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

Shulei Wang, T. Tony Cai and Hongzhe Li
University of Pennsylvania

In this supplementary material, we provide the proof for the main results (Section S1) and all the technical lemmas (Section S2).

S1 Proofs of Main Results

In this section, we present detailed proofs for the main results. To distinguish from the constants appeared in the previous sections, we shall use the capital letters C and c to denote generic positive constants that may take different values at each appearance.

S1.1 Proof of Proposition 1

We firstly show the upper bound. Observe that

$$
\begin{aligned}
\mathbb{E}(D(\hat{P},\hat{Q})-D(P,Q))^2 =& \mathbb{E}\left(\sum_{e\in E} L_e\left(|\hat{P}_e-\hat{Q}_e|-|P_e-Q_e|\right)\right)^2 \\
\le& \mathbb{E}\left(\sum_{e\in E} L_e\left(|\hat{P}_e-P_e|+|\hat{Q}_e-Q_e|\right)\right)^2 \\
\le& 2\mathbb{E}\left(D(\hat{P},P)^2+D(\hat{Q},Q)^2\right)
\end{aligned}
$$

Thus, it is sufficient to obtain an upper bound for $\mathbb{E}D(\hat{P},P)^2$. Decomposing $\mathbb{E}\left(D(\hat{P},P)^2\right)$ into bias and variance parts yields

$$\mathbb{E}\left(D(\hat{P},P)^2\right) = \left(\mathbb{E}D(\hat{P},P)\right)^2 + \mathrm{Var}\left(D(\hat{P},P)\right).$$

Since $n\hat{P}_e \sim \mathrm{Poi}(nP_e)$ and Lemma 9,

$$\mathbb{E}D(\hat{P},P) = \sum_{e\in E} L_e \mathbb{E}\left(|\hat{P}_e - P_e|\right) \leq 2\sum_{e\in E} L_e\left(P_e \wedge \sqrt{\frac{P_e}{n}}\right) \leq 2M\sum_{e\in E}\left(P_e \wedge \sqrt{\frac{P_e}{n}}\right).$$

To analyze the variance, we have

$$\mathrm{Var}\left(D(\hat{P},P)\right) = \sum_{e\in E} L_e^2 \mathrm{Var}\left(|\hat{P}_e - P_e|\right) + \sum_{e_1,e_2\in E} L_{e_1}L_{e_2}\mathrm{Cov}\left(|\hat{P}_{e_1} - P_{e_1}|, |\hat{P}_{e_2} - P_{e_2}|\right).$$

Hereafter, we write $e_1 \in \tau(e_2)$ if $e_2 \in [\rho, v]$ for all $v \in \tau(e_1)$. Since two edges on tree T share descendants if and only if one edge is descendant of other edge. In other word, $\tau(e') \subset \tau(e)$ if and only if $e' \in \tau(e)$. Application of Lemma 11 suggests that

$$\mathrm{Cov}\left(|\hat{P}_{e_1} - P_{e_1}|, |\hat{P}_{e_2} - P_{e_2}|\right) \begin{cases} \leq \frac{P_{e_1}}{n} & e_1 \in \tau(e_2) \\ \leq \frac{P_{e_2}}{n} & e_2 \in \tau(e_1) \\ = 0 & \text{otherwise} \end{cases}.$$

This implies that

$$\mathrm{Var}\left(D(\hat{P},P)\right) \leq \left(\sum_{e\in E}\frac{P_e}{n} + 2\sum_{e_1\in\tau(e_2)}\frac{P_{e_1}}{n}\right) \leq \frac{3d^2}{n}$$

Putting bias and variance together yields

$$\mathbb{E}\left(D(\hat{P},P)^2\right) \leq C\left(\left(\sum_{e\in E} P_e \wedge \sqrt{\frac{P_e}{n}}\right)^2 + \frac{d^2}{n}\right)$$

for some constant C. This implies

$$\mathbb{E}(D(\hat{P},\hat{Q}) - D(P,Q))^2 \leq C\left(\left(\sum_{e\in E} P_e \wedge \sqrt{\frac{P_e}{n}}\right)^2 + \left(\sum_{e\in E} Q_e \wedge \sqrt{\frac{Q_e}{n}}\right)^2 + \frac{d^2}{n}\right).$$

Next, we show the lower bound. Let v be the leaf with the largest $d(\rho, v)$ on a tree T, i.e. $d(\rho, v) = d$. Let P_1 be a distribution on tree T with probability $1/2$ at v and $1/2$ at root

ρ, i.e. $p_v = p_\rho = 1/2$ and $Q_1 = P_1$. P_2 is a distribution by putting probability $1/2 + \epsilon$ at v and $1/2 - \epsilon$ at ρ and $Q_2 = P_1$. By construction, we could know that

$$D(P_1, Q_1) = 0 \qquad \text{and} \qquad D(P_2, Q_2) = d\epsilon.$$

The Kullback-Leibler divergence between observations of (T, P_1, Q_1) and (T, P_2, Q_2) is

$$\begin{aligned} KL(P^n_{(T,P_1,Q_1)} || P^n_{(T,P_2,Q_2)}) &= n \left[\frac{1}{2} \log \left(\frac{1}{1+2\epsilon} \right) + \frac{1}{2} \log \left(\frac{1}{1-2\epsilon} \right) \right] \\ &\leq \frac{4n\epsilon^2}{1-4\epsilon^2} \end{aligned}$$

Choosing $\epsilon^2 = 1/n$ and applying Theorem 2.2 in Tsybakov (2009) yields

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

S1.2 Proof of Proposition 2

Proof. For each edge $e \in \tilde{E}(w)$, we can prove that there is at most one children edge of e belonging to $\tilde{E}(w)$. Otherwise, suppose there are two children edge of e belonging to $\tilde{E}(w)$, naming them e_1 and e_2. Then, we could know that $\sum_{v \in \tau(e_1)} x_v, \sum_{v \in \tau(e_2)} x_v > w/2$. Since e_1 and e_2 are not on the paths to root ρ of each other, Lemma 4 suggests that $\tau(e_1) \cap \tau(e_2) = \emptyset$. This suggests that $\sum_{v \in \tau(e)} x_v \geq (\sum_{v \in \tau(e_1)} x_v) + (\sum_{v \in \tau(e_2)} x_v) > w$, which contradicts that $e \in \tilde{E}(w)$.

Next, let

$$\widetilde{E}'(w) = \left\{ e \in \tilde{E}(w) : \text{no children edge of } e \text{ is in } \tilde{E}(w) \right\}.$$

For each $\tilde{e} \in \widetilde{E}'(w)$, we define its ancestor in $\tilde{E}(w)$

$$E^p_{\tilde{e}}(w) := \{ e \in \tilde{E}(w) : e \in [\tilde{v}, \rho], \ \forall \tilde{v} \in \tau(\tilde{e}) \}.$$

As $\sum_{v \in \tau(e)} x_v$ is nondecreasing along $[\rho, \tilde{v}]$, we can conclude that $E^p_{\tilde{e}}$ is connected. We can conclude that $E^p_{\tilde{e}}(w)$ is actually a path as there is at most one children edge of e belonging to $\tilde{E}(w)$ for any $e \in \tilde{E}(w)$. Therefore, we can know that

$$\tilde{E}(w) = \bigcup_{\tilde{e} \in \widetilde{E}'(w)} E^p_{\tilde{e}}(w).$$

Now, we prove $E^p_{\tilde{e}_1}(w) \cap E^p_{\tilde{e}_2}(w) = \emptyset, \forall\ \tilde{e}_1 \neq \tilde{e}_2$. Suppose there exists some $\tilde{e}_1$ and $\tilde{e}_2$ such that $E^p_{\tilde{e}_1}(w) \cap E^p_{\tilde{e}_2}(w) \neq \emptyset$. Let e' be an edge in $E^p_{\tilde{e}_1}(w) \cap E^p_{\tilde{e}_2}(w)$. Since every node in $\tilde{E}(w)$ has at most one parent and at most children in $\tilde{E}(w)$. We can conclude that $E^p_{\tilde{e}_1}(w) = E^p_{\tilde{e}_2}(w)$ and thus $\tilde{e}_1 = \tilde{e}_2$. With the same arguments, we could also prove that no edge from $E^p_{\tilde{e}_1}(w)$ is predecessor of edge in $E^p_{\tilde{e}_2}(w)$. This implies

$$\tau(e_1) \cap \tau(e_2) = \emptyset,$$

if $e_1 \in E^p_{\tilde{e}_1}(w)$ and $e_2 \in E^p_{\tilde{e}_2}(w)$ for any $\tilde{e}_1 \neq \tilde{e}_2$.

By definition, we can know that $S = |\widetilde{E}'(w)|$. Because $\tilde{e}_1 \neq \tilde{e}_2 \in \widetilde{E}'(w)$ implies $\tau(\tilde{e}_1) \cap \tau(\tilde{e}_2) = \emptyset$, we have

$$\sum_{\tilde{e} \in \widetilde{E}'} \sum_{v \in \tau(\tilde{e})} x_v \leq W.$$

As $\sum_{v \in \tau(\tilde{e})} x_v > w/2$ for $\tilde{e} \in \tilde{E}(w)$, we can conclude that $wS \leq 2W$. $\square$

S1.3 Proof of Theorem 1

We define the following events

$$B_0 = \left\{ P_e + Q_e \leq \frac{2c_1 \log n}{n}, \forall\ e \in E_0 \right\},$$

$$B_j = \left\{ |P_e - Q_e| \leq \sqrt{\frac{2c_1 \log n}{n}} \left(\sqrt{P_e + Q_e} \right), \frac{1}{2^{j+1}} \leq P_e + Q_e \leq \frac{1}{2^{j-2}},\ \forall\ e \in E_j \right\}$$

and

$$B' = \left\{ P_e < Q_e \quad \text{or} \quad Q_e < P_e, \forall\ e \in E_c \right\}.$$

Based on B_j, $j = 0, \ldots, J$ and B', we can define event

$$B = \left(\bigcap_{j=0}^{J} B_j \right) \bigcap B'.$$

By Lemma 2, all the analysis can be conducted conditioned on B as $\mathbb{P}(B) \geq 1 - 5s/n^{c_1/10}$.

Define the following events

$$\tilde{B}_0 = \left\{ \{(P_e, Q_e)\}_{e \in E_0} \in I_0 \right\}, \quad \tilde{B}_j = \left\{ \{P_e - Q_e\}_{e \in E_j} \in I_j \right\}$$

and

$$\tilde{B} = \bigcap_{j=1}^{J} \tilde{B}_j.$$

Lemma 1 suggests that

$$\mathbb{P}\left(B \bigcap \tilde{B}\right) \geq 1 - \frac{8\log n}{n^4}$$

if we choose $c_1 \geq 40$. Hereafter, we conduct the analysis conditioned on $B \bigcap \tilde{B}$.

Define the following random variables

$$L_j = \sum_{e \in E_j} L_e \left(|\tilde{P}_e - \tilde{Q}_e| - |P_e - Q_e|\right), \qquad j = 0, \dots, J$$

and

$$L' = \sum_{e \in E_c} L_e \left(|\hat{P}_e - \hat{Q}_e| - |P_e - Q_e|\right).$$

Thus,

$$\begin{aligned}
&\mathbb{E}\left(\hat{D}_{\mathrm{MET}} - D(P,Q)\right)^2 \\
\leq& 3\mathbb{E}\left(L_0^2 \mathbb{I}_{B\bigcap\tilde{B}}\right) + 3\mathbb{E}\left(\left(\sum_{j=1}^{J} L_j\right)^2 \mathbb{I}_{B\bigcap\tilde{B}}\right) + 3\mathbb{E}\left(L'^2 \mathbb{I}_{B\bigcap\tilde{B}}\right) + \frac{16d^2M^2\log n}{n^4}.
\end{aligned}$$

Here we use the fact that $D(P,Q), \hat{D}_{\mathrm{MET}} \leq dM$ for any P and Q and Cauchy-Schwarz inequality. We now bound above three terms one by one.

Firstly, we bound $\mathbb{E}\left(L_0^2 \mathbb{I}_{B\bigcap\tilde{B}}\right)$. Let $F_K^{(1)}(x,y)$ be an approximated K-polynomial of $|x-y|$ within $[0, 2c_1\log n/n]^2$ such that

$$\left||x-y| - F_K^{(1)}(x,y)\right| \leq \frac{\sqrt{x}+\sqrt{y}}{K}\sqrt{\frac{2c_1\log n}{n}} + \frac{1}{K^2}\left(\frac{2c_1\log n}{n}\right), \ \forall x, y \in [0, 2c_1\log n/n].$$

The existence of $F_K^{(1)}(x,y)$ has been shown in Lemma 20. Write

$$F_K^{(1)}(x,y) = \sum_{k_1,k_2=0}^{K} f^{(1)}(k_1,k_2) x^{k_1} y^{k_2}$$

and the coefficients $f^{(1)}(k_1,k_2)$ can be bounded by $\tilde{C}^K(2c_1\log n/n)^{1-k_1-k_2}$ for some constant $\tilde{C}$. On event $\tilde{B}$, we have

$$\left|\sum_{e \in E_0} L_e \left(\tilde{P}_e^{k_1}\tilde{Q}_e^{k_2} - P_e^{k_1}Q_e^{k_2}\right)\right| \leq 2dM\sqrt{2.5n\log^2 n}\left(\frac{76c_1\log n}{n}\right)^{k_1+k_2}, \quad 0 \leq k_1, k_2 \leq K.$$

Thus,

$$
\begin{aligned}
&\left|\sum_{e\in E_0} L_e\left(F_K^{(1)}(\tilde{P}_e,\tilde{Q}_e)-F_K^{(1)}(P_e,Q_e)\right)\right|\\
\le&\left|\sum_{e\in E_0} L_e\sum_{k_1,k_2=0}^{K} f^{(1)}(k_1,k_2)\left(\tilde{P}_e^{k_1}\tilde{Q}_e^{k_2}-P_e^{k_1}Q_e^{k_2}\right)\right|\\
\le&\sum_{k_1,k_2=0}^{K} 2dM\tilde{C}^K\sqrt{2.5n\log^2 n}\left(\frac{2c_1\log n}{n}\right)^{1-k_1-k_2}\left(\frac{76c_1\log n}{n}\right)^{k_1+k_2}\\
\le& C\frac{d(38\tilde{C})^K K^2\log^2 n}{\sqrt{n}}.
\end{aligned}
$$

On event $B\bigcap\tilde{B}$, we have

$$
\begin{aligned}
|L_0|=&\left|\sum_{e\in E_0} L_e\left(|\tilde{P}_e-\tilde{Q}_e|-|P_e-Q_e|\right)\right|\\
\le&\left|\sum_{e\in E_0} L_e\left(|\tilde{P}_e-\tilde{Q}_e|-F_K^{(1)}(\tilde{P}_e,\tilde{Q}_e)+F_K^{(1)}(\tilde{P}_e,\tilde{Q}_e)-F_K^{(1)}(P_e,Q_e)+F_K^{(1)}(P_e,Q_e)-|P_e-Q_e|\right)\right|\\
\le& 2\sum_{e\in E_0} L_e\left(\frac{\sqrt{P_e+Q_e}}{K}\sqrt{\frac{2c_1\log n}{n}}+\frac{1}{K^2}\left(\frac{2c_1\log n}{n}\right)\right)+C\frac{d(38\tilde{C})^K K^2\log^2 n}{\sqrt{n}}.
\end{aligned}
$$

As $K=c_2\log n$ for small enough constant c_2, Lemma 12 suggests

$$
\begin{aligned}
\sup_{(T,P,Q)\in\Theta(s,d)}\mathbb{E}\left(L_0^2\mathbb{I}_{B\bigcap\tilde{B}}\right)\le&\left(\sup_{(T,P,Q)\in\Theta(s,d)} 4\sum_{e\in E_0} L_e\sqrt{P_e}\sqrt{\frac{2c_1}{c_2 n\log n}}+C\frac{c_2^2 d\log^4 n}{n^{1/2-c_2\log 38\tilde{C}}}\right)^2\\
\le& C\left(\sqrt{\frac{s\log(2^{d+2}/s)}{n\log n}}+C\frac{d\log^4 n}{n^{1/2-c_2\log 38\tilde{C}}}\right)^2\\
\le& C\left(\frac{s\log(2^{d+2}/s)}{n\log n}+\frac{d^2}{n^{1-\gamma}}\right).
\end{aligned}
$$

Here $\gamma=2c_2\log 38\tilde{C}$.

