Supplementary Material for: Entropy Partial Transport with Tree Metrics: Theory and Practice

Tam Le[*]
RIKEN AIP

Truyen Nguyen[*]
The University of Akron

This is the supplementary material for the main text in [11]. We organize it as follows:

- We give detailed proofs of the theoretical results in the main text for the entropy partial transport (EPT) problem for nonnegative measures on a tree having different masses in §A.
- We provide further experimental results in §B, e.g.,
  - about more setups for the efficient approximation of $\widetilde{\mathrm{ET}}_{\lambda}^{0}$ for $\widetilde{\mathrm{ET}}_{\lambda}^{\alpha}$,
  - about different values of α,
  - about different numbers of slices,
  - and about different parameters in tree metric sampling.
- We next give more details and discussions in §C, e.g.,
  - more details about experiments (e.g., softwares, datasets, more details about the experiment setup).
  - some brief review about kernels, and more referred details (e.g., for tree metric sampling, persistence diagrams and related mathematical definitions in topological data analysis).
  - more discussions about other relations with other work.

We note that we have released code for our proposals at

`https://github.com/lttam/EntropyPartialTransport`.

A Detailed Proofs

In this section, we present detailed proofs of the theoretical results in the main text.

[*]: Two authors contributed equally.

A.1 Proof for Theorem 3.1 in the main text

Proof. i) Note that $\mathsf{ET}_{c,\lambda}(\mu,\nu)$ is a concave function in λ since it is the infimum of a family of concave functions in λ. Therefore, u is convex on R. In particular, u is differentiable almost everywhere on R.

Let $\lambda \in \mathsf{R}$, recall the definition of $\mathcal{C}_\lambda(\gamma)$ in Equation (4) in the main text. Then for any $\gamma \in \;^0(\lambda)$, we have

$$\mathsf{ET}_{c,\lambda+\delta}(\mu,\nu) \leq \mathcal{C}_{\lambda+\delta}(\gamma) = \mathcal{C}_\lambda(\gamma) - b\delta\gamma(\mathcal{T}\times\mathcal{T}) = \mathsf{ET}_{c,\lambda}(\mu,\nu) - b\delta\gamma(\mathcal{T}\times\mathcal{T}) \;\; \forall\delta \in \mathsf{R}. \tag{1}$$

This implies that

$$\left\{b\,\gamma(\mathcal{T}\times\mathcal{T}) : \gamma \in \;^0(\lambda)\right\} \subset \partial u(\lambda).$$

We next show that the opposite inclusion is also true, i.e., $\left\{b\,\gamma(\mathcal{T}\times\mathcal{T}) : \gamma \in \;^0(\lambda)\right\} = \partial u(\lambda)$. This is obviously holds if $\partial u(\lambda)$ is singleton and hence we only need to consider λ for which the convex set $\partial u(\lambda)$ has more than one element.

Let $m \in \partial u(\lambda)$, then m can be expressed as a convex combination of extreme points $m_1, \ldots, m_N$ of $\partial u(\lambda)$, i.e., $m = \sum_{i=1}^N t_i m_i$ with $0 \leq t_i \leq 1$ and $\sum_{i=1}^N t_i = 1$. As m_i is an extreme point of $\partial u(\lambda)$, there exists a sequence $\lambda_n \to \lambda$ such that λ_n is a differentiable point of u and $u'(\lambda_n) \to m_i$.

Let $\gamma^n \in \;^0(\lambda_n)$, then $b\,\gamma^n(\mathcal{T}\times\mathcal{T}) = u'(\lambda_n) \to m_i$. By compactness, there exists a subsequence $\{\gamma^{n_k}\}$ and $\boldsymbol{\gamma}^i \in \;_{\leq}(\mu,\nu)$ such that $\gamma^{n_k} \to \boldsymbol{\gamma}^i$ weakly. It follows that $\gamma^{n_k}(\mathcal{T}\times\mathcal{T}) \to \boldsymbol{\gamma}^i(\mathcal{T}\times\mathcal{T})$, and hence we must have $b\,\boldsymbol{\gamma}^i(\mathcal{T}\times\mathcal{T}) = m_i$. We have

