar manner we have we have that with probability at least $1 - 2(2^{(1/\epsilon'')^d} - 2)e^{-N\eta^2/2}$

$$\text{PV}(P, Q, \epsilon) - \widehat{\text{PV}}(S_1, S_2, \epsilon) \leq m(T) + \eta. \qquad (18)$$

For $T \gg \epsilon$ we have that $\epsilon' \approx \epsilon'' = \epsilon$, and therefore the value of $m(T) \to 0$ as $T \to \infty$. Taking $T \to \infty$ in (17) and (18) concludes the proof. □

Proofs of Lemmas 14,15

Proof. Lemma 14

Let sample $x_i \in S_1$ belong to the element a_k in the ν-discretization, that is $x_i \in B(a_k, \|\cdot\|_\infty, \nu = \frac{\epsilon}{T})$. Recall that the ϵ-neighborhood of a sample x_i is the set $\text{ng}(x_i, \epsilon) = \{z : d(x_i, z) \leq \epsilon\}$, and the $\frac{\epsilon(T+1)}{T}$-neighborhood of bin a_k is the set $\text{ng}(a_k, \frac{\epsilon(T+1)}{T}) = \{z : d(a_k, z) \leq \frac{\epsilon(T+1)}{T}\}$. For the left side of (10), observe that for any such x_i there exists values of z such that $\|z - a_k\|_\infty \leq \frac{\epsilon(T+1)}{T}$

but $\|z - x_i\|_\infty > \epsilon$, while for any z for which $\|z - x_i\|_\infty \leq \epsilon$ also $\|z - a_k\|_\infty \leq \frac{\epsilon(T+1)}{T}$. As a result, $\text{ng}(x_i, \epsilon) \subseteq \text{ng}(a_k, \frac{\epsilon(T+1)}{T})$. Enlarging the number of neighbors adds edges to the bipartite graph describing the problem, and accordingly, a matching with a larger cardinality may be found. In such a case, the number of unmatched samples is decreased, and therefore the PV is decreased, as it is the normalized sum of the unmatched samples.

For the right hand side of (10), observe that when the discretization is $\frac{\epsilon(T-1)}{T}$, for any point $x_i \in B(a_k, \|\cdot\|_\infty, \nu)$ we have that $\text{ng}(x_i, \epsilon) \supseteq \text{ng}(a_k, \frac{\epsilon(T-1)}{T})$, as the ϵ-neighborhood of each point mass encloses the $\frac{\epsilon(T-1)}{T}$-neighborhood of its ascribed bin. As a result, the PV between the histograms $\hat{\mu}_1$ and $\hat{\mu}_2$ may correspond to a graph that has less edges, which may result in a maximum matching with a smaller cardinality. As a result, the discrete version may have a larger PV.
Inequalities (11) hold, as the same claims apply for the discretization of the distributions. □

Proof. Lemma 15

First note that any solution of Problem (2) is a feasible solution of Problem (12), and so we have that the optimum $\text{PV}(\mu_1(\nu), \mu_2(\nu), \epsilon) \geq \text{PV}_{eq}(\mu_1(\nu), \mu_2(\nu), \epsilon)$. We construct a solution of (2) that realizes the equality, and therefore is optimal. Namely, to show the problems are equivalent it is sufficient to show that any solution of (12) has a corresponding solution of (2) with the same objective value.