Next, we bound $\mathbb{E}\left(\left(\sum_{j=1}^J L_j\right)^2\mathbb{I}_{B\bigcap\tilde{B}}\right)$. For each j, let $F_K^{(2,j)}(x)$ be a K-polynomial of $|x|$ within $[-\sqrt{4c_1\log n/2^j n},\sqrt{4c_1\log n/2^j n}]$ such that

$$
\sup_{x\in[-\sqrt{4c_1\log n/2^j n},\sqrt{4c_1\log n/2^j n}]}\left||x|-F_K^{(2,j)}(x)\right|\le\frac{1}{K}\sqrt{\frac{4c_1\log n}{2^j n}}.
$$

The existence of such polynomial has been discussed in Lemma 21. If we write

$$
F_K^{(2,j)}(x)=\sum_{k=0}^{K} f^{(2,j)}(k)x^k,
$$

then the coefficients $f^{(2,j)}(k)$ can be bounded by $\tilde{C}^K(4c_1\log n/2^jn)^{1-k}$ for some constant $\tilde{C}$. On event $\tilde{B}$, we have

$$\left|\sum_{e\in E_j} L_e\left((\tilde{P}_e-\tilde{Q}_e)^k-(P_e-Q_e)^k\right)\right| \le 4dM\sqrt{10S_j\log n}\left(\frac{48c_1\log n}{2^jn}\right)^{k/2}, \quad 0\le k\le K.$$

This suggests that

$$\begin{aligned}
&\left|\sum_{e\in E_j} L_e\left(F_K^{(2,j)}(\tilde{P}_e-\tilde{Q}_e)-F_K^{(2,j)}(P_e-Q_e)\right)\right| \\
\le&\left|\sum_{e\in E_j} L_e\sum_{k=0}^{K} f^{(2,j)}(k)\left((\tilde{P}_e-\tilde{Q}_e)^k-(P_e-Q_e)^k\right)\right| \\
\le&\sum_{k=0}^{K} 4dM\tilde{C}^K\sqrt{10S_j\log n}\left(\frac{4c_1\log n}{2^jn}\right)^{1/2-k/2}\left(\frac{48c_1\log n}{2^jn}\right)^{k/2} \\
\le& C\frac{d(12\tilde{C})^K K\log n}{\sqrt{n}}.
\end{aligned}$$

Here we use $S_j\le 2^{j+2}$. On event $B\bigcap\tilde{B}$, we have

$$\begin{aligned}
\left|\sum_{j=1}^{J} L_j\right| &\le \sum_j\left|\sum_{e\in E_j} L_e\left(|\Delta_e|-|P_e-Q_e|\right)\right| \\
&\le \sum_j\left|\sum_{e\in E_j} L_e\left(|\Delta_e|-F_K^{(j)}(\Delta_e)+F_K^{(2,j)}(\Delta_e)-F_K^{(2,j)}(P_e-Q_e)+F_K^{(2,j)}(P_e-Q_e)-|P_e-Q_e|\right)\right| \\
&\le \sum_j\left(2\sum_{e\in E_j} L_e\left(\frac{1}{K}\sqrt{\frac{4c_1\log n}{2^jn}}\right)+C\frac{d(12\tilde{C})^K K\log n}{\sqrt{n}}\right).
\end{aligned}$$

Here $\Delta_e := \tilde{P}_e-\tilde{Q}_e$. On event $B\bigcap\tilde{B}$, we also know that $P_e+Q_e\ge 2^{-(j+1)}$ when $e\in E_j$. Thus,

$$\left|\sum_{j=1}^{J} L_j\right| \le \sum_j\left(2\sum_{e\in E_j} L_e\left(\frac{1}{K}\sqrt{\frac{8c_1(P_e+Q_e)\log n}{n}}\right)+C\frac{d(12\tilde{C})^K K\log n}{\sqrt{n}}\right).$$

Together with choice of K and Lemma 12, we have

$$\begin{aligned}
&\sup_{(T,P,Q)\in\Theta(s,d)} \mathbb{E}\left(\left(\sum_{j=1}^{J} L_j\right)^2 \mathbb{I}_{B\bigcap\tilde{B}}\right) \\
\leq &\left(4 \sup_{(T,P,Q)\in\Theta(s,d)} \sum_{e\in E_r} L_e\sqrt{P_e}\sqrt{\frac{c_1}{c_2 n\log n}} + C\frac{d\log^2 n}{n^{1/2-c_2\log 12\tilde{C}}}\right)^2 \\
\leq &C\left(\sqrt{\frac{s\log(2^{d+2}/s)}{n\log n}} + \frac{d\log^2 n}{n^{1/2-c_2\log 12\tilde{C}}}\right)^2 \\
\leq &C\left(\frac{s\log(2^{d+2}/s)}{n\log n} + \frac{d^2}{n^{1-\gamma}}\right).
\end{aligned}$$

Finally, we bound the last term $\mathbb{E}\left(L'^2\mathbb{I}_{B\bigcap\tilde{B}}\right)$. As $\hat{P}_e - \hat{Q}_e$ is unbiased estimator on event $B\bigcap\tilde{B}$ when $e\in E_c$. With the same arguments in proof of Proposition 1, we have

$$\mathbb{E}\left(L'^2\mathbb{I}_{B\bigcap\tilde{B}}\right) \leq \frac{d^2}{n}.$$

We now put three terms together.

$$\sup_{(T,P,Q)\in\Theta(s,d)} \mathbb{E}\left(\hat{D}_{\mathrm{MET}} - D(P,Q)\right)^2 \leq C\left(\frac{s\log(2^{d+2}/s)}{n\log n} + \frac{d^2}{n^{1-\gamma}}\right) + \frac{d^2}{n} + \frac{16d^2M^2\log n}{n^4}.$$

Because $\log n \leq C_1\log(s/d)$, we can choose c_2 small enough so that

$$\sup_{(T,P,Q)\in\Theta(s,d)} \mathbb{E}\left(\hat{D}_{\mathrm{MET}} - D(P,Q)\right)^2 \leq C\frac{s\log(2^{d+2}/s)}{n\log n}.$$

S1.4 Proof of Theorem 2

We now prove lower bound $s\log(2^{d+2}/s)/n\log n$. To the end, we provide a lower bound when Q is known, i.e. we have infinite number of sample from Q. The minimax risk when Q is known can be defined as

$$R^*(s,d,Q) = \inf_{\hat{D}} \sup_{(T,P,Q)\in\Theta(s,d,Q)} \mathbb{E}(\hat{D} - D(P,Q))^2,$$

where

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

Clearly,

$$R^*(s,d) \geq \sup_Q R^*(s,d,Q).$$

Thus, the rest of proof aims to find the hardest case Q and show a lower bound of $R^*(s,d,Q)$.

Let $\mathcal{M}_s(\epsilon)$ be an vector set

$$\mathcal{M}_s(\epsilon) := \left\{ P : \left| \sum_{v \in V} p_v - 1 \right| \leq \epsilon \right\}$$

and

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

The minimax rate under Poisson model can be generalized accordingly

$$\tilde{R}^*(s,d,Q,\epsilon) := \inf_{\hat{D}} \sup_{(T,P,Q) \in \Theta(s,d,Q,\epsilon)} \mathbb{E}(\hat{D} - D(P,Q))^2.$$

Lemma 5 suggests that it is sufficient to provide a lower bound of $\tilde{R}^*(s,d,Q,\epsilon)$ where ϵ is specified later.

To show a lower bound of $\tilde{R}^*(s,d,Q,\epsilon)$, we adopt the method of two fuzzy hypothesis in Tsybakov (2009). Our strategy is first to construct a least favorable tree and then construct two prior probability measures for P and Q. Our construction of least favorable tree relies on two elementary tree: full binary tree and chain tree. A full binary tree is a tree in which every non-leaf node has exactly two children. A typical example is shown in Figure 1. A full binary tree with depth d has $2^{d+1} - 1$ nodes and 2^d leaves. A chain tree is a binary tree in which right children of non-leaf node is a leaf. An example of chain tree can be found in Figure 2. A chain tree with depth d has $2d - 1$ nodes and d leaves.

Now we construct the least favorable tree $T_0(k_1,k_2)$ for some constant k_1 and k_2. The top part of $T_0(k_1,k_2)$ is a complete binary tree T_1 with depth k_1. At each leaf of T_1, we link a chain tree with depth k_2. There are totally 2^{k_1} chain tree attached to T_1, naming them as $T_{2,i}$, $i = 1,\ldots,2^{k_1}$. An example of $T_0(k_1,k_2)$ is shown in Figure 3. We choose $k_1 = \lfloor \log_2(s/\log(2^{d+2}/s)) \rfloor$ and $k_2 = \lfloor \log(2^{d+2}/s) \rfloor$. Choices of k_1 and k_2 suggests that $k_1 + k_2 \leq d$ and $2^{k_1}(k_2+2) \leq s$. Clearly, each subtree $T_{2,i}$ has only two leaves with depth $k_1 + k_2$ and name the left one of them as $v_{0,i}$. Let V_0 be a collection of $v_{0,i}$, i.e. $V_0 = \{v_{0,i}, 1 \leq i \leq 2^{k_1}\}$. Observe that $|V_0| = 2^{k_1}$.

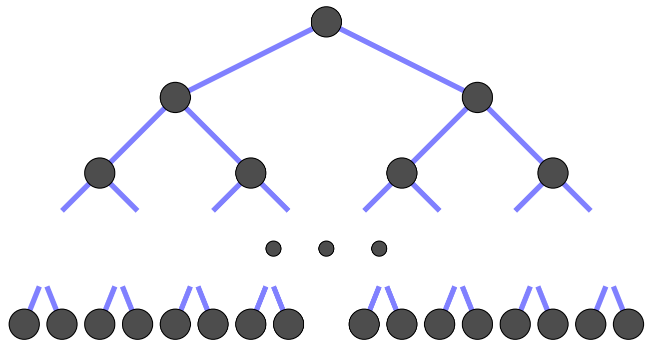

Figure 1: Full Binary Tree

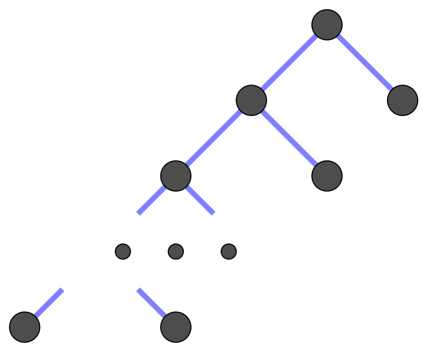

Figure 2: Chain Tree

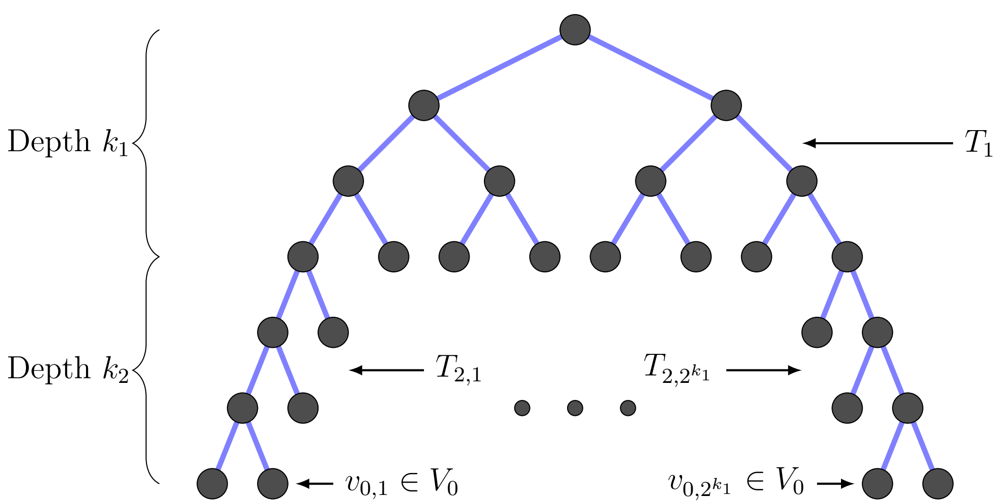

Figure 3: Least Favorable Tree $T_0(k_1, k_2)$

Now, we construct the probability distribution on $T_0(k_1, k_2)$. The probability distribution $Q = Q_1 = Q_2$ put probability $q = 2^{-k_1}$ at each node in V_0 and 0 at other nodes, i.e.

$$q_v = \begin{cases} q & v \in V_0 \\ 0 & v \in V \setminus V_0 \end{cases}.$$

We fix the distribution Q and construct the two prior probability measures $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ on distribution P. We assume the prior distribution on each node are independent, i.e.

$$\boldsymbol{\mu}_1 = \prod_{v \in V} \mu_{1,v} \qquad \text{and} \qquad \boldsymbol{\mu}_2 = \prod_{v \in V} \mu_{2,v}.$$

Similar with construction of Q, we assume p_v is always 0 when $v \notin V_0$ and the prior distributions are the same when $v \in V_0$, i.e.

$$\mu_{i,v} = \begin{cases} \mu_i & v \in V_0 \\ \delta_{(0)} & v \in V \setminus V_0 \end{cases}, \qquad i = 1, 2$$

where $\delta_{(0)}$ is a probability distribution with probability 1 being 0. Suppose ν_1 and ν_2 are two distributions in Lemma 22 and $f(x) = q + x\lambda$. Then we define μ_1 and μ_2 as $\mu_i(A) = \nu_i(f^{-1}(A))$. Then μ_1 and μ_2 are a pair of distributions on $[q - \lambda, q + \lambda]$ such that

$$\int t\mu_1(dt) = \int t\mu_2(dt) = q,$$

$$\int t^k \mu_1(dt) = \int t^k \mu_2(dt), \quad k = 2, \dots, K$$

and

$$\int |t - q|\mu_1(dt) - \int |t - q|\mu_2(dt) = c\frac{\lambda}{K}.$$

Under prior probability measures $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, we have

$$\Delta := \mathbb{E}_{\boldsymbol{\mu}_1} D(P, Q) - \mathbb{E}_{\boldsymbol{\mu}_2} D(P, Q) = ck_2 2^{k_1} \frac{\lambda}{K} = c\frac{s\lambda}{K}$$

and

$$\mathbb{E}_{\boldsymbol{\mu}_1}\left(\sum_{v \in V} p_v\right) = \mathbb{E}_{\boldsymbol{\mu}_2}\left(\sum_{v \in V} p_v\right) = 1.$$

As the support of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are larger than $\mathcal{M}_{|V|}(\epsilon)$, we consider the following subset

$$B_i = \left\{ P : \left| \sum_{v \in V} p_v - 1 \right| \leq \epsilon, |D(P,Q) - \mathbb{E}_{\boldsymbol{\mu}_i} D(P,Q)| \leq \frac{\Delta}{4} \right\}, \qquad i = 1, 2.$$

Hoeffding inequality and McDiarmid inequality (see, e.g. Boucheron et al., 2013) suggests that

$$\mathbb{P}_{\boldsymbol{\mu}_i} \left(\left| \sum_{v \in V} p_v - 1 \right| > t \right) \leq 2 \exp \left(- \frac{Ct^2}{2^{k_1} \lambda^2} \right)$$

and

$$\mathbb{P}_{\boldsymbol{\mu}_i} \left(\left| D(P,Q) - \mathbb{E}_{\boldsymbol{\mu}_i} D(P,Q) \right| > t \right) \leq 2 \exp \left(- \frac{Ct^2}{2^{k_1} k_2^2 \lambda^2} \right) \leq 2 \exp \left(- \frac{Ct^2}{s k_2 \lambda^2} \right).$$