$$\begin{aligned}\mathcal{C}_{\lambda_{n_k}}(\gamma^{\lambda_{n_k}}) = \mathcal{C}_\lambda(\gamma^{\lambda_{n_k}}) + b(\lambda - \lambda_{n_k})\gamma^{n_k}(\mathcal{T}\times\mathcal{T}) &\geq \mathsf{ET}_{c,\lambda}(\mu,\nu) + b(\lambda - \lambda_{n_k})\gamma^{n_k}(\mathcal{T}\times\mathcal{T})\\ &\geq \mathsf{ET}_{c,\lambda}(\mu,\nu) - bm|\lambda - \lambda_{n_k}|\end{aligned}$$

and for any $\gamma \in \;^0(\lambda)$, there holds

$$\mathcal{C}_{\lambda_{n_k}}(\gamma^{\lambda_{n_k}}) \leq \mathcal{C}_{\lambda_{n_k}}(\gamma) = \mathcal{C}_\lambda(\gamma) + b(\lambda - \lambda_{n_k})\gamma(\mathcal{T}\times\mathcal{T}) = \mathsf{ET}_{c,\lambda}(\mu,\nu) + b(\lambda - \lambda_{n_k})\gamma(\mathcal{T}\times\mathcal{T}).$$

We thus deduce that $\lim_{k\to\infty} \mathcal{C}_{\lambda_{n_k}}(\gamma^{\lambda_{n_k}}) = \mathsf{ET}_{c,\lambda}(\mu,\nu)$. These together with the lower semicontinuity of $\mathcal{C}_\lambda$ give

$$\begin{aligned}\mathsf{ET}_{c,\lambda}(\mu,\nu) = \liminf_{k\to\infty} \mathcal{C}_{\lambda_{n_k}}(\gamma^{\lambda_{n_k}}) &= \liminf_{k\to\infty}\left[\mathcal{C}_\lambda(\gamma^{\lambda_{n_k}}) + b(\lambda - \lambda_{n_k})\gamma^{n_k}(\mathcal{T}\times\mathcal{T})\right]\\ &= \liminf_{k\to\infty} \mathcal{C}_\lambda(\gamma^{\lambda_{n_k}}) \geq \mathcal{C}_\lambda(\boldsymbol{\gamma}^i).\end{aligned}$$

Therefore, $\boldsymbol{\gamma}^i \in \;^0(\lambda)$ with mass $b\,\boldsymbol{\gamma}^i(\mathcal{T}\times\mathcal{T}) = m_i$. Due to the convexity of $\;^0(\lambda)$, we have $\gamma := \sum_{i=1}^N t_i \boldsymbol{\gamma}^i \in \;^0(\lambda)$ with $b\,\gamma(\mathcal{T}\times\mathcal{T}) = \sum_{i=1}^N t_i m_i = m$. That is,

$$\partial u(\lambda) \subset \left\{b\,\gamma(\mathcal{T}\times\mathcal{T}) : \gamma \in \;^0(\lambda)\right\},$$

and we thus infer that $\left\{b\,\gamma(\mathcal{T}\times\mathcal{T}) : \gamma \in \;^0(\lambda)\right\} = \partial u(\lambda)$ for all $\lambda \in \mathsf{R}$.

In order to prove the second part of i), let $\gamma \in \;^0(\lambda_1)$ and $\boldsymbol{\gamma} \in \;^0(\lambda_2)$ be arbitrary. We have

$$\begin{aligned}\mathsf{ET}_{c,\lambda_2}(\mu,\nu) = \mathcal{C}_{\lambda_2}(\boldsymbol{\gamma}) &= \mathcal{C}_{\lambda_1}(\boldsymbol{\gamma}) - b(\lambda_2 - \lambda_1)\boldsymbol{\gamma}(\mathcal{T}\times\mathcal{T})\\ &\geq \mathsf{ET}_{c,\lambda_1}(\mu,\nu) - b(\lambda_2 - \lambda_1)\boldsymbol{\gamma}(\mathcal{T}\times\mathcal{T}).\end{aligned} \tag{2}$$