Let w_i, v_j, Z_{ij} be the solution to (12). In the following, we construct a feasible solution $\tilde{w}_i, \tilde{v}_i, \tilde{Z}_{ij}$ to (2):

If $w_i < 0$ and $v_i > 0$ set $\Delta_i = |w_i|$ and

$$\tilde{w}_i = w_i + \Delta_i = 0,\ \tilde{v}_i = v_i + \Delta_i > 0,\ \sum_{a_j \in \text{ng}(a_i)} \tilde{Z}_{ij} = \sum_{a_j \in \text{ng}(a_i)} Z_{ij} - \Delta_i.$$

If $v_i < 0$ and $w_i > 0$ set $\Gamma_j = |v_j|$ and

$$\tilde{v}_i = v_i + \Gamma_i = 0,\ \tilde{w}_i = w_i + \Gamma_i > 0,\ \sum_{a_j \in \text{ng}(a_i)} \tilde{Z}_{ji} = \sum_{a_j \in \text{ng}(a_i)} Z_{ji} - \Gamma_i.$$

If both $w_i < 0$ and $v_i < 0$ set

$$\tilde{w}_i = w_i + \Delta_i + \Gamma_i > 0,\ \tilde{v}_i = v_i + \Delta_i + \Gamma_i > 0,\ \sum_{a_j \in \text{ng}(a_i)} (\tilde{Z}_{ij} + \tilde{Z}_{ji}) = \sum_{a_j \in \text{ng}(a_i)} (Z_{ij} + Z_{ji}) - \Delta_i - \Gamma_i.$$

Otherwise, set $\tilde{w}_i = w_i$, $\tilde{v}_j = v_j$, and $\tilde{Z}_{ij} = Z_{ij}$.

The resulting $\tilde{w}_i, \tilde{v}_j, \tilde{Z}_{ij}$ obey the equality constraints in (2) while fixing $\tilde{w}_i \geq 0,\ \tilde{v}_j \geq 0$. It is easy to show that there exists $\tilde{Z}_{ij} \geq 0$ that obeys the equalities above. The objective value of (12) with w_i, v_j, Z_{ij} and of (2) with $\tilde{w}_i, \tilde{v}_j, \tilde{Z}_{ij}$ is equal:

$$\sum_{i=1}^{N} \tilde{w}_i + \sum_{j=1}^{N} \tilde{v}_j = \sum_{i=1}^{N} (w_i + v_i) 1_{[w_i \geq 0\,,\, v_i \geq 0]} + \sum_{i=1}^{N} ((w_i + \Delta_i) + v_i + \Delta_i) 1_{[w_i < 0\,,\, v_i \geq 0]} +$$
$$\sum_{j=1}^{N} (w_j + \Gamma_j + (v_j + \Gamma_j)) 1_{[w_j \geq 0\,,\, v_j < 0]} + \sum_{i=1}^{N} ((w_i + \Delta_i + \Gamma_i) + (v_i + \Gamma_i + \Delta_i)) 1_{[w_i < 0\,,\, v_i < 0]}$$
$$= \sum_{i=1}^{N} |w_i| + \sum_{j=1}^{N} |v_j|.$$

We conclude that $\tilde{w}_i, \tilde{v}_j, \tilde{Z}_{ij}$ attains the optimal solution to Problem (2). □

A.4 Proof of Theorem 4

We restate the theorem:

Theorem 4. *Let $P = Q$ be the uniform distribution on $\mathbb{S}^{d-1}$, a unit $(d-1)$–dimensional hypersphere. Let $S_1 = \{x_1, ..., x_N\} \sim P$ and $S_2 = \{y_1, ..., y_N\} \sim Q$ be two i.i.d. samples. For any $\epsilon, \epsilon', \delta \in (0,1)$, $0 \leq \eta < 2/3$ and sample size $\frac{\log(1/\delta)}{2(1-3\eta/2)^2} \leq N \leq \eta/2e^{d(1-\frac{\epsilon^2}{2})/2}$, we have $PV(P, Q, \epsilon') = 0$ and*

$$\mathbb{P}(\widehat{PV}(S_1, S_2, \epsilon) > \eta) \geq 1 - \delta. \tag{19}$$

Proof. We use the following definitions and lemmas.