Choosing $\epsilon = \Delta / 4d$ yields

$$\boldsymbol{\mu}_i(B_i) \geq 1 - 4 \exp \left(- \frac{Cs}{K^2 k_2} \right).$$

Let π_1 and π_2 be a pair of prior distribution measures conditioned on B_1 and B_2

$$\pi_i(A) = \frac{\boldsymbol{\mu}_i(A \cap B_i)}{\boldsymbol{\mu}_i(B_i)}.$$

When the prior distribution is π_i or $\boldsymbol{\mu}_i$, we define $\mathbb{P}_{\pi_i}$ and $\mathbb{P}_{\boldsymbol{\mu}_i}$ as corresponding marginal distribution of observed data. Then, Lemma 7 implies

$$\begin{aligned} TV(\mathbb{P}_{\pi_1}, \mathbb{P}_{\pi_2}) &\leq TV(\mathbb{P}_{\pi_1}, \mathbb{P}_{\boldsymbol{\mu}_1}) + TV(\mathbb{P}_{\boldsymbol{\mu}_1}, \mathbb{P}_{\boldsymbol{\mu}_2}) + TV(\mathbb{P}_{\pi_2}, \mathbb{P}_{\boldsymbol{\mu}_2}) \\ &\leq 1 - \boldsymbol{\mu}_1(B_1) + 1 - \boldsymbol{\mu}_2(B_2) + 2^{k_1+1} \left(\frac{e\lambda\sqrt{n}}{\sqrt{q(K+1)}} \right)^{K+1} \\ &\leq 8 \exp \left(- \frac{Cs}{K^2 k_2} \right) + 2^{k_1+1} \left(\frac{e\lambda\sqrt{n}}{\sqrt{q(K+1)}} \right)^{K+1} \end{aligned}$$

We are ready to use prior distribution measures π_1 and π_2 and Lemma 6 to lower bound $\tilde{R}^*(s,d,Q,\epsilon)$. Lemma 6 suggests that

$$\begin{aligned} &\inf_{\hat{D}} \sup_{\theta \in \Theta(s,d,Q,\epsilon)} \mathbb{P}_\theta \left(|\hat{D} - D(P,Q)| \geq c \frac{s\lambda}{K} \right) \\ \geq &\frac{1}{2} \left[1 - 8 \exp \left(- \frac{Cs}{K^2 \log(2^{d+2}/s)} \right) + \frac{2s}{\log(2^{d+2}/s)} \left(\frac{e\lambda\sqrt{n}}{\sqrt{q(K+1)}} \right)^{K+1} \right]. \end{aligned}$$

Now we choose λ and K. We only need to show lower bound when

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

When $s/\log(2^{d+2}/s)<\log^2 n$, let

$$K=c_1\sqrt{\log n}\qquad\text{and}\qquad\lambda=c_2\sqrt{\frac{k_2}{sn}}.$$

We can obtain

$$\tilde{R}^*(s,d,Q,\epsilon)\geq c\left(\frac{s\log(2^{d+2}/s)}{n\log n}\right).$$

When $s/\log(2^{d+2}/s)\geq\log^2 n$, we choose

$$K=c_1\log n\qquad\text{and}\qquad\lambda=c_2\sqrt{\frac{k_2\log n}{sn}}.$$

Small enough c_1 and c_2 suggests that

$$\tilde{R}^*(s,d,Q,\epsilon)\geq c\left(\frac{s\log(2^{d+2}/s)}{n\log n}\right).$$

We can complete proof by applying Lemma 5.

S1.5 Proof of Theorem 3

We first prove the upper bound. Proposition 1 and Lemma 12 suggest that suggests

$$\sup_{P\in\mathcal{M}_s}\left(\sum_{e\in E}P_e\wedge\sqrt{\frac{P_e}{n}}\right)^2\leq C\frac{s\log(2^{d+2}/s)}{n}.$$

So we have

$$\sup_{(T,P,Q)\in\Theta(s,d)}\mathbb{E}(D(\hat{P},\hat{Q})-D(P,Q))^2\leq C\left(\frac{s\log(2^{d+2}/s)}{n}+\frac{d^2}{n}\right)$$

Now, let's turn to the lower bound. Because of Jensen's inequality,

$$\mathbb{E}(D(\hat{P},\hat{Q})-D(P,Q))^2\geq\left(\sum_{e\in E}L_e\mathbb{E}\left(|\hat{P}_e-\hat{Q}_e|-|P_e-Q_e|\right)\right)^2.$$

Application of conditional Jensen's inequality and Lemma 9 suggests

$$\begin{aligned}\sup_{P\in\mathcal{M}_s}\mathbb{E}\left(|\hat{P}_e-\hat{Q}_e|-|P_e-Q_e|\right)&\geq\mathbb{E}\left(|\hat{Q}_e-Q_e|\right)\\&\geq\frac{1}{\sqrt{2}}\left(Q_e\wedge\sqrt{\frac{Q_e}{n}}\right).\end{aligned}$$

So we have

$$\sup_{P\in\mathcal{M}_s} \mathbb{E}(D(\hat{P},\hat{Q})-D(P,Q))^2 \geq C\left(\sum_{e\in E} L_e \left(Q_e \wedge \sqrt{\frac{Q_e}{n}}\right)\right)^2.$$

Taking supreme with respect to $Q\in\mathcal{M}_s$, Lemma 12 implies

$$\sup_{(T,P,Q)\in\Theta(s,d)} \mathbb{E}(D(\hat{P},\hat{Q})-D(P,Q))^2 \geq C\frac{s\log(2^{d+2}/s)}{n}$$

With lower bound in Proposition 1, we could know that

$$\sup_{(T,P,Q)\in\Theta(s,d)} \mathbb{E}(D(\hat{P},\hat{Q})-D(P,Q))^2 \geq C\left(\frac{s\log(2^{d+2}/s)}{n}+\frac{d^2}{n}\right).$$

Because $d^2 \leq s\log(2^{d+2}/s)$, we prove

$$\sup_{(T,P,Q)\in\Theta(s,d)} \mathbb{E}(D(\hat{P},\hat{Q})-D(P,Q))^2 \asymp \frac{s\log(2^{d+2}/s)}{n}.$$

S1.6 Proof of Theorem 4

We firstly show the bias of classical plugin estimator can be bounded by d/n. By Lemma 14,

$$\begin{aligned}\left|\mathbb{E}(D_\alpha(\hat{P},\hat{Q}))-D_\alpha(P,Q)\right| &\leq \sum_{e\in E} L_e \left|\mathbb{E}|\hat{P}_e-\hat{Q}_e|^\alpha - |P_e-Q_e|^\alpha\right| \\ &\leq C\sum_{e\in E} L_e\frac{P_e+Q_e}{n} \\ &\leq C\frac{d}{n}.\end{aligned}$$

Next, we show the variance of $D_\alpha(\hat{P},\hat{Q})$ is always bounded by d^2/n. To the end, we would like to apply Lemma 17 directly. Putting bias and variance together yields

$$\mathbb{E}(D_\alpha(\hat{P},\hat{Q})-D_\alpha(P,Q))^2 \leq C\frac{d^2}{n}.$$

We omit the proof of lower bound as it can be proven in the exactly same way in Proposition 1.

S1.7 Proof of Theorem 5

We now show the upper bound and lower bound of MET when $0<\alpha<1$.

S1.7.1 Upper bound

We follow the same notation and proof pipeline in proof of Theorem 1. Following the arguments there yields

$$\mathbb{E}\left(L_0^2 \mathbb{I}_{B\bigcap \tilde{B}}\right) \le C\left(\frac{s^{2-\alpha}\log^{\alpha}(2^{d+2}/s)}{(n\log n)^{\alpha}} + \frac{d^2}{n^{1-\gamma}}\right)$$

and

$$\mathbb{E}\left(\left(\sum_{j=1}^{J} L_j\right)^2 \mathbb{I}_{B\bigcap \tilde{B}}\right) \le C\left(\frac{s^{2-\alpha}\log^{\alpha}(2^{d+2}/s)}{(n\log n)^{\alpha}} + \frac{d^2}{n^{1-\gamma}}\right).$$

The main difference is the definition of L' and the way to bound it. More concretely, L' can be defined as

$$L' = \sum_{e\in E_c} L_e\left(U_\alpha(\hat{P}_{e,1},\hat{Q}_{e,1}) - |P_e - Q_e|^{\alpha}\right).$$

To get a bound for $\mathbb{E}\left(L'^2\mathbb{I}_{B\bigcap\tilde{B}}\right)$, we work on the bias and variance separately. We firstly work on bias. As

$$\mathbb{E}\left(L'|B\bigcap\tilde{B}\right) = \sum_{e\in E_c} L_e\left(\mathbb{E}U_\alpha(\hat{P}_{e,1},\hat{Q}_{e,1}) - |P_e - Q_e|^{\alpha}\right),$$

an application of Lemma 13 and Lemma 12 yields

$$\begin{aligned}\mathbb{E}\left(L'|B\bigcap\tilde{B}\right) &\le C\sum_{e\in E_c} L_e\left(\frac{(P_e+Q_e)^{\alpha/2}}{n^{\alpha/2}\log^{(4-\alpha)/2} n} + \frac{P_e+Q_e}{n^{c_1-4}}\right)\\ &\le C\frac{s^{(2-\alpha)/2}\log^{\alpha/2}(2^{d+2}/s)}{(n\log n)^{\alpha/2}}.\end{aligned}$$

Next, we work on the variance of L'. Observe

$$\begin{aligned}\mathrm{Var}(L') &= \sum_{e\in E_c} L_e^2\mathrm{Var}\left(U_\alpha(\hat{P}_{e,1},\hat{Q}_{e,1})\right) + \sum_{e_1,e_2\in E_c} L_{e_1}L_{e_2}\mathrm{Cov}\left(U_\alpha(\hat{P}_{e_1,1},\hat{Q}_{e_1,1}), U_\alpha(\hat{P}_{e_2,1},\hat{Q}_{e_2,1})\right)\\ &\le 2\sum_{e\in E_c} L_e\left(L_e\mathrm{Var}\left(U_\alpha(\hat{P}_{e,1},\hat{Q}_{e,1})\right) + \sum_{e'\in \mathrm{P}(e)} L_{e'}\mathrm{Cov}\left(U_\alpha(\hat{P}_{e_1,1},\hat{Q}_{e_1,1}), U_\alpha(\hat{P}_{e_2,1},\hat{Q}_{e_2,1})\right)\right),\end{aligned}$$

we apply Lemma 15

$$
\begin{aligned}
\operatorname{Var}(L') \leq & 2C\Bigg(\sum_{e\in E_c} L_e|P_e-Q_e|^{\alpha-1}\frac{P_e+Q_e}{n}\sum_{e'\in \mathrm{P}(e)} L_{e'}|P_{e'}-Q_{e'}|^{\alpha-1} \\
& \quad + \sum_{e\in E_c} L_e|P_e-Q_e|^{\alpha-1}\frac{\sqrt{P_e+Q_e}}{n\log^2 n}\sum_{e'\in \mathrm{P}(e)} L_{e'}|P_{e'}-Q_{e'}|^{\alpha-1}\sqrt{P_{e'}+Q_{e'}}\Bigg) \\
=: & 2C(T_1+T_2)
\end{aligned}
$$

Here $\mathrm{P}(e)=\{e'\in E_c : e'\in[\rho,v], \forall v \text{ such that } e\in[\rho,v]\}$. In other words, $\mathrm{P}(e)$ is all parent edges. We bound T_1 and T_2 with different strategies. In particular,

$$
\begin{aligned}
T_1 \leq & \sum_{e\in E_c} L_e\left(\frac{c_1(P_e+Q_e)\log n}{n}\right)^{(\alpha-1)/2}\frac{P_e+Q_e}{n}\sum_{e'\in \mathrm{P}(e)} L_{e'}\left(\frac{c_1(P_{e'}+Q_{e'})\log n}{n}\right)^{(\alpha-1)/2} \\
\leq & C\sum_{e\in E_c} L_e\frac{(P_e+Q_e)^{(\alpha+1)/2}}{n^{(\alpha+1)/2}\log^{(\alpha+1)/2} n}\cdot dM\left(\frac{(P_e+Q_e)\log n}{n}\right)^{(\alpha-1)/2} \\
\leq & C\sum_{e\in E_c} L_e\frac{d(P_e+Q_e)^{\alpha}}{n^{\alpha}\log^{\alpha} n} \\
\leq & C\frac{s^{1-\alpha}\log^{\alpha}(2^{d+2}/s)d}{n^{\alpha}\log^{\alpha} n}.
\end{aligned}
$$

Next, we work on T_2. Clearly, an application of Lemma 12 suggests

$$
\begin{aligned}
T_2 & \leq \frac{C}{n\log^2 n}\left(\sum_{e\in E_c}|P_e-Q_e|^{\alpha-1}\sqrt{P_e+Q_e}\right)^2 \\
& \leq \frac{C}{n\log^2 n}\left(\sum_{e\in E_c}\left(\frac{(P_e+Q_e)\log n}{n}\right)^{(\alpha-1)/2}\sqrt{P_e+Q_e}\right)^2 \\
& \leq \frac{C}{n^{\alpha}\log^{3-\alpha} n}\left(\sum_{e\in E_c}(P_e+Q_e)^{\alpha/2}\right)^2 \\
& \leq \frac{C}{n^{\alpha}\log^{3-\alpha} n}s^{2-\alpha}\log^{\alpha}(2^{d+2}/s)
\end{aligned}
$$

Putting T_1, T_2 and $\mathbb{E}\left(L'|B\bigcap\tilde{B}\right)$ together yields

$$
\mathbb{E}\left(L'^2\mathbb{I}_{B\bigcap\tilde{B}}\right) \leq C\frac{s^{2-\alpha}\log^{\alpha}(2^{d+2}/s)}{(n\log n)^{\alpha}}
$$

When we choose c_2 small enough to make $1-\gamma>\alpha$, we show that

$$
\mathbb{E}(\hat{D}_{\mathrm{MET},\alpha}-D_\alpha(P,Q))^2 \leq C\frac{s^{2-\alpha}\log^{\alpha}(2^{d+2}/s)}{(n\log n)^{\alpha}}.
$$

S1.7.2 Lower bound

Next, we show the lower bound. We follow the pipeline in proof of Theorem 2. Let $T_0(k_1, k_2)$ still be the least favorable tree with $k_1 = \lfloor \log_s(s/\log(2^{d+2}/s)) \rfloor$ and $k_2 = \lfloor \log(2^{d+2}/s) \rfloor$. We also put uniform probability in V_0 for Q_1 and Q_2 and 0 other nodes. We still construct distribution P with prior distribution $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. In particular, by Lemma 21 and Lemma 22 μ_1 and μ_2 are a pair of distributions on $[q-\lambda, q+\lambda]$ such that

$$\int t\mu_1(dt) = \int t\mu_2(dt) = q,$$

$$\int t^k\mu_1(dt) = \int t^k\mu_2(dt), \qquad l = 2, \dots, K$$

and

$$\int |t-q|^\alpha \mu_1(dt) - \int |t-q|^\alpha \mu_2(dt) = c\left(\frac{\lambda}{K}\right)^\alpha.$$

Under these two prior measures $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, we have

$$\Delta_\alpha := \mathbb{E}_{\boldsymbol{\mu}_1} D_\alpha(P,Q) - \mathbb{E}_{\boldsymbol{\mu}_2} D_\alpha(P,Q) = cs\left(\frac{\lambda}{K}\right)^\alpha.$$

With the same arguments in proof of Theorem 2, we have

$$R^*_\alpha(s,d) \geq \frac{s^2\lambda^{2\alpha}}{2K^{2\alpha}}\left[1 - 8\exp\left(-\frac{Cs\lambda^{2\alpha-2}}{K^{2\alpha}k_2}\right) + \frac{2s}{k_2}\left(\frac{e\lambda\sqrt{n}}{\sqrt{q(K+1)}}\right)^{K+1}\right].$$

We choose K and λ as

$$K = c_1 \log n \qquad \text{and} \qquad \lambda = c_2\sqrt{\frac{q\log n}{n}}.$$

By choice of K and λ, we have

$$\frac{s\lambda^{2\alpha-2}}{K^{2\alpha}k_2} \asymp \frac{n^{1-\alpha}(s/k_2)^{2-\alpha}}{\log^{1+\alpha} n} \qquad \text{and} \qquad \frac{2s}{k_2}\left(\frac{e\lambda\sqrt{n}}{\sqrt{q(K+1)}}\right)^{K+1} \asymp \frac{s}{k_2 n^{C_1}},$$

where $C_1 = c_1(1 + \log c_2 - \log(c_1)/2)$. If we choose c_2 small enough and c_1 large enough, we have

$$R^*_\alpha(s,d) \geq C_2 \frac{s^{2-\alpha}\log^\alpha(2^{d+2}/s)}{(n\log n)^\alpha},$$

when n is large enough. Here, $C_2 = (c_2/c_1)^{2\alpha}/4$. This also suggests that $R^*(s,d) \geq C_2$, when $(n\log n)^\alpha \leq s^{2-\alpha}\log^\alpha(2^{d+2}/s)$.