Hence by combining with (1), we deduce that

$$\mathsf{ET}_{c,\lambda_1}(\mu,\nu) - b(\lambda_2 - \lambda_1)\boldsymbol{\gamma}(\mathcal{T}\times\mathcal{T}) \leq \mathsf{ET}_{c,\lambda_2}(\mu,\nu) \leq \mathsf{ET}_{c,\lambda_1}(\mu,\nu) - b(\lambda_2 - \lambda_1)\gamma(\mathcal{T}\times\mathcal{T}),$$

which yields $\gamma(\mathcal{T}\times\mathcal{T}) \leq \boldsymbol{\gamma}(\mathcal{T}\times\mathcal{T})$. This together with the above characterization of $\partial u(\lambda)$ implies the second part of i).

ii) If u is differentiable at λ, then $\partial u(\lambda)$ is a singleton set. However, as $\partial u(\lambda) = \left\{b\,\gamma(\mathcal{T}\times\mathcal{T}) : \gamma \in \;^0(\lambda)\right\}$ by i), we thus infer that the mass $\gamma(\mathcal{T}\times\mathcal{T})$ must be the same for every $\gamma \in \;^0(\lambda)$.

Next assume that every element in $\;^0(\lambda)$ has the same mass, say m. For $\delta \neq 0$, let $\gamma^{\lambda+\delta} \in \;^0(\lambda+\delta)$ and $m(\lambda+\delta) := \gamma^{\lambda+\delta}(\mathcal{T}\times\mathcal{T})$. Then, we claim that

$$\lim_{\delta\to 0} m(\lambda+\delta) = m. \tag{3}$$

Assume the claim for the moment, and let $\delta > 0$. Then, as in (1)–(2), we have

$$\mathsf{ET}_{c,\lambda+\delta}(\mu,\nu) \leq \mathsf{ET}_{c,\lambda}(\mu,\nu) - b\delta m \quad \text{and} \quad \mathsf{ET}_{c,\lambda+\delta}(\mu,\nu) \geq \mathsf{ET}_{c,\lambda}(\mu,\nu) - b\delta m(\lambda+\delta).$$

It follows that

$$-bm(\lambda+\delta) \leq \frac{\mathsf{ET}_{c,\lambda+\delta}(\mu,\nu) - \mathsf{ET}_{c,\lambda}(\mu,\nu)}{\delta} \leq -bm.$$

This together with claim (3) gives $\lim_{\delta\to 0^+} \frac{\mathrm{ET}_{c,\lambda+\delta}(\mu,\nu)-\mathrm{ET}_{c,\lambda}(\mu,\nu)}{\delta} = -bm$. By the same argument, we also have $\lim_{\delta\to 0} \frac{\mathrm{ET}_{c,\lambda+\delta}(\mu,\nu)-\mathrm{ET}_{c,\lambda}(\mu,\nu)}{\delta} = -bm$. Thus, we infer that u is differentiable at λ with $u'(\lambda) = bm$. Therefore, it remains to prove claim (3).

Indeed, by compactness there exists a subsequence, still labeled by $\gamma^{\lambda+\delta}$, and $\gamma \in {}_{\leq}(\mu,\nu)$ such that $\gamma^{\lambda+\delta} \to \gamma$ weakly as $\delta \to 0$. As in i), we can show that $\gamma \in {}^{0}(\lambda)$. Then, as the mass functional is weakly continuous, we obtain $m(\lambda+\delta) = \gamma^{\lambda+\delta}(\mathcal{T}\times\mathcal{T}) \to \gamma(\mathcal{T}\times\mathcal{T}) = m$. We in fact have shown that any subsequence of $\{m(\lambda+\delta)\}_\delta$ has a further subsequence converging to the same number m. Therefore, the full sequence $\{m(\lambda+\delta)\}_\delta$ must converge to m, and hence (3) is proved.