Definition 18. *The spherical cap of radius r about a point x is*

$$C(r, x) = \left\{ z \in S^{d-1} : d(z, x) \leq r \right\}.$$

Lemma 19. *The spherical cap of radius r about a point x on a unit sphere is equal to*

$$C(r, x) = \left\{ z \in S^{d-1} :< z, x > \geq \sqrt{1 - \frac{r^2}{2}} \right\}.$$

Lemma 20. *Let $\eta = \sqrt{1 - \frac{r^2}{2}}$. For $0 \leq \eta < 1$, the cap $C(r, x)$ on $S^d - 1$ has a measure at most $e^{-d\eta^2/2}$.*

Let $p = \mathbb{P}(\text{ng}_{S_2}(x) = \emptyset)$ be the probability of an empty neighbor set. The next lemma bounds this probability.

Lemma 21. *The probability of an empty neighbor set $\mathbb{P}(ng_{S_2}(x) = \emptyset) \geq 1 - Ne^{-d(1-\frac{\epsilon^2}{2})/2}$.*

Proof.

$$\begin{aligned} p =& \mathbb{P}(\text{ng}_{S_2}(x) = \emptyset) = 1 - \mathbb{P}(\text{ng}_{S_2}(x) \neq \emptyset) = 1 - \mathbb{P}(\exists y_j \in S_2 \, ; y_j \in C(\epsilon, x_i)) \\ & \geq 1 - N\mathbb{P}(y \in C(\epsilon, x)) \geq 1 - Ne^{-d(1-\frac{\epsilon^2}{2})/2}, \end{aligned}$$

where the first inequality is due to the union bound, and the second by Lemma 20. □

We consider the probability that the $\widehat{PV}$ is grater than some $0 \leq \eta < 1$. Note, that since $PV(P, Q) = 0$ this is also the difference between the empirical and distributional PV. Let $e = \{x_i \in S_1 : \text{ng}_{S_2}(x_i) = \emptyset\}$ be the set of samples in S_1 without neighbors, and N_e its cardinality.

$$\begin{aligned} \mathbb{P}(\widehat{PV}(S_1, S_2, \epsilon) > \eta) \geq \mathbb{P}(\frac{N_e}{N} > \eta) = 1 - \mathbb{P}(N_e \leq N\eta) \geq 1 - \mathbb{P}(N_e \leq \lceil N\eta \rceil) \qquad (20) \\ = 1 - \sum_{i=0}^{\lceil N\eta \rceil} \binom{N}{i} (p)^i (1-p)^{N-i}. \end{aligned}$$

The first inequality holds, as $\widehat{PV}(S_1, S_2, \epsilon) > \eta$ is obtained when $N_e > \eta N$ samples from S_1 have no neighbors from S_2 in their ϵ-neighborhood. Note that since $n = m$ there are also exactly N_e sample from S_2 which are not matched.

By Chernoff's inequality we have that

$$\sum_{i=0}^{\lceil N\eta \rceil} \binom{n}{i} (1-p)^i p^{N-i} \leq \exp(-2N(p-\eta)^2). \tag{21}$$

Combining Equations (20) and (21) we get

$$\mathbb{P}(\widehat{PV}(S_1, S_2, \epsilon) > \eta) \geq 1 - \exp(-2N(p-\eta)^2). \tag{22}$$

By Lemma 21, we have that $p \geq 1 - Ne^{-d(1-\frac{\epsilon^2}{2})/2}$.

If $0 \leq \eta < 2/3$ and $Ne^{-d(1-\frac{\epsilon^2}{2})/2} < \eta/2$, we have that

$$p - \eta \geq 1 - Ne^{-d(1-\frac{\epsilon^2}{2})/2} - \eta > 1 - 3\eta/2 > 0.$$

Substituting the last inequality to (22):

$$\mathbb{P}(\widehat{PV}(S_1, S_2, \epsilon) > \eta) \geq 1 - \exp(-2N(1-3\eta/2)^2).$$

The theorem statement is obtained for any N, d and η for which $2N(1-3\eta/2)^2 \geq \log(\frac{1}{\delta})$. $\square$