S1.8 Proof of Theorem 6

We now work on the case where $1 < \alpha < 2$.

S1.8.1 Upper bound

Performance of MET when $r(s,d,n) < T(\alpha)$ The proof is the same with previous case ($0 < \alpha < 1$) except the way to bound variance of L'. To bound the variance, we adopt Efron-Stein inequality, which is also used in Lemma 17. More specifically, if we follow the notation in Lemma 17, we apply Efron-Stein inequality with respect to $\hat{p}_{v,1}$ and $\hat{q}_{v,1}$. For arbitrary $v_0 \in V$, $\hat{P}'$ is the sample where $\hat{p}_{v_0}$ is replaced independent copy $\hat{p}'_{v_0}$. For any $e \in E_c$ such that $v_0 \in \tau(e)$, an application of Lemma 16 yields

$$\mathbb{E}(U_\alpha(\hat{P}_{e,1}, \hat{Q}_{e,1}) - U_\alpha(\hat{P}'_{e,1}, \hat{Q}_{e,1}))^2 \le C\left(\frac{p_{v_0}}{n} + \frac{1}{n^{\alpha+c_1/4}}\right).$$

The Efron-Stein inequality suggests that

$$\mathrm{Var}(L') \le C\left(\sum_{v\in V} d^2\left(\frac{p_v + q_v}{n} + \frac{2}{n^{\alpha+c_1/4}}\right)\right) \le C\left(\frac{d^2}{n} + \frac{s}{n^{\alpha+c_1/4}}\right) \le C\frac{d^2}{n}.$$

Therefore, if we choose c_2 small enough to make $1-\gamma > \alpha$, putting all terms together yields

$$\mathbb{E}(\hat{D}_{\mathrm{MET},\alpha} - D_\alpha(P,Q))^2 \le C\left(\frac{s^{2-\alpha}\log^\alpha(2^{d+2}/s)}{(n\log n)^\alpha}\right).$$

Performance of plugin estimator when $r(s,d,n) \ge T(\alpha)$ We follow the similar strategy in Proposition 1. Since $|x|^\alpha$ is a convex function, we have

$$\begin{aligned}
\mathbb{E}\left(D_\alpha(\hat{P},\hat{Q}) - D_\alpha(P,Q)\right)^2 &= \mathbb{E}\left(\sum_{e\in E} L_e\left(|\hat{P}_e - \hat{Q}_e|^\alpha - |P_e - Q_e|^\alpha\right)\right)^2 \\
&\le \mathbb{E}\left(3\sum_{e\in E} L_e\left(|\hat{P}_e - P_e|^\alpha + |\hat{Q}_e - Q_e|^\alpha\right)\right)^2 \\
&\le 18\mathbb{E}\left(D_\alpha(\hat{P},P)^2 + D_\alpha(\hat{Q},Q)^2\right).
\end{aligned}$$

Thus, it is enough to show an upper bound for $\mathbb{E}D_\alpha(\hat{P},P)^2$. To the end, we work on the bias and variance separately. First, we work on the bias

$$\mathbb{E}D_\alpha(\hat{P},P) = \sum_{e\in E} L_e\mathbb{E}\left(|\hat{P}_e - P_e|^\alpha\right) \le \sum_{e\in E} L_e\left(\mathbb{E}|\hat{P}_e - P_e|^2\right)^{\alpha/2} \le \sum_{e\in E} L_e\left(\frac{P_e}{n}\right)^{\alpha/2}.$$

Here, we use the Jensen's inequality. Next, we apply Efron-Setein inequality to bound variance as in Lemma 17. With the same arguments there, we can get

$$\mathrm{Var}(D_\alpha(\hat{P}, P)) \leq C\frac{d^2}{n}.$$

By Lemma 12, putting bias and variance together yields

$$\begin{aligned}
&\sup_{(T,P,Q)\in\Theta(s,d)} \mathbb{E}\left(D_\alpha(\hat{P},\hat{Q}) - D_\alpha(P,Q)\right)^2 \\
\leq& C \sup_{(T,P,Q)\in\Theta(s,d)} \left(\frac{(\sum_{e\in E} P_e^{\alpha/2})^2}{n^\alpha} + \frac{(\sum_{e\in E} Q_e^{\alpha/2})^2}{n^\alpha} + \frac{d^2}{n}\right) \\
\leq& C\left(\frac{s^{2-\alpha}\log^\alpha(2^{d+2}/s)}{n^\alpha} + \frac{d^2}{n}\right).
\end{aligned}$$

Because $(\alpha-1)\log n \geq (2-\alpha)\log(s/d)$, we can conclude

$$\sup_{(T,P,Q)\in\Theta(s,d)} \mathbb{E}\left(D_\alpha(\hat{P},\hat{Q}) - D_\alpha(P,Q)\right)^2 \leq C\frac{d^2}{n}.$$

S1.8.2 Lower bound

The lower bound of d^2/n part can be proven in the exactly same way in Proposition 1. So, we only focus the bias part when $1 < \alpha < 2$. It is sufficient to prove the lower bound when

$$\frac{s^{2-\alpha}\log^\alpha(2^{d+2}/s)}{(n\log n)^\alpha} \gg \frac{d^2}{n}. \tag{S1.1}$$

As in the last regimes, we could follow the exact steps in proof of Theorem 2 and obtain

$$R_\alpha^*(s,d) \geq \frac{s^2\lambda^{2\alpha}}{2K^{2\alpha}}\left[1 - 8\exp\left(-\frac{Cs\lambda^{2\alpha-2}}{K^{2\alpha}k_2}\right) + \frac{2s}{k_2}\left(\frac{e\lambda\sqrt{n}}{\sqrt{q(K+1)}}\right)^{K+1}\right].$$

When $(s/k_2)^{2-\alpha} \geq n^{\alpha-1}\log^{1+\alpha} n$, if choose

$$K = c_1\log n \qquad \text{and} \qquad \lambda = c_2\sqrt{\frac{q\log n}{n}},$$

then we have

$$R_\alpha^*(s,d) \geq C_2\frac{s^{2-\alpha}\log^\alpha(2^{d+2}/s)}{(n\log n)^\alpha}.$$

On the other hand, if $(s/k_2)^{2-\alpha} < n^{\alpha-1}\log^{1+\alpha} n$, we can choose

$$K = c_1\sqrt{\log n} \qquad \text{and} \qquad \lambda = c_2\sqrt{\frac{q}{n}}.$$

As (S1.1) suggests that $(s/k_2)^{2-\alpha} \gg n^{\alpha-1}\log^\alpha n$, we can know

$$R_\alpha^*(s,d) \geq C_2 \frac{s^{2-\alpha}\log^\alpha(2^{d+2}/s)}{(n\log n)^\alpha}$$

when c_1 and c_2 are chosen in the previous case.

S2 Auxiliary Results and Proofs

In this section, we present all the technical lemmas.

Lemma 1. *Suppose for any $e \in E_0$, (P_e, Q_e) satisfies $n(P_e + Q_e) \leq 2c_1 \log n$. We further assume $n \geq |E|/\log|E|$. If $K = c\log n$ for some constant $c < c_1$, then*

$$\mathbb{P}\bigg(\{(P_e, Q_e)\}_{e\in E_0} \in I_0\bigg) \geq 1 - \frac{1}{n^4}. \tag{S2.2}$$

Furthermore, when $K = c\log n$, $|P_e - Q_e| \leq \sqrt{2c_1(P_e+Q_e)\log n/n}$ and $2^{-(j+1)} \leq P_e + Q_e \leq 2^{-(j-2)}$ for any $e \in E_j$, then

$$\mathbb{P}\bigg(\{P_e - Q_e\}_{e\in E_j} \in I_j\bigg) \geq 1 - \frac{1}{n^4} \tag{S2.3}$$

for $j = 1, \ldots, J$.

Proof. Before we show (S2.2), we first show that

$$\mathbb{P}\left(\left|\sum_{e\in E_0} L_e(P_e^k - H_k(\hat{P}_{e,1}))\right| \leq dM\sqrt{2.5n\log^2 n}\left(\frac{76c_1\log n}{n}\right)^k\right) \geq 1 - \frac{2}{n^5}. \tag{S2.4}$$

To the end, we define $\hat{P}'_{e,1} = \min(\hat{P}_{e,1}, 38c_1\log n/n)$ and event $B_t := \Big\{\hat{P}'_{e,1} = \hat{P}_{e,1}, \quad \forall\ e \in E_0\Big\}$. As $P_e \leq 2c_1\log n/n$ and $n \geq |E|/\log|E|$, applying Lemma 8 yields that

$$\mathbb{P}(B_t) \geq 1 - \sum_e \mathbb{P}\left(\hat{P}_{e,1} > \frac{38c_1\log n}{n}\right) \geq 1 - \frac{1}{n^5}.$$

In order to show that $\sum_{e\in E_0} L_e H_k(\hat{P}'_{e,1})$ is a difference bounded function with respect to each $\hat{p}_v$, we apply Lemma 30 in Han et al. (2018). More specifically, Lemma 30 in Han et al. (2018) suggests

$$0 \leq |H_k(x)| \leq \left(2\max\left\{x, \sqrt{\frac{4xk}{n}}\right\}\right)^k.$$

Because perturbing any $\hat{p}_v$ only results in change at most d terms in $\sum_{e\in E_0} L_e H_k(\hat{P}'_{e,1})$, we can know that

$$\left|\sum_{e\in E_0} L_e H_k(\hat{P}'_{e,1}) - \sum_{e\in E_0} L_e H_k(\hat{P}''_{e,1})\right| \le dM \left(\frac{76c_1 \log n}{n}\right)^k,$$

where $\hat{P}''_{e,1}$ is just replace some $\hat{p}_v$ by $\hat{p}'_v$. After showing difference bounded function, we are now ready to apply McDiarmid inequality (see, e.g. Boucheron et al., 2013)

$$\mathbb{P}\left(\sum_{e\in E_0} L_e \left(H_k(\hat{P}'_{e,1}) - \mathbb{E}(H_k(\hat{P}'_{e,1}))\right) > t\right) \le \exp\left(\frac{-2t^2}{d^2M^2|E|(76c_1\log n/n)^{2k}}\right).$$

With the similar arguments in proof of Lemma 18 in Han et al. (2018), we have

$$\left|\mathbb{E}(H_k(\hat{P}'_{e,1})) - P_e^k\right| \le \frac{C}{n^5}\left(\frac{76c_1 \log n}{n}\right)^k.$$

Since $\mathbb{P}(B_t) \ge 1 - n^{-5}$ and $|E| \le n\log n$, we can conclude that

$$\mathbb{P}\left(\sum_{e\in E_0} L_e \left(H_k(\hat{P}_{e,1}) - P_e^k\right) > dM\sqrt{2.5n\log^2 n}\left(\frac{76c_1 \log n}{n}\right)^k\right) \ge 1 - \frac{2}{n^5}.$$

Applying the exact same argument to Q_e yields

$$\mathbb{P}\left(\sum_{e\in E_0} L_e \left(H_k(\hat{Q}_{e,1}) - Q_e^k\right) > dM\sqrt{2.5n\log^2 n}\left(\frac{76c_1 \log n}{n}\right)^k\right) \ge 1 - \frac{2}{n^5}.$$

On the event B_t, we have

$$\max_{e\in E_0} P_e^{k_1} \le \left(\frac{2c_1\log n}{n}\right)^{k_1} \quad \text{and} \quad \max_{e\in E_0} H_{k_2}(\hat{Q}_{e,1}) \le \left(\frac{76c_1 \log n}{n}\right)^{k_2},$$

for $k_1, k_2 = 0, \ldots, K$. Therefore, with probability at least $1 - 4n^{-5}$,

$$\begin{aligned}
&\left|\sum_{e\in E_0} L_e(P_e^{k_1}Q_e^{k_2} - H_{k_1}(\hat{P}_{e,1})H_{k_2}(\hat{Q}_{e,1}))\right| \\
\le&\left(\max_{e\in E_0} P_e^{k_1}\right)\left|\sum_{e\in E_0} L_e(Q_e^{k_2} - H_{k_2}(\hat{Q}_{e,1}))\right| + \left(\max_{e\in E_0} H_{k_2}(\hat{Q}_{e,1})\right)\left|\sum_{e\in E_0} L_e(P_e^{k_1} - H_{k_1}(\hat{P}_{e,1}))\right| \\
\le&2dM\sqrt{2.5n\log^2 n}\left(\frac{76c_1\log n}{n}\right)^{k_1+k_2}.
\end{aligned}$$

Then, we complete proof of (S2.2).

Next, we aim to show (S2.3). We consider the event

$$B_j = \left\{ |\hat{P}_{e,1} - P_e| \le \sqrt{\frac{5(P_e+Q_e)\log n}{n}} \quad \text{and} \quad |\hat{Q}_{e,1} - Q_e| \le \sqrt{\frac{5(P_e+Q_e)\log n}{n}},\ e \in E_j \right\}$$

Based on concentration inequality in Lemma 8, we have

$$\mathbb{P}(B_j) \ge 1 - \frac{2}{n^5}.$$

Hereafter, we conduct analysis conditioned on event B_j. Recall that

$$|P_e - Q_e| \le \sqrt{\frac{2c_1(P_e+Q_e)\log n}{n}}.$$

Conditioned on event B_j, we know that

$$|\hat{P}_{e,1} - \hat{Q}_{e,1}| \le \sqrt{\frac{3c_1(P_e+Q_e)\log n}{n}}.$$

Together with lemma 3, this suggests that

$$|G_k(\hat{P}_{e,1}, \hat{Q}_{e,1})| \le \left(\frac{12c_1(P_e+Q_e)\log n}{n}\right)^{k/2}.$$

As $P_e + Q_e \le 2^{-(j-2)}$, this naturally leads to

$$\left| G_k(\hat{P}_{e,1}, \hat{Q}_{e,1}) - (P_e - Q_e)^k \right| \le 2\left(\frac{48c_1 \log n}{2^j n}\right)^{k/2}.$$

Proposition 2 suggests that E_j can be decomposed S_j subset of disjoint path. More concretely, let $E_{j,1}, \ldots, E_{j,S_j}$ be these subsets of paths. Since each $E_{j,i}$ is a subset of a path, we can know that $|E_{j,i}| \le d$ and, on event B_j,

$$\left| \sum_{e \in E_{j,i}} L_e \left(G_k(\hat{P}_{e,1}, \hat{Q}_{e,1}) - (P_e - Q_e)^k \right) \right| \le 2dM \left(\frac{48c_1 \log n}{2^j n}\right)^{k/2}, \qquad 1 \le i \le S_j.$$

Thus, for each $1 \le i \le S_j$, we define a random variable

$$Z_i = \sum_{e \in E_{j,i}} L_e \left(G_k(\hat{P}_{e,1}, \hat{Q}_{e,1}) - (P_e - Q_e)^k \right)$$

and its truncated version

$$\tilde{Z}_i = \begin{cases} T & Z_i > T \\ Z_i & -T \le Z_i \le T \\ -T & Z_i < -T \end{cases},$$

where $T = 2dM\left(48c_1 \log n/2^j n\right)^{k/2}$. As suggested by proposition 2, $\hat{P}_{e_1,1}$ and $\hat{P}_{e_2,1}$ are independent for any two edges e_1 and e_2 coming from different $E_{j,i}$. Thus, we can know that $(Z_i, \tilde{Z}_i)$ are independent for different i. An application of Hoeffding inequality yields

$$\mathbb{P}\left(\left|\sum_{i=1}^{S_j}\left(\tilde{Z}_i - \mathbb{E}(\tilde{Z}_i)\right)\right| > t\right) \leq \exp\left(\frac{-t^2}{2S_j T^2}\right).$$