iii) For any $\lambda \in \mathsf{R}$, we have $\partial u(\lambda) = \big\{b\,\gamma(\mathcal{T}\times\mathcal{T}) : \gamma \in {}^{0}(\lambda)\big\} \subset [0, b\,m]$. Thus, we only need to prove $[0, b\,m] \subset \partial u(\mathsf{R})$. First, note that as $\partial u(\lambda) \subset \mathsf{R}$ is a compact and convex set, it must be a finite and closed interval. Therefore, if we let

$$\gamma^\lambda_{min} := \operatorname*{arg\,min}_{\gamma\in\Gamma^0(\lambda)} \gamma(\mathcal{T}\times\mathcal{T}) \quad \text{and} \quad \gamma^\lambda_{max} := \operatorname*{arg\,max}_{\gamma\in\Gamma^0(\lambda)} \gamma(\mathcal{T}\times\mathcal{T}),$$

then it follows from ii) that $\partial u(\lambda) = \big[b\,\gamma^\lambda_{min}(\mathcal{T}\times\mathcal{T}), b\,\gamma^\lambda_{max}(\mathcal{T}\times\mathcal{T})\big]$ for every $\lambda \in \mathsf{R}$. From Equation (4) in the main text, it is clear that $\partial u(\lambda) = \{0\}$ for λ negative enough. Indeed, if we take $\lambda < -M$, then as $w_1(x) + w_2(y) \leq b\,c(x,y) + M$, we have $0 < b\,c(x,y) - w_1(x) - w_2(y) - \lambda$ for all $x, y \in \mathcal{T}$. Then, we obtain from Equation (4) in the main text that $\mathcal{C}_\lambda(0) \leq \mathcal{C}_\lambda(\gamma)$ for every $\gamma \in {}_{\leq}(\mu,\nu)$ and the strict inequality holds if $\gamma \neq 0$. Thus, ${}^{0}(\lambda) = \{0\}$ which gives $\partial u(\lambda) = \{0\}$ and $u(\lambda) = -\int_{\mathcal{T}} w_1 \mu(dx) - \int_{\mathcal{T}} w_2 \nu(dx)$.

We next show that $\partial u(\lambda) = \{b\,m\}$ for λ positive enough. Since $c(x,y)$ is bounded due to its continuity on $\mathcal{T}\times\mathcal{T}$, we can choose $\lambda \in \mathsf{R}$ such that $c(x,y) - \lambda < 0$ for all $x, y \in \mathcal{T}$. Let $\gamma \in {}^{0}(\lambda)$. We claim that either $\gamma_1 = \mu$ or $\gamma_2 = \nu$. Indeed, since otherwise we have $\gamma_1(A_0) < \mu(A_0)$ and $\gamma_2(B_0) < \nu(B_0)$ for some Borel sets $A_0, B_0 \subset \mathcal{T}$. Let $\boldsymbol{\gamma} := \gamma + [(\mu - \gamma_1)\chi_{A_0}] \otimes [(\nu - \gamma_2)\chi_{B_0}]$. Then, for any Borel set $A \subset \mathcal{T}$ we have

$$\begin{aligned}\boldsymbol{\gamma}_1(A) = \gamma_1(A) + \mu(A\cap A_0) - \gamma_1(A\cap A_0) &= \gamma_1(A\setminus A_0) + \mu(A\cap A_0)\\ &\leq \mu(A\setminus A_0) + \mu(A\cap A_0) = \mu(A).\end{aligned}$$

Likewise, $\boldsymbol{\gamma}_2(B) \leq \nu(B)$ for any Borel set $B \subset \mathcal{T}$. Thus $\boldsymbol{\gamma} \in {}_{\leq}(\mu,\nu)$. On the other hand, it is clear from Equation (4) in the main text and the facts $\gamma_1 \leq \boldsymbol{\gamma}_1$, $\gamma_2 \leq \boldsymbol{\gamma}_2$, and $c - \lambda < 0$ that $\mathcal{C}_\lambda(\boldsymbol{\gamma}) < \mathcal{C}_\lambda(\gamma)$. This is impossible and so the claim is proved. That is, either $\gamma_1 = \mu$ or $\gamma_2 = \nu$. It follows that $\gamma(\mathcal{T}\times\mathcal{T}) = m$ for every $\gamma \in {}^{0}(\lambda)$, and hence $\partial u(\lambda) = \{b\,m\}$. This also means that u is differentiable at λ with $u'(\lambda) = b\,m$.