Now, we would like to show that $|\mathbb{E}(Z_i - \tilde{Z}_i)|$ is small following the similar arguments in proof of Lemma 18 in Han et al. (2018). Clearly,

$$\begin{aligned}|\mathbb{E}(Z_i - \tilde{Z}_i)| &\leq \mathbb{E}\left(|Z_i - \tilde{Z}_i|\mathbb{I}_{(|\hat{P}_{e,1}-P_e|,|\hat{Q}_{e,1}-Q_e|>\sqrt{5(P_e+Q_e)\log n/n}, e\in E_{j,i})}\right)\\ &\leq M\sum_{e\in E_{j,i}}\mathbb{E}\left(|G_k(\hat{P}_{e,1},\hat{Q}_{e,1})|\mathbb{I}_{(|\hat{P}_{e,1}-P_e|,|\hat{Q}_{e,1}-Q_e|>\sqrt{5(P_e+Q_e)\log n/n})}\right).\end{aligned}$$

We now bound $\mathbb{E}\left(|G_k(\hat{P}_{e,1},\hat{Q}_{e,1})|\mathbb{I}_{(|\hat{P}_{e,1}-P_e|,|\hat{Q}_{e,1}-Q_e|>\sqrt{5(P_e+Q_e)\log n/n})}\right)$ for different cases. If we write $\Delta_j = \sqrt{48c_1 \log n/2^j n}$, then

$$\begin{aligned}&\mathbb{E}\left(|G_k(\hat{P}_{e,1},\hat{Q}_{e,1})|\mathbb{I}_{(\hat{P}_{e,1}-P_e,\hat{Q}_{e,1}-Q_e>\sqrt{5(P_e+Q_e)\log n/n})}\right)\\ \leq &\sum_{m_p-nP_e>\sqrt{5nP_e\log n}}\ \sum_{m_q-nQ_e>\sqrt{5nQ_e\log n}}\left(\frac{2|m_p-m_q|}{n}\right)^k \mathbb{P}\left(\mathrm{Pois}(nP_e)=m_p\right)\mathbb{P}\left(\mathrm{Pois}(nQ_e)=m_q\right)\\ \leq & n^{-10}\sum_{l_p,l_q=0}^{\infty}\left(\Delta_j + \frac{l_p+l_q}{n}\right)^k\left(1-\sqrt{\frac{5\log n}{nP_e}}\right)^{l_1+l_2}\\ \leq & n^{-10}\Delta_j^k\left(1-\exp\left(-\sqrt{\frac{5\log n}{nP_e}}+\frac{k}{n\Delta_j}\right)\right)^{-2}\\ \leq & n^{-10}\Delta_j^k\left(1-\exp\left(-\sqrt{\frac{2^j\log n}{2n}}\right)\right)^{-2}\\ \leq & n^{-10}\Delta_j^k.\end{aligned}$$

The other three cases including $\hat{P}_{e,1}-P_e < \sqrt{5(P_e+Q_e)\log n/n}$ or $\hat{Q}_{e,1}-Q_e < \sqrt{5(P_e+Q_e)\log n/n}$ can be treated similarly. Thus, we can conclude that

$$|\mathbb{E}(Z_i - \tilde{Z}_i)| \leq 4dMn^{-10}\left(\frac{48c_1\log n}{2^j n}\right)^{k/2}.$$

Since $Z_i = \tilde{Z}_i$ on the event B_j, we can have

$$\mathbb{P}\left(\left|\sum_{i=1}^{S_j} Z_i\right| \leq 2dM\sqrt{10S_j\log n}\left(\frac{48c_1\log n}{2^j n}\right)^{k/2}\right) \geq 1-\frac{1}{n^4}.$$

The proof of (S2.3) is complete. $\square$

Lemma 2. *If we define the event B as in proof of Theorem 1, then*

$$\mathbb{P}(B^c) \leq 5|E|n^{-c_1/10}$$

Proof. We follow the similar strategy in Lemma 4 of Jiao et al. (2018). Because

$$\mathbb{P}(B^c) \leq \sum_{j=0}^{J} \mathbb{P}(B_j^c) + \mathbb{P}(B'^c), \tag{S2.5}$$

we bound above terms separately. For B_0, we have

$$\mathbb{P}(B_0^c) \leq \sum_e \mathbb{P}\left(P_e + Q_e \geq \frac{2c_1 \log n}{n}, \hat{P}_{e,0} + \hat{Q}_{e,0} < \frac{c_1 \log n}{n}\right).$$

Since $n(\hat{P}_{e,0} + \hat{Q}_{e,0})$ follows a Poisson distribution with mean $n(P_e + Q_e)$, we apply concentration inequality in Lemma 8 and obtain

$$\mathbb{P}\left(P_e + Q_e \geq \frac{2c_1 \log n}{n}, \hat{P}_{e,0} + \hat{Q}_{e,0} < \frac{c_1 \log n}{n}\right) \leq n^{-c_1/4}.$$

An application of union bound suggests that

$$\mathbb{P}(B_0^c) \leq |E|n^{-c_1/4}.$$

For each B_j, we have

$$\begin{aligned}\mathbb{P}(B_j^c) \leq & \sum_e \mathbb{P}\left(|P_e - Q_e| > \sqrt{\frac{2c_1(P_e + Q_e)\log n}{n}}, \left|\hat{P}_{e,0} - \hat{Q}_{e,0}\right| \leq \sqrt{\frac{1.1c_1 \log n}{n}}\left(\sqrt{\hat{P}_{e,0} + \hat{Q}_{e,0}}\right)\right) \\ & + \sum_e \mathbb{P}\left(P_e + Q_e \geq \frac{1}{2^{j-2}}, \hat{P}_{e,0} + \hat{Q}_{e,0} < \frac{1}{2^{j-1}}\right) + \sum_e \mathbb{P}\left(P_e + Q_e \leq \frac{1}{2^{j+1}}, \hat{P}_{e,0} + \hat{Q}_{e,0} > \frac{1}{2^j}\right).\end{aligned}$$

Following the similar arguments in proof of Lemma 4 in Jiao et al. (2018), we obtain the bound for the first term

$$\mathbb{P}\left(|P_e - Q_e| > \sqrt{\frac{2c_1(P_e + Q_e)\log n}{n}}, \left|\hat{P}_{e,0} - \hat{Q}_{e,0}\right| \leq \sqrt{\frac{1.1c_1 \log n}{n}}\left(\sqrt{\hat{P}_{e,0} + \hat{Q}_{e,0}}\right)\right) \leq 4n^{-c_1/3}.$$

As $n(\hat{P}_{e,0} + \hat{Q}_{e,0})$ follows a Poisson distribution, we have

$$\mathbb{P}\left(P_e + Q_e \geq \frac{1}{2^{j-2}}, \hat{P}_{e,0} + \hat{Q}_{e,0} < \frac{1}{2^{j-1}}\right) \leq n^{-c_1/4}$$

and

$$\mathbb{P}\left(P_e + Q_e \le \frac{1}{2^{j+1}}, \hat{P}_{e,0} + \hat{Q}_{e,0} > \frac{1}{2^j}\right) \le n^{-c_1/4}.$$

Putting all these terms together yields

$$\mathbb{P}(B_j^c) \le 6|E|n^{-c_1/4}.$$

Finally, we work on the last term

$$\begin{aligned}
\mathbb{P}(B'^c) &\le \sum_e 2\mathbb{P}\left(P_e = Q_e, \hat{P}_{e,0} - \hat{Q}_{e,0} > \sqrt{\frac{1.1c_1(\hat{P}_{e,0} + \hat{Q}_{e,0})\log n}{n}}\right) \\
&\le 2\sum_e \mathbb{P}\left(P_e = Q_e, \sqrt{\hat{P}_{e,0}} - \sqrt{\hat{Q}_{e,0}} > \sqrt{\frac{1.1c_1 \log n}{2n}}\right) \\
&\le 4\sum_e \mathbb{P}\left(\left|\hat{P}_{e,0} - P_e\right| > \sqrt{\frac{1.1c_1 P_e \log n}{4n}}\right) \\
&\le \frac{4|E|}{n^{c_1/10}}.
\end{aligned}$$

Putting all above terms back into (S2.5), we have

$$\mathbb{P}(B^c) \le 5|E|n^{-c_1/10}.$$

□

Lemma 3.

$$|G_k(P,Q)| \le \left(2|P-Q| + \sqrt{\frac{4k}{n}}\left(\sqrt{P} + \sqrt{Q}\right)\right)^k$$

Proof. We define

$$G_{k,Q}(P) = \sum_{m=0}^{k} \binom{k}{m}(-Q)^{k-m}\prod_{m'=0}^{m-1}\left(P - \frac{m'}{n}\right).$$

As proof of Lemma 19 in Jiao et al. (2018) suggests

$$G_k(P,Q) = \sum_{m=0}^{k}\binom{k}{m} G_{m,(P+Q)/2}(P)(-1)^{k-m}G_{k-m,(P+Q)/2}(Q).$$

Lemma 30 in Han et al. (2018) implies

$$|G_{m,(P+Q)/2}(P)| \le \left(|P-Q| + \sqrt{\frac{4mP}{n}}\right)^m$$

and

$$|G_{k-m,(P+Q)/2}(Q)| \leq \left(|P-Q| + \sqrt{\frac{4(k-m)Q}{n}}\right)^{k-m}.$$

Thus, we could know that

$$\begin{aligned}
|G_k(P,Q)| &\leq \sum_{m=0}^{k} \binom{k}{m} |G_{m,(P+Q)/2}(P)||G_{k-m,(P+Q)/2}(Q)| \\
&\leq \sum_{m=0}^{k} \binom{k}{m} \left(|P-Q| + \sqrt{\frac{4mP}{n}}\right)^{m} \left(|P-Q| + \sqrt{\frac{4(k-m)Q}{n}}\right)^{k-m} \\
&\leq \sum_{m=0}^{k} \binom{k}{m} \left(|P-Q| + \sqrt{\frac{4kP}{n}}\right)^{m} \left(|P-Q| + \sqrt{\frac{4kQ}{n}}\right)^{k-m} \\
&\leq \left(2|P-Q| + \sqrt{\frac{4k}{n}}\left(\sqrt{P}+\sqrt{Q}\right)\right)^{k}
\end{aligned}$$

□

Lemma 4. *For any pair of edges on tree $e_1, e_2 \in E$, $\tau(e_1)$ and $\tau(e_2)$ satisfy one and only one of following relationships*

- $\tau(e_1) \cap \tau(e_2) = \emptyset$;
- $\tau(e_1) \subset \tau(e_2)$;
- $\tau(e_2) \subset \tau(e_1)$.

Proof. If $e_1 \in [\rho, v]$ for all $v \in \tau(e_2)$, then we can know that $\tau(e_2) \subset \tau(e_1)$. Similarly, $\tau(e_1) \subset \tau(e_2)$ if $e_2 \in [\rho, v]$ for all $v \in \tau(e_1)$. Supposing there exists $v_1 \in \tau(e_1)$ such that $e_2 \notin [\rho, v_1]$ and $v_2 \in \tau(e_2)$ such that $e_1 \notin [\rho, v_2]$, we can conclude that $\tau(e_1) \cap \tau(e_2) = \emptyset$. Otherwise, let $v' \in \tau(e_1) \cap \tau(e_2)$. Then, there are two paths connecting v_1 and v_2: one is through ρ and the other is through v'. This contradicts with the fact there is one and only one path connect a pair of nodes on tree. □

Lemma 5. *For any s, d, n,*

$$R^*_{n/2}(s,d,Q) \geq \frac{1}{2}\tilde{R}^*_n(s,d,Q,\epsilon) - d^2 e^{-n/8} - d^2\epsilon^2.$$

Proof. Given $\delta > 0$, suppose $\hat{D}$ is an estimator such that

$$\sup_{(T,P,Q)\in\Theta(s,d,Q)} \mathbb{E}(\hat{D} - D(P,Q)) \leq \delta + R_n^*(s,d,Q).$$

Here, the sample are drawn from multinomial distribution with sample size n.

Fixing $P \in \mathcal{M}_{|V|}(\epsilon)$, the sample $X = (X_v)_{v\in V}$ are drawn from $\{\mathrm{Pois}(np_v)\}_{v\in V}$. Here, $n' = \sum_v X_v$ is "sample size". Since X can be seen as a sample drawn from multinomial conditioned on n', X can also be regard as input of $\hat{D}$. Let $\tilde{P} = \{p_v / \sum_v p_v\}_{v\in V}$. So, we have

$$\mathbb{E}_P(\hat{D} - D(P,Q))^2 \leq 2\mathbb{E}_P(\hat{D} - D(\tilde{P},Q))^2 + 2(D(P,Q) - D(\tilde{P},Q))^2.$$

Note

$$\begin{aligned}
D(P,Q) - D(\tilde{P},Q) &\leq \sum_{e\in E} L_e \left(|\tilde{P}_e - Q_e| - |P_e - Q_e|\right) \\
&\leq M \sum_{e\in E} |\tilde{P}_e - P_e| \\
&\leq M \sum_{e\in E} \frac{P_e}{\sum_v p_v}\epsilon \\
&\leq dM\epsilon,
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_P(\hat{D} - D(\tilde{P},Q))^2 &\leq \sum_m R_m^*(s,d,Q)\mathbb{P}(n' = m) + \delta \\
&\leq R_{n/2}^*(s,d,Q) + d^2\mathbb{P}(n' < n/2) + \delta \\
&\leq R_{n/2}^*(s,d,Q) + d^2 e^{-n/8} + \delta.
\end{aligned}$$

Since δ can be arbitrarily small,

$$\mathbb{E}_P(\hat{D} - D(P,Q))^2 \leq 2R_{n/2}^*(s,d,Q) + 2d^2 e^{-n/8} + 2d^2M^2\epsilon^2.$$

This immediately suggests that

$$\tilde{R}_n^*(s,d,Q,\epsilon) \leq 2R_{n/2}^*(s,d,Q) + 2d^2 e^{-n/8} + 2d^2M^2\epsilon^2.$$

□

Lemma 6 (Tsybakov (2009))**.** *Suppose* π_i $i = 1, 2$ *are two prior distributions on parameter space* Θ*,* $F(\theta)$ *is a functional on parameter space and* $\mathbb{P}_i$*s are corresponding marginal distributions of observed data. Suppose there exists* $c \in R$*,* $\delta > 0$*,* $0 \leq \beta_1, \beta_2 < 1$ *such that*

$$\pi_1(F(\theta) \leq c) \geq 1 - \beta_1 \qquad \text{and} \qquad \pi_2(F(\theta) \geq c + 2\delta) \geq 1 - \beta_2.$$

If $TV(\mathbb{P}_1, \mathbb{P}_2) \leq \eta < 1$*, then*

$$\inf_{\hat{F}} \sup_{\theta \in \Theta} \mathbb{P}_\theta(|\hat{F} - F(\theta)| \geq \delta) \geq \frac{1 - \eta - \beta_0 - \beta_1}{2}.$$

Here $TV(\mathbb{P}_1, \mathbb{P}_2)$ *is total variation distance defined as*

$$TV(\mathbb{P}_1, \mathbb{P}_2) = \sup_A |\mathbb{P}_1(A) - \mathbb{P}_2(A)|.$$

Lemma 7 (Jiao et al. (2018))**.** *Suppose* U_1 *and* U_2 *are two random variables supported on* $[nq - n\lambda, nq + n\lambda]$*, where* $q \geq \lambda \geq 0$*. Suppose* $\mathbb{E}(U_1^j) = \mathbb{E}(U_2^j)$*,* $0 \leq j \leq L$*. Denote the marginal distribution of* X *where* $X|\eta \sim \mathrm{Pois}(\eta)$*,* $\eta \sim U_i$ *as* F_i*. If* $L + 1 \geq (2e\lambda)^/ q$*, then*

$$TV(F_1, F_2) \leq 2 \left(\frac{e\lambda\sqrt{n}}{\sqrt{q(L+1)}} \right)^{L+1}.$$

Lemma 8 (Mitzenmacher and Upfal (2005))**.** *If* $X \sim \mathrm{Pois}(\lambda)$*, the for any* $\delta > 0$*, we have*

$$\mathbb{P}(X \geq (1 + \delta)\lambda) \leq \exp\left(-\frac{(\delta^2 \wedge \delta)\lambda}{3} \right)$$

and

$$\mathbb{P}(X \leq (1 - \delta)\lambda) \leq \exp\left(-\frac{\delta^2\lambda}{2} \right).$$

Lemma 9 (Jiao et al. (2018))**.** *Suppose* $n\hat{p} \sim \mathrm{Poi}(np)$*. Then,*

$$\frac{1}{\sqrt{2}} \left(p \wedge \sqrt{\frac{p}{n}} \right) \leq \mathbb{E}|\hat{p} - p| \leq 2 \left(p \wedge \sqrt{\frac{p}{n}} \right)$$

and

$$\mathrm{Var}\left(|\hat{p} - p|\right) \leq \frac{p}{n}$$

Lemma 10. *If* $x, y > 0$ *and* $0 < \alpha < 1$*, then*

$$|x^\alpha - y^\alpha| < |x - y|^\alpha$$

Proof. Without loss of generality, we assume $x > y$ and write $z = x - y > 0$.It is sufficient to show

$$(y+z)^\alpha < z^\alpha + y^\alpha.$$

We can assume $y > z$ for simplicity as y and z are exchangeable in above inequality. Thus,

$$(y+z)^\alpha - y^\alpha = \alpha\xi^{\alpha-1}z \leq \alpha z^\alpha < z^\alpha.$$

Here ξ is some number between z and y. $\square$

Lemma 11. *Suppose $n\hat{p}_i \sim \mathrm{Poi}(np_i)$, $i = 1,2$ and $\hat{p}_1$ and $\hat{p}_2$ are independent. Then,*

$$\mathrm{Cov}\left(|\hat{p}_1 + \hat{p}_2 - p_1 - p_2|, |\hat{p}_1 - p_1|\right) \leq \frac{p_1}{n}.$$

Proof. Write $A = \hat{p}_1 - p_1$ and $B = \hat{p}_2 - p_2$. Then

$$\begin{aligned}
\mathrm{Cov}\left(|A+B|, |A|\right) &= \mathbb{E}(|A^2 + AB|) - \mathbb{E}|A+B|\mathbb{E}|A| \\
&\leq \mathbb{E}\left(|B| - |A+B|\right)\mathbb{E}|A| + \mathbb{E}(A^2) \\
&\leq \mathbb{E}(A^2),
\end{aligned}$$

where we use the fact that $\mathbb{E}|B| < \mathbb{E}|A+B|$, see the exact analytic expression of $\mathbb{E}(X-\lambda)$ for Poisson random variable $X \sim \mathrm{Poi}(\lambda)$ in Diaconis and Zabell (1991). Property of poisson distribution suggests that

$$\mathbb{E}(A^2) = \mathbb{E}(\hat{p}_1 - p_1)^2 \leq \frac{p_1}{n}.$$

$\square$

Lemma 12. *Suppose the branch length $L_e = 1$. For any tree T with height d and $0 < \alpha < 1$, we have*

$$\sup_{P\in\mathcal{M}_s}\sum_{e\in E} P_e^\alpha \leq Cs^{1-\alpha}\log^\alpha(2^{d+2}/s)$$

for some constant C. Furthermore, there exists some $T \in \mathcal{T}(s,d)$ such that

$$\sup_{P\in\mathcal{M}_s}\sum_{e\in E} P_e^\alpha \geq cs^{1-\alpha}\log^\alpha(2^{d+2}/s)$$

Proof. We define $U_i = \{e : d(\rho, v) \geq i, \forall e \in [\rho, v]\}$, $i = 1, \ldots, d$. Because the tree is binary tree, $|U_i| \leq 2^i$. By the definition of U_i, we know that $\tau(e_1) \cap \tau(e_2) = \emptyset$ if $e_1, e_2 \in U_i$. Since $\sum_{e \in U_i} P_e \leq 1$ and $|U_i| \leq 2^i$, Holder inequality yields

$$\begin{aligned}
\sum_{e \in E} P_e^\alpha &\leq \sum_{i=1}^{d} \left(\sum_{e \in U_i} P_e^\alpha \right) \\
&\leq \sum_{i=1}^{d} \left(\sum_{e \in U_i} (P_e^\alpha)^{1/\alpha} \right)^\alpha \left(\sum_{e \in U_i} 1^{1/(1-\alpha)} \right)^{1-\alpha} \\
&\leq \sum_{i=1}^{d} |U_i|^{1-\alpha} \\
&\leq \sum_{i \geq 1} 2^{i\beta} |U_i|^{1-\alpha-\beta}
\end{aligned}$$

for some $0 < \beta < 1 - \alpha$ which is specified later. By Holder's inequality again,

$$\begin{aligned}
\sum_{1 \leq i \leq d} 2^{i\beta} |U_i|^{1-\alpha-\beta} &\leq \left(\sum_{1 \leq i \leq d} 2^{i\beta/(\alpha+\beta)} \right)^{\alpha+\beta} \left(\sum_{1 \leq i \leq d} |U_i| \right)^{1-\alpha-\beta} \\
&\leq \left(\frac{2 - 2^{\beta d/(\alpha+\beta)}}{1 - 2^{\beta/(\alpha+\beta)}} \right)^{\alpha+\beta} s^{1-\alpha-\beta} \\
&\leq \left(\frac{1}{2^{\beta/(\alpha+\beta)} - 1} \right)^{\alpha+\beta} \left(\frac{2^d}{s} \right)^\beta s^{1-\alpha}
\end{aligned}$$

Choosing $\beta = \alpha / \log(2^d/s)$ yields

$$\left(\frac{1}{2^{\beta/(\alpha+\beta)} - 1} \right)^{\alpha+\beta} \left(\frac{2^d}{s} \right)^\beta \leq C \log^\alpha(2^{d+2}/s).$$

Thus, we can conclude that

$$\sum_{e \in E} P_e^\alpha \leq C s^{1-\alpha} \log^\alpha(2^{d+2}/s).$$

We now prove the converse side. Suppose T is $T_0(k_1, k_2)$ in lower bound proof and k_1, k_2 are chosen in the same way. When we put uniform probability in V_0, then we complete proof. $\square$

Lemma 13. *Suppose* $n\hat{P} \sim \mathrm{Pois}(nP)$ *and* $n\hat{Q} \sim \mathrm{Pois}(nQ)$. *Assume* $|P - Q| > \sqrt{c(P+Q)\log n/n}$ *and* $P + Q > c \log n/n$. *Then, for any* $0 < \alpha < 2$,

$$\left| \mathbb{E}\left(U_\alpha(\hat{P}, \hat{Q}) \right) - |P - Q|^\alpha \right| \leq C \left(\frac{(P+Q)^{\alpha/2}}{n^{\alpha/2} \log^{2-\alpha/2} n} + \frac{P+Q}{n^{c-4}} \right)$$

for some constant C.

Proof. Write $\Delta = P - Q$, $\Sigma = P + Q$, $\hat{\Delta} = \hat{P} - \hat{Q}$, $\hat{\Sigma} = \hat{P} + \hat{Q}$ and $\hat{I} = I_n(\hat{P}, \hat{Q})$. We only focus the situation $\Delta > 0$ in the rest of proof and other case can be treated similarly. As

$$U_\alpha(\hat{P}, \hat{Q}) = |\hat{\Delta}|^\alpha + \frac{\alpha(1-\alpha)}{2n}|\hat{\Delta}|^{\alpha-2}\hat{\Sigma}\hat{I},$$

we do Taylor expansion for $|\hat{\Delta}|^\alpha$ and $|\hat{\Delta}|^{\alpha-2}$. More concretely, the Taylor expansion of $|\hat{\Delta}|^\alpha$ at Δ can be written as

$$\begin{aligned} T_3(|\hat{\Delta}|^\alpha; \Delta) =& |\Delta|^\alpha + \alpha|\Delta|^{\alpha-1}(\hat{\Delta} - \Delta) + \frac{\alpha(\alpha-1)}{2}\Delta^{\alpha-2}(\hat{\Delta} - \Delta)^2 \\ &+ \frac{\alpha(\alpha-1)(\alpha-2)}{6}\Delta^{\alpha-3}(\hat{\Delta} - \Delta)^3. \end{aligned}$$

Then, the residue of above Taylor expansion is denoted by $R_3(|\hat{\Delta}|^\alpha; \Delta) = |\hat{\Delta}|^\alpha - T_3(|\hat{\Delta}|^\alpha; \Delta)$. We know bound the residue term at different regimes. When $\hat{\Delta} \geq 0$, the residue term can be represented in integral form

$$R_3(|\hat{\Delta}|^\alpha; \Delta) = \frac{1}{6}\int_\Delta^{\hat{\Delta}} C_1(\alpha)(\hat{\Delta} - u)^3 u^{\alpha-4} du,$$

where $C_1(\alpha) = \alpha(\alpha-1)(\alpha-2)(\alpha-3)$. In particular, when $\hat{\Delta} > \Delta/2$, we have

$$|R_3(|\hat{\Delta}|^\alpha; \Delta)| \leq \frac{C_1(\alpha)}{24}\left(\frac{\Delta}{2}\right)^{\alpha-4}(\hat{\Delta} - \Delta)^4.$$

If $0 \leq \hat{\Delta} < \Delta/2$, then

$$\begin{aligned} |R_3(|\hat{\Delta}|^\alpha; \Delta)| &\leq \frac{C_1(\alpha)}{6}\int_\Delta^{\hat{\Delta}} (\hat{\Delta}^3 - 3\hat{\Delta}^2 u + 3\hat{\Delta} u^2 - u^3) u^{\alpha-4} du \\ &\leq \frac{C_1(\alpha)}{6}\left(\frac{3\hat{\Delta}^2}{2-\alpha}(\hat{\Delta}^{\alpha-2} - \Delta^{\alpha-2}) + \frac{1}{\alpha}(\Delta^\alpha - \hat{\Delta}^\alpha)\right) \\ &\leq \frac{C_1(\alpha)(\alpha+1)}{3\alpha(2-\alpha)}\Delta^\alpha. \end{aligned}$$

On the other hand, if $\hat{\Delta} < 0$, we could work on $|\hat{\Delta}|^\alpha - T_3(|\hat{\Delta}|^\alpha; \Delta)$ directly. If $\hat{\Delta} > -\Delta$, then

$$|R_3(|\hat{\Delta}|^\alpha; \Delta)| \leq 10\Delta^\alpha.$$

When $\hat{\Delta} < -\Delta$, we have

$$|R_3(|\hat{\Delta}|^\alpha; \Delta)| \leq 10\Delta^\alpha \left(\frac{\hat{\Delta}}{\Delta}\right)^3.$$

Thus, the expectation of $|R_3(|\hat{\Delta}|^\alpha;\Delta)|$ can be decomposed as

$$\begin{aligned}\mathbb{E}(|R_3(|\hat{\Delta}|^\alpha;\Delta)|) =& \mathbb{E}(|R_3(|\hat{\Delta}|^\alpha;\Delta)|\mathbf{I}(\hat{\Delta} > \Delta/2)) + \mathbb{E}(|R_3(|\hat{\Delta}|^\alpha;\Delta)|\mathbf{I}(\hat{\Delta} < -\Delta)) \\ &+ \mathbb{E}(|R_3(|\hat{\Delta}|^\alpha;\Delta)|\mathbf{I}(-\Delta \le \hat{\Delta} \le \Delta/2)) \\ \le& \frac{C_1(\alpha)}{24}\left(\frac{\Delta}{2}\right)^{\alpha-4}\mathbb{E}(\hat{\Delta}-\Delta)^4 + 10\Delta^{\alpha-3}\mathbb{E}(|\hat{\Delta}|^3\mathbf{I}(\hat{\Delta} < -\Delta)) \\ &+ 10\Delta^\alpha\mathbb{P}(\hat{\Delta} \le \Delta/2) \\ \le& \frac{2C_1(\alpha)}{3\cdot 2^\alpha}\Delta^{\alpha-4}\left(\frac{\Sigma}{n^3}+\frac{3\Sigma^2}{n^2}\right) + 10\Delta^\alpha\frac{(n\Sigma)^{7/2}+1}{n^c}.\end{aligned}$$

Similarly, we can write $|\hat{\Delta}|^{\alpha-2}\hat{\Sigma}\hat{I}$ as

$$|\hat{\Delta}|^{\alpha-2}\hat{\Sigma}\hat{I} = T_1(|\hat{\Delta}|^{\alpha-2};\Delta)\hat{\Sigma}\hat{I} + R_1(|\hat{\Delta}|^{\alpha-2};\Delta)\hat{\Sigma}\hat{I},$$

where $T_1(|\hat{\Delta}|^{\alpha-2};\Delta) = |\Delta|^{\alpha-2} + (\alpha-2)|\Delta|^{\alpha-3}(\hat{\Delta}-\Delta)$ and $R_1(|\hat{\Delta}|^{\alpha-2};\Delta) = |\hat{\Delta}|^{\alpha-2} - T_1(|\hat{\Delta}|^{\alpha-2};\Delta)$. Bounding $R_1(|\hat{\Delta}|^{\alpha-2};\Delta)$ in different regimes, we can have

$$\mathbb{E}(|R_1(|\hat{\Delta}|^{\alpha-2};\Delta)\hat{\Sigma}\hat{I}|) \le 3\left(\frac{\Delta}{2}\right)^{\alpha-4}\left(\frac{\Sigma}{n^2}+\frac{\Sigma^2}{n}\right). + 10\Delta^{\alpha-2}\Sigma\frac{(n\Sigma)^{3/2}+1}{n^c}$$

Putting two Taylor expansion together yields

$$\begin{aligned}\left|\mathbb{E}\left(U_\alpha(\hat{P},\hat{Q})\right) - |P-Q|^\alpha\right| &\le \frac{1}{3n^2}\Delta^{\alpha-2} + \mathbb{E}(|R_3(|\hat{\Delta}|^\alpha;\Delta)|) + \frac{1}{8n}\mathbb{E}(|R_1(|\hat{\Delta}|^{\alpha-2};\Delta)\hat{\Sigma}\hat{I}|) \\ &\le \frac{1}{3n^2}\Delta^{\alpha-2} + \frac{12}{2^\alpha}\Delta^{\alpha-4}\left(\frac{2\Sigma}{n^3}+\frac{4\Sigma^2}{n^2}\right) + 22\Delta^\alpha\frac{(n\Sigma)^{7/2}}{n^c} \\ &\le C\frac{\Sigma^{\alpha/2}}{n^{\alpha/2}\log^{2-\alpha/2}n} + \frac{\Sigma}{n^{c-4}}\end{aligned}$$

Here, we use $|\Delta| > \sqrt{c\log n\Sigma/n}$ and $\Sigma > c\log n/n$. $\square$

Lemma 14. *Suppose $n\hat{P} \sim \mathrm{Pois}(nP)$ and $n\hat{Q} \sim \mathrm{Pois}(nQ)$. If $0 < P, Q < 1$ and $\alpha \ge 2$, then*

$$\left|\mathbb{E}\left(|\hat{P}-\hat{Q}|^\alpha\right) - |P-Q|^\alpha\right| \le C\frac{P+Q}{n}$$

for some constant C.

Proof. If $\alpha = 2$, then we can directly calculate

$$\left|\mathbb{E}\left(|\hat{P}-\hat{Q}|^2\right) - |P-Q|^2\right| = \frac{P+Q}{n}.$$

When $\alpha > 2$, $|x|^\alpha$ is twice differential continuous function on $[-1, 1]$. Thus, we can have Taylor expansion

$$|y|^\alpha = |x|^\alpha + \alpha|x|^{\alpha-1}(y-x) + \frac{\alpha(\alpha-1)|tx+(1-t)y|^{\alpha-2}}{2}(y-x)^2$$

for some $t \in (0, 1)$. This suggests that

$$\left|\mathbb{E}\left(|\hat{P}-\hat{Q}|^\alpha\right) - |P-Q|^\alpha\right| \le \frac{\alpha(\alpha-1)}{2}\left(\mathbb{E}(\hat{P}-P)^2 + \mathbb{E}(\hat{Q}-Q)^2\right) \le C\frac{P+Q}{n}.$$

We now complete proof. □

Lemma 15. *Suppose $n\hat{P}_i \sim \mathrm{Pois}(nP_i)$ and $n\hat{Q}_i \sim \mathrm{Pois}(nQ_i)$ for $i = 1, 2$. We also assume $|P_i - Q_i| > \sqrt{c(P_i+Q_i)\log n/n}$ and $P_i + Q_i > c\log n/n$. Then, for any $0 < \alpha < 2$,*

$$\mathrm{Cov}\left(U_\alpha(\hat{P}_1, \hat{Q}_1), U_\alpha(\hat{P}_1+\hat{P}_2, \hat{Q}_1+\hat{Q}_2)\right) \le C|\Delta_1\Delta_2|^{\alpha-1}\left(\frac{\Sigma_1}{n} + \frac{\sqrt{\Sigma_1\Sigma_2}}{n\log^2 n} + \frac{1}{n^{c/2-4}}\right),$$

where $\Sigma_i = \sum_{j=1}^i (P_j + Q_j)$, $\Delta_i = \sum_{j=1}^i (P_j - Q_j)$ and C is some constant. In particular,

$$\mathrm{Var}\left(U_\alpha(\hat{P}_1, \hat{Q}_1)\right) \le C|\Delta_1|^{2\alpha-2}\left(\frac{\Sigma_1}{n} + \frac{1}{n^{c/2-4}}\right).$$

Proof. We write $U_{\alpha,i} = U_\alpha(\sum_{j=1}^i \hat{P}_j, \sum_{j=1}^i \hat{Q}_i)$, $\hat{\Delta}_i = \sum_{j=1}^i(\hat{P}_j - \hat{Q}_j)$, $\hat{\Sigma}_i = \sum_{j=1}^i(\hat{P}_j + \hat{Q}_j)$ and $\hat{I}_i = I_n(\sum_{j=1}^i \hat{P}_j, \sum_{j=1}^i \hat{Q}_i)$ for $i = 1, 2$. In particular, we only focus on the cases $\Delta_i > 0$ for $i = 1, 2$. We represent $U_{\alpha,i}$ in Taylor expansion

$$U_{\alpha,i} = |\Delta_i|^\alpha + \alpha|\Delta_i|^{\alpha-1}(\hat{\Delta}_i - \Delta_i) + R_1(|\hat{\Delta}_i|^\alpha; \Delta_i) + \frac{\alpha(1-\alpha)}{2n}|\hat{\Delta}_i|^{\alpha-2}\hat{\Sigma}_i\hat{I}_i$$

where $R_1(|\hat{\Delta}_i|^\alpha; \Delta_i) = |\hat{\Delta}_i|^\alpha - [|\Delta_i|^\alpha + \alpha|\Delta_i|^{\alpha-1}(\hat{\Delta}_i - \Delta_i)]$. If we write

$$R_{1,i} = R_1(|\hat{\Delta}_i|^\alpha; \Delta_i) + \frac{\alpha(1-\alpha)}{2n}|\hat{\Delta}_i|^{\alpha-2}\hat{\Sigma}_i\hat{I}_i,$$

then the covariance between $U_{\alpha,1}$ and $U_{\alpha,2}$ can be decomposed as

$$\begin{aligned}\mathrm{Cov}(U_{\alpha,1}, U_{\alpha,2}) =& \alpha\left(|\Delta_1|^{\alpha-1}\mathrm{Cov}(\hat{\Delta}_1, R_{1,2}) + |\Delta_2|^{\alpha-1}\mathrm{Cov}(\hat{\Delta}_2, R_{1,1})\right) \\ &+ \alpha^2|\Delta_1\Delta_2|^{\alpha-1}\mathrm{Var}(\hat{\Delta}_1) + \mathrm{Cov}(R_{1,1}, R_{1,2}).\end{aligned}$$

We now bound above terms one by one. Firstly, we work on $\mathrm{Cov}(\hat{\Delta}_1, R_{1,2})$. We rewrite $R_{1,2}$ as

$$R_{1,2} = \frac{\alpha(\alpha-1)}{2}|\Delta_2|^{\alpha-2}\left((\hat{\Delta}_2 - \Delta_2)^2 - \frac{\hat{\Sigma}_2\hat{I}_2}{n}\right) + R_2(|\hat{\Delta}_2|^\alpha; \Delta_2) + \frac{\alpha(1-\alpha)}{2n}R_0(|\hat{\Delta}_2|^{\alpha-2}; \Delta_2)\hat{\Sigma}_2\hat{I}_2.$$

Thus, we have

$$
\begin{aligned}
\mathrm{Cov}(\hat{\Delta}_1, R_{1,2}) =& \frac{\alpha(\alpha-1)}{2}|\Delta_2|^{\alpha-2}\mathrm{Cov}\left(\hat{\Delta}_1, (\hat{\Delta}_2-\Delta_2)^2 - \frac{\hat{\Sigma}_2\hat{I}_2}{n}\right) \\
&+ \mathrm{Cov}\left(\hat{\Delta}_1, R_2(|\hat{\Delta}_2|^{\alpha};\Delta_2)\right) + \frac{\alpha(1-\alpha)}{2n}\mathrm{Cov}\left(\hat{\Delta}_1, R_0(|\hat{\Delta}_2|^{\alpha-2};\Delta_2)\hat{\Sigma}_2\hat{I}_2\right) \\
\leq& \mathrm{Cov}\left(\hat{\Delta}_1, R_2(|\hat{\Delta}_2|^{\alpha};\Delta_2)\right) + \frac{\alpha(1-\alpha)}{2n}\mathrm{Cov}\left(\hat{\Delta}_1, R_0(|\hat{\Delta}_2|^{\alpha-2};\Delta_2)\hat{\Sigma}_2\hat{I}_2\right)
\end{aligned}
$$

We could further expand $R_2(|\hat{\Delta}_2|^{\alpha};\Delta_2)$ by Taylor expansion

$$
\mathrm{Cov}\left(\hat{\Delta}_1, R_2(|\hat{\Delta}_2|^{\alpha};\Delta_2)\right) \leq C\left(\frac{\Delta_2^{\alpha-3}\Sigma_1\Sigma_2}{n^2} + \frac{\Delta_2^{\alpha-4}\Sigma_1\Sigma_2^2}{n^3}\right) + \mathrm{Cov}\left(\hat{\Delta}_1, R_4(|\hat{\Delta}_2|^{\alpha};\Delta_2)\right)
$$

By the similar bound technique in proof of Lemma 13, we have

$$
|R_4(|\hat{\Delta}_2|^{\alpha};\Delta_2)| \leq \begin{cases} \Delta_2^{\alpha-5}(\hat{\Delta}_2-\Delta_2)^5 & \hat{\Delta}_2 > \Delta_2/2 \\ 10\Delta_2^{\alpha} & -\Delta_2 < \hat{\Delta}_2 < \Delta_2/2 \\ 10\Delta_2^{\alpha}(\hat{\Delta}_2/\Delta_2)^4 & \hat{\Delta}_2 < -\Delta \end{cases}
$$

Thus,

$$
\mathrm{Var}(R_4(|\hat{\Delta}_2|^{\alpha};\Delta_2)) \leq \mathbb{E}(R_4(|\hat{\Delta}_2|^{\alpha};\Delta_2)^2) \leq C\left(\frac{\Delta_2^{\alpha-5}\Sigma_2^5}{n^5} + \frac{\Delta_2^{2\alpha}}{n^{c-4}}\right).
$$

This suggests that

$$
\begin{aligned}
\mathrm{Cov}\left(\hat{\Delta}_1, R_2(|\hat{\Delta}_2|^{\alpha};\Delta_2)\right) &\leq C\left(\frac{\Delta_2^{\alpha-3}\Sigma_1\Sigma_2}{n^2} + \frac{\Delta_2^{\alpha-4}\Sigma_1\Sigma_2^2}{n^3} + \frac{\Delta_2^{\alpha-5}\Sigma_1^{1/2}\Sigma_2^{5/2}}{n^3} + \frac{\Delta_2^{\alpha}}{n^{c/2-4}}\right) \\
&\leq C\left(\frac{\Delta_2^{\alpha-1}\Sigma_1}{n\log n} + \frac{\Delta_2^{\alpha-1}\Sigma_1^{1/2}\Sigma_2^{1/2}}{n\log^2 n} + \frac{\Delta_2^{\alpha}}{n^{c/2-4}}\right)
\end{aligned}
$$

Similarly, we could also obtain

$$
\mathrm{Cov}(\hat{\Delta}_1, R_0(|\hat{\Delta}_2|^{\alpha-2};\Delta_2)\hat{\Sigma}_2\hat{I}_2) \leq C\left(\frac{\Delta_2^{\alpha-1}\Sigma_1}{\log n} + \frac{\Delta_2^{\alpha-1}\Sigma_1^{1/2}\Sigma_2^{1/2}}{\log^2 n} + \frac{\Delta_2^{\alpha}}{n^{c/2}}\right).
$$

Therefore, we can know

$$
\mathrm{Cov}(\hat{\Delta}_1, R_{1,2}) \leq C\left(\frac{\Delta_2^{\alpha-1}\Sigma_1}{n\log n} + \frac{\Delta_2^{\alpha-1}\Sigma_1^{1/2}\Sigma_2^{1/2}}{n\log^2 n} + \frac{\Delta_2^{\alpha}}{n^{c/2-4}}\right).
$$

With the same strategy, we can show

$$
\mathrm{Cov}(\hat{\Delta}_2, R_{1,1}) = \mathrm{Cov}(\hat{\Delta}_1, R_{1,1}) \leq C\frac{\sqrt{\Sigma_1}}{n^2}\left(\Delta_1^{\alpha-3}\sqrt{\Sigma_1^3} + \frac{\Delta_1^{\alpha}}{n^{c/2-4}}\right) \leq C\left(\frac{\Delta_1^{\alpha-1}\Sigma_1}{n\log n} + \frac{\Delta_1^{\alpha}}{n^{c/2}}\right)
$$

and

$$
\begin{aligned}
\mathrm{Cov}(R_{1,1}, R_{1,2}) \leq & C\Bigg[\frac{\Delta_2^{\alpha-2}\Sigma_2}{n^{5/2}}\left(\Delta_1^{\alpha-3}\sqrt{\Sigma_1^3} + \frac{\Delta_1^{\alpha}}{n^{c/2-4}}\right) + \frac{\Delta_1^{\alpha-2}\Sigma_1}{n^{5/2}}\left(\Delta_2^{\alpha-3}\sqrt{\Sigma_2^3} + \frac{\Delta_2^{\alpha}}{n^{c/2-4}}\right) \\
& + \frac{(\Delta_1\Delta_2)^{\alpha-2}\Sigma_1^2}{n^2} + \frac{1}{n^3}\left(\Delta_1^{\alpha-3}\sqrt{\Sigma_1^3} + \frac{\Delta_1^{\alpha}}{n^{c/2-4}}\right)\left(\Delta_2^{\alpha-3}\sqrt{\Sigma_2^3} + \frac{\Delta_2^{\alpha}}{n^{c/2-4}}\right)\Bigg] \\
\leq & C\left[\frac{(\Delta_1\Delta_2)^{\alpha-1}\Sigma_1}{n\log n} + \frac{\Delta_2^{\alpha}}{n^{c/2-4}}\right]
\end{aligned}
$$

Putting all these terms together yields

$$
\mathrm{Cov}(U_{\alpha,1}, U_{\alpha,2}) \leq C|\Delta_1\Delta_2|^{\alpha-1}\left[\frac{\Sigma_1}{n} + \frac{\sqrt{\Sigma_1\Sigma_2}}{n\log^2 n} + \frac{1}{n^{c/2-4}}\right]
$$

We complete the proof. $\square$

Lemma 16. *Suppose $n\hat{P}_i \sim \mathrm{Pois}(nP_i)$ for $i = 1, 2, 3$ and $n\hat{Q} \sim \mathrm{Pois}(nQ)$. Assume $P_2 = P_3$, $P_1 + P_2 - Q \geq \sqrt{c_1(P_1 + P_2 + Q)\log n/2n}$ and $P_1 + P_2 + Q \geq c_1 \log n/2n$. Then, when $1 < \alpha < 2$,*

$$
\mathbb{E}\left(U_\alpha(\hat{P}_1 + \hat{P}_2, \hat{Q}) - U_\alpha(\hat{P}_1 + \hat{P}_3, \hat{Q})\right)^2 \leq C\left(\frac{P_2}{n} + \frac{1}{n^{\alpha+c_1/4}}\right).
$$

Proof. In this proof, we also adopt the following notations: $\Delta = P_1 + P_2 - Q$, $\Sigma = P_1 + P_2 + Q$, $\hat{\Delta}_1 = \hat{P}_1 + \hat{P}_2 - \hat{Q}$, $\hat{\Delta}_2 = \hat{P}_1 + \hat{P}_3 - \hat{Q}$, $\hat{\Sigma}_1 = \hat{P}_1 + \hat{P}_2 + \hat{Q}$ and $\hat{\Sigma}_2 = \hat{P}_1 + \hat{P}_3 + \hat{Q}$. We also define $\hat{I}_1 = I_n(\hat{P}_1 + \hat{P}_2, \hat{Q})$ and $\hat{I}_2 = I_n(\hat{P}_1 + \hat{P}_3, \hat{Q})$. The definitions of $U_\alpha(\hat{P}_1 + \hat{P}_2, \hat{Q})$ and $U_\alpha(\hat{P}_1 + \hat{P}_3, \hat{Q})$ suggest that

$$
\begin{aligned}
& \mathbb{E}\left(U_\alpha(\hat{P}_1 + \hat{P}_2, \hat{Q}) - U_\alpha(\hat{P}_1 + \hat{P}_3, \hat{Q})\right)^2 \\
\leq & \mathbb{E}\left(|\hat{\Delta}_1|^\alpha - |\hat{\Delta}_2|^\alpha + \frac{\alpha(1-\alpha)}{2n}\left(|\hat{\Delta}_1|^{\alpha-2}\hat{\Sigma}_1\hat{I}_1 - |\hat{\Delta}_2|^{\alpha-2}\hat{\Sigma}_2\hat{I}_2\right)\right)^2 \\
\leq & 2\mathbb{E}\left(\frac{\alpha(1-\alpha)}{2n}\left(|\hat{\Delta}_1|^{\alpha-2}\hat{\Sigma}_1\hat{I}_1 - |\hat{\Delta}_2|^{\alpha-2}\hat{\Sigma}_2\hat{I}_2\right)\right)^2 + 2\mathbb{E}\left(|\hat{\Delta}_1|^\alpha - |\hat{\Delta}_2|^\alpha\right)^2.
\end{aligned}
$$

We now bound the above two terms separately. As $|x|^\alpha$ is a Lipschitz function, we have

$$
\mathbb{E}\left(|\hat{\Delta}_1|^\alpha - |\hat{\Delta}_2|^\alpha\right)^2 \leq C\mathbb{E}\left(\hat{\Delta}_1 - \hat{\Delta}_2\right)^2 \leq C\frac{P_2}{n}.
$$

It is sufficient to bound the first term. For the first terms, observe that

$$
\begin{aligned}
& \mathbb{E}\left(|\hat{\Delta}_1|^{\alpha-2}\hat{\Sigma}_1\hat{I}_1 - |\hat{\Delta}_2|^{\alpha-2}\hat{\Sigma}_2\hat{I}_2\right)^2 \\
\leq & 2\mathbb{E}\left(|\hat{\Delta}_1|^{\alpha-2}\hat{\Sigma}_1\hat{I}_1 - |\hat{\Delta}_2|^{\alpha-2}\hat{\Sigma}_1\hat{I}_2\right)^2 + 2\mathbb{E}\left(|\hat{\Delta}_2|^{\alpha-2}\hat{\Sigma}_1\hat{I}_2 - |\hat{\Delta}_2|^{\alpha-2}\hat{\Sigma}_2\hat{I}_2\right)^2 \\
:= & 2(T_1 + T_2).
\end{aligned}
$$

Because $|\hat{\Delta}_2|^{\alpha-2}\hat{I}_2 \le (c_1 \log n/4n)^{\alpha-2}$,

$$T_2 = \mathbb{E}\left(|\hat{\Delta}_2|^{\alpha-2}\hat{\Sigma}_1\hat{I}_2 - |\hat{\Delta}_2|^{\alpha-2}\hat{\Sigma}_2\hat{I}_2\right)^2 \le C\left(\frac{c_1\log n}{4n}\right)^{2(\alpha-2)}\frac{P_2}{n}.$$

For T_1, we define the event $B := \{\hat{\Delta}_1, \hat{\Delta}_2 > \sqrt{c_1 \log n\Sigma/4n}, \Sigma_1/2 \le \hat{\Sigma}_1 \le 2\Sigma_1\}$. Thus,

$$\begin{aligned}T_1 &= \mathbb{E}\left(\left(|\hat{\Delta}_1|^{\alpha-2}\hat{\Sigma}_1\hat{I}_1 - |\hat{\Delta}_2|^{\alpha-2}\hat{\Sigma}_1\hat{I}_2\right)^2\mathbf{I}_B\right) + \mathbb{E}\left(\left(|\hat{\Delta}_1|^{\alpha-2}\hat{\Sigma}_1\hat{I}_1 - |\hat{\Delta}_2|^{\alpha-2}\hat{\Sigma}_1\hat{I}_2\right)^2\mathbf{I}_{B^c}\right)\\
&\le C\left(\frac{c_1\log n\Sigma}{4n}\right)^{\alpha-3}\frac{\Sigma^2P_2}{n} + C\frac{n^{2-\alpha}}{n^{c_1/4}}.\end{aligned}$$

Here, we apply the Taylor expansion to obtain

$$|\hat{\Delta}_2|^{\alpha-2} = |\hat{\Delta}_1|^{\alpha-2} + (\alpha-2)|t\hat{\Delta}_1 + (1-t)\hat{\Delta}_2|^{\alpha-3}(\hat{\Delta}_2 - \hat{\Delta}_1)$$

for $0 \le t \le 1$, when $\hat{\Delta}_2, \hat{\Delta}_1 > 0$. Putting T_1 and T_2 together yields

$$\begin{aligned}&\mathbb{E}\left(U_\alpha(\hat{P}_1+\hat{P}_2,\hat{Q}) - U_\alpha(\hat{P}_1+\hat{P}_3,\hat{Q})\right)^2\\
\le& C\frac{P_2}{n} + C\left(\frac{c_1\log n}{4n}\right)^{2(\alpha-2)}\frac{P_2}{n^3} + C\left(\frac{c_1\log n\Sigma}{4n}\right)^{\alpha-3}\frac{\Sigma^2P_2}{n^3} + C\frac{1}{n^{\alpha+c_1/4}}\\
\le& C\frac{P_2}{n} + C\frac{1}{n^{\alpha+c_1/4}}.\end{aligned}$$

The proof is complete. □

Lemma 17. *Suppose $\{\hat{P}_e.\hat{Q}_e\}_{e\in E}$ are the empirical distribution of sample drawn from Poisson-multinomial model. Then, for any $\alpha > 1$, there exists a constant C such that*

$$\mathrm{Var}\left(D_\alpha(\hat{P},\hat{Q})\right) \le C\frac{d^2}{n},$$

where d is the height of tree.

Proof. The basic idea of proof is to apply the Efron-Stein inequality (see Boucheron et al., 2013). Because $\hat{p}_v$ and $\hat{q}_v$ are independent, the Efron-Stein inequality can be applied with respect to them. For arbitrary $v_0 \in V$, $D_\alpha(\hat{P}',\hat{Q})$ is the distance between $\hat{P}'$ and $\hat{Q}$, where $\hat{p}_{v_0}$ is replaced independent copy $\hat{p}'_{v_0}$ in $\hat{P}'$. For any $e \in E$ such that $v_0 \in \tau(e)$, we have

$$\mathbb{E}(|\hat{P}_e - \hat{Q}_e|^\alpha - |\hat{P}'_e - \hat{Q}_e|^\alpha)^2 \le \mathbb{E}(\alpha|\hat{p}_{v_0} - \hat{p}'_{v_0}|)^2 \le \frac{2\alpha^2 p_{v_0}}{n}.$$

Here, we appeal to the fact that $|x|^\alpha$ is Lipschitz function with Lipschitz constant α on $[-1,1]$, i.e.

$$||x|^\alpha - |y|^\alpha| \le \alpha |x-y|, \quad x, y \in [-1,1].$$

Since there are at most d terms involving v_0, thus

$$\begin{aligned}\mathbb{E}\left(D_\alpha(\hat{P},\hat{Q}) - D_\alpha(\hat{P}',\hat{Q})\right)^2 &\le dM^2 \sum_{v_0\in\tau(e)} \mathbb{E}(|\hat{P}_e - \hat{Q}_e|^\alpha - |\hat{P}'_e - \hat{Q}_e|^\alpha)^2 \\ &\le (dM)^2 \frac{2\alpha^2 p_{v_0}}{n}.\end{aligned}$$

By Efron-Stein inequality, we can know that

$$\mathrm{Var}\left(D_\alpha(\hat{P},\hat{Q})\right) \le \frac{1}{2}\sum_{v\in V}(dM)^2 \frac{2\alpha^2(p_v+q_v)}{n} \le C\frac{d^2}{n}.$$

Then, we complete proof. □

S2.1 Lemmas on Approximation Theory

To introduce lemmas on approximation theory, we need the following definitions. The first order symmetric difference of a function f is defined as

$$\Delta_h^1 f(x) = f\left(x+\frac{h}{2}\right) - f\left(x-\frac{h}{2}\right),$$

and the second order symmetric difference of a function f is defined as

$$\Delta_h^2 f(x) = \Delta_h(\Delta_h^1 f(x)) = f(x+h) + f(x-h) - 2f(x).$$

The rth order symmetric difference the can be defined as $\Delta_h^r f(x) = \Delta_h(\Delta_h^{r-1}(x))$. Denoted by $\varphi(x) = \sqrt{x(1-x)}$, the rth order Ditzian-Totik modulus of smoothness of function $f : [0,1] \to \mathbb{R}$ is defined as

$$\omega_\varphi^r(f,t) = \sup_{0<h\le t} \|\Delta_{h\varphi}^r f(x)\|_\infty.$$

If f is a function defined on $[0,1]^2$, then rth order Ditzian-Totik modulus of smoothness can be defined similarly

$$\omega_{[0,1]^2}^r(f,t) = \sup_{i=1,2,0<h\le t, x\in[0,1]^2} |\Delta_{i,h\varphi(x_i)}^r f(x)|,$$

where $\Delta_{i,h}$ denotes the symmetric difference with respect to the i-th coordinate. The next lemma shows the best polynomial approximation error can be upper bounded by Ditzian-Totik modulus.

Lemma 18 (Ditzian and Totik (2012), see also Jiao et al. (2018))**.** *There exists a constant $M(r) > 0$ such that for any function $f \in C[0,1]$,*

$$E_K(f, [0,1]) \leq M(r)\omega_\varphi^r(f, K^{-1}), \quad K > r,$$

where $E_K(f, I)$ denotes the distance of function f to the space of polynomials at most degree K in the uniform norm $\|\cdot\|_\infty$ on set I. Moreover, if $f(x) : [0,1]^2 \to R$, we have

$$E_K(f, [0,1]^2) \leq M\omega_{[0,1]^2}^r(f, K^{-1}), \quad K > r,$$

where M is a constant independent from f and K.

Lemma 19. *Suppose $0 < \alpha < 2$ and $x, y \in [0,1]$. Then,*

$$\omega_{[0,1]}^2((\sqrt{x} + \sqrt{y})^\alpha, t) \leq Ct^\alpha \qquad \text{and} \qquad \omega_{[0,1]}^2(|\sqrt{x} - \sqrt{y}|^\alpha, t) \leq Ct^\alpha$$

for some constant C.

Proof. We first work on $f(x,y) = (\sqrt{x} + \sqrt{y})^\alpha$. Since x and y exchangeable in f, it is sufficient to show that

$$g_1(t) := \sup_{0<h\leq t,(x,y)\in[0,1]^2} |(\sqrt{x + h\varphi(x)} + \sqrt{y})^\alpha + (\sqrt{x - h\varphi(x)} + \sqrt{y})^\alpha - 2(\sqrt{x} + \sqrt{y})^\alpha| \leq Ct^\alpha$$

for some constant C, value of which could be different place from place. With Thoerem 4.1.1 Ditzian and Totik (2012), we can show that

$$\begin{aligned}
g_1(t) \leq& C \sup_{0<h\leq t, x\geq 4h^2, y\in[0,1]} |(\sqrt{x + hx^{1/2}} + \sqrt{y})^\alpha + (\sqrt{x - hx^{1/2}} + \sqrt{y})^\alpha - 2(\sqrt{x} + \sqrt{y})^\alpha| \\
\leq& C \sup_{0<h\leq t, x\geq 4h^2, y\in[0,1]} |(\sqrt{x + \xi_1 hx^{1/2}} + \sqrt{y})^{\alpha-1} - (\sqrt{x - \xi_2 hx^{1/2}} + \sqrt{y})^{\alpha-1}|h \\
\leq& C \sup_{0<h\leq t, x\geq 4h^2, y\in[0,1]} |(\sqrt{x + \xi_3 hx^{1/2}} + \sqrt{y})^{\alpha-2}|h^2 \\
\leq& Ct^\alpha.
\end{aligned}$$

Here, ξ_1 and ξ_2 are two constants between 0 and 1 and ξ_3 is constant between -1 and 1.

Next, we work on $f(x,y) = |\sqrt{x} - \sqrt{y}|^\alpha$. Define

$$g_2(t) := \sup_{0<h\le t,(x,y)\in[0,1]^2} \left| |\sqrt{x+hx^{1/2}} - \sqrt{y}|^\alpha + |\sqrt{x-hx^{1/2}} - \sqrt{y}|^\alpha - 2|\sqrt{x}-\sqrt{y}|^\alpha \right|.$$

To bound $g_2(t)$, we consider two cases. First, we assume $x - hx^{1/2} > y$ or $x + hx^{1/2} < y$. With the same arguments in bounding $g_1(t)$, we can show

$$\left| |\sqrt{x+hx^{1/2}} - \sqrt{y}|^\alpha + |\sqrt{x-hx^{1/2}} - \sqrt{y}|^\alpha - 2|\sqrt{x}-\sqrt{y}|^\alpha \right| \le Ch^\alpha.$$

Next, we assume $x - hx^{1/2} < y < x + hx^{1/2}$. Then,

$$\begin{aligned}
&\left| |\sqrt{x+hx^{1/2}} - \sqrt{y}|^\alpha + |\sqrt{x-hx^{1/2}} - \sqrt{y}|^\alpha - 2|\sqrt{x}-\sqrt{y}|^\alpha \right| \\
\le& 4\left(\sqrt{x+hx^{1/2}} - \sqrt{x-hx^{1/2}}\right)^\alpha \\
\le& 4h^\alpha.
\end{aligned}$$

Thus, we can conclude that

$$g_2(t) \le Ct^\alpha.$$

□

Lemma 20. *For any $0 < \alpha < 2$, there exists polynomial of degree at most $2K$ $F_K^M(x,y)$ such that*

$$|F_K^M(x,y) - |x-y|^\alpha| \le C_1\left(\frac{M^{\alpha/2}(x+y)^{\alpha/2}}{K^\alpha} + \frac{M^\alpha}{K^{2\alpha}}\right), \quad \forall (x,y)\in[0,M]^2.$$

for constant C_1. Furthermore, if

$$F_K^M(x,y) = \sum_{n_1,n_2=0}^{K} f(n_1,n_2)x^{n_1}y^{n_2},$$

then the coefficients of $f(n_1,n_2)$ are bounded by $C_2(\sqrt{2}+1)^{8K}M^{\alpha-n_1-n_2}$.

Proof. As $|x-y|^\alpha = (\sqrt{x}+\sqrt{y})^\alpha|\sqrt{x}-\sqrt{y}|^\alpha$, we approximate $(\sqrt{x}+\sqrt{y})^\alpha$ and $|\sqrt{x}-\sqrt{y}|^\alpha$ separately. More concretely, Lemma 18 and Lemma 19 suggest that there exist polynomials U_K and V_K such that

$$\sup_{(x,y)\in[0,1]^2} |U_K(x,y) - (\sqrt{x}+\sqrt{y})^\alpha| \le \frac{C_1}{K^\alpha} \text{ and } \sup_{(x,y)\in[0,1]^2} |V_K(x,y) - |\sqrt{x}-\sqrt{y}|^\alpha| \le \frac{C_2}{K^\alpha}$$

for constants C_1 and C_2. Thus, we could use $U_K V_K$ to approximate $|x-y|^\alpha$. Since

$$\begin{aligned}
&|U_K(x,y)V_K(x,y) - |x-y|^\alpha| \\
=&\left|U_K(x,y)V_K(x,y) - U_K(x,y)|\sqrt{x}-\sqrt{y}|^\alpha + U_K(x,y)|\sqrt{x}-\sqrt{y}|^\alpha - |x-y|^\alpha\right| \\
\leq&|U_K(x,y)||V_K(x,y) - |\sqrt{x}-\sqrt{y}|^\alpha| + |\sqrt{x}-\sqrt{y}|^\alpha|U_K(x,y) - (\sqrt{x}+\sqrt{y})^\alpha| \\
\leq&|\sqrt{x}-\sqrt{y}|^\alpha|U_K(x,y) - (\sqrt{x}+\sqrt{y})^\alpha| + (\sqrt{x}+\sqrt{y})^\alpha|V_K(x,y) - |\sqrt{x}-\sqrt{y}|^\alpha| \\
&+ |U_K(x,y) - (\sqrt{x}+\sqrt{y})^\alpha||V_K(x,y) - |\sqrt{x}-\sqrt{y}|^\alpha|,
\end{aligned}$$

we can know

$$\sup_{(x,y)\in[0,1]^2} |U_K(x,y)V_K(x,y) - |x-y|^\alpha| \leq \frac{4(C_1+C_2)(x+y)^{\alpha/2}}{K^\alpha} + \frac{C_1C_2}{K^{2\alpha}}.$$

By scaling $\tilde{x} = xM$ and $\tilde{y} = yM$,

$$\sup_{(\tilde{x},\tilde{y})\in[0,M]^2} \left|M^\alpha U_K\left(\frac{\tilde{x}}{M},\frac{\tilde{y}}{M}\right)V_K\left(\frac{\tilde{x}}{M},\frac{\tilde{y}}{M}\right) - |\tilde{x}-\tilde{y}|^\alpha\right| \leq C\left(\frac{M^{\alpha/2}(\tilde{x}+\tilde{y})^{\alpha/2}}{K^\alpha} + \frac{M^\alpha}{K^{2\alpha}}\right).$$

Therefore, we have already constructed a polynomial $F_K^M(\tilde{x},\tilde{y}) = M^\alpha U_K(\tilde{x}/M,\tilde{y}/M)\,V_K(\tilde{x}/M,\tilde{y}/M)$. An application of Lemma 17 in Jiao et al. (2018) could yields the conclusion on coefficients of F_K^M. □

Lemma 21 (Timan (2014)). *If $\alpha > 0$, there exists polynomial of degree at most K $F_K^M(x)$ such that*

$$C_1\left(\frac{M}{K}\right)^\alpha \leq \sup_{-M\leq x\leq M} |F_K^M(x) - |x|^\alpha| \leq C_2\left(\frac{M}{K}\right)^\alpha.$$

for constant C_1 and C_2.

Lemma 22 (Cai and Low (2011)). *For any given even integer $K > 0$, there exist two probability measures ν_1 and ν_2 on $[-1,1]$ that satisfy the following conditions:*

- *ν_1 and ν_2 are symmetric around 0;*
- *$\int t^k \nu_1(dt) = \int t^k \nu_2(dt)$, for $k = 0,1,\ldots,K$;*
- *$\int f(t)\nu_1(dt) - \int f(t)\nu_2(dt) = 2\delta_K$,*

where δ_K is the distance in the uniform norm on $[-1,1]$ from function $f(x)$ to the space of polynomials of no more than degree K.

References

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

T. T. Cai and M. Low. Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.

P. Diaconis and S. Zabell. Closed form summation for classical distributions: variations on a theme of de moivre. *Statistical Science*, pages 284–302, 1991.

Z. Ditzian and V. Totik. *Moduli of smoothness*, volume 9. Springer Science & Business Media, 2012.

Y. Han, J. Jiao, and T. Weissman. Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under wasserstein distance. *arXiv preprint arXiv:1802.08405*, 2018.

J. Jiao, Y. Han, and T. Weissman. Minimax estimation of the l1 distance. *IEEE Transactions on Information Theory*, 2018.

M. Mitzenmacher and E. Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge university press, 2005.

A. F. Timan. *Theory of approximation of functions of a real variable*, volume 34. Elsevier, 2014.

A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.