Therefore, it remains to show that

$$(0, b\,m) \subset \partial u(\mathsf{R}) = \bigcup_{\lambda\in\mathsf{R}} \big[b\,\gamma^\lambda_{min}(\mathcal{T}\times\mathcal{T}), b\,\gamma^\lambda_{max}(\mathcal{T}\times\mathcal{T})\big]. \tag{4}$$

Assume by contradiction that there exists $m \in (0, b\,m)$ such that $m \notin \partial u(\lambda)$ for every $\lambda \in \mathsf{R}$. For convenience, we adopt the following notation: for sets $A, B \subset \mathsf{R}$ and $r \in \mathsf{R}$, we write $A < r$ if $a < r$ for every $a \in A$, and $A < B$ if $a < b$ for every $a \in A$ and $b \in B$. Let us consider the following two sets

$$S_1 := \{\lambda : \partial u(\lambda) < m\} \quad \text{and} \quad S_2 := \{\lambda : \partial u(\lambda) > m\}.$$

Then $\lambda \in S_1$ if λ is negative enough, and $\lambda \in S_2$ if λ is positive enough. For any $\lambda_1 \in S_1$ and $\lambda_2 \in S_2$, we have $\partial u(\lambda_1) < m < \partial u(\lambda_2)$, and hence $\lambda_1 < \lambda_2$ by the monotonicity in i). That is, $S_1 < S_2$ and so we obtain

$$\lambda^* := \sup\{\lambda : \lambda \in S_1\} \leq \inf\{\lambda : \lambda \in S_2\} =: \lambda^{**}.$$

If $\lambda^* < \lambda^{**}$, then for any $\lambda \in (\lambda^*, \lambda^{**})$ we have $\lambda \notin S_1$ and $\lambda \notin S_2$. Therefore, $\partial u(\lambda) \not< m$ and $\partial u(\lambda) \not> m$. Hence, we can find $m_1, m_2 \in \partial u(\lambda)$ such that $m_1 \geq m$ and $m_2 \leq m$. Thus, $m \in [m_2, m_1] \subset \partial u(\lambda)$ due to the convexity of the set $\partial u(\lambda)$. This contradicts our hypothesis, and we conclude that $\lambda^* = \lambda^{**}$.

We next select sequences $\{\lambda_n^1\} \subset S_1$ and $\{\lambda_n^2\} \subset S_2$ such that $\lambda_n^1 \to \lambda^*$ and $\lambda_n^2 \to \lambda^{**} = \lambda^*$. For each n, let

$$\gamma_{min}^n := \operatorname*{arg\,min}_{\gamma \in \Gamma^0(\lambda_n^1)} \gamma(\mathcal{T} \times \mathcal{T}) \quad \text{and} \quad \gamma_{max}^n := \operatorname*{arg\,max}_{\gamma \in \Gamma^0(\lambda_n^2)} \gamma(\mathcal{T} \times \mathcal{T}).$$

By compactness, there exist subsequences, still labeled as $\{\gamma_{min}^n\}$ and $\{\gamma_{max}^n\}$, and $\gamma^*, \gamma^{**} \in {}_{\leq}(\mu, \nu)$ such that $\gamma_{min}^n \to \gamma^*$ weakly and $\gamma_{max}^n \to \gamma^{**}$ weakly. By arguing exactly as in i), we then obtain $\gamma^*, \gamma^{**} \in {}^{0}(\lambda^*)$, $\gamma_{min}^n(\mathcal{T} \times \mathcal{T}) \to \gamma^*(\mathcal{T} \times \mathcal{T})$, and $\gamma_{max}^n(\mathcal{T} \times \mathcal{T}) \to \gamma^{**}(\mathcal{T} \times \mathcal{T})$. As $b\gamma_{min}^n(\mathcal{T} \times \mathcal{T}) < m$ due to $\lambda_n^1 \in S_1$, we must have $b\gamma^*(\mathcal{T} \times \mathcal{T}) \leq m$. Likewise, we have $b\gamma^{**}(\mathcal{T} \times \mathcal{T}) \geq m$ as $b\gamma_{max}^n(\mathcal{T} \times \mathcal{T}) > m$ for all n. Hence, $m \in [b\gamma^*(\mathcal{T} \times \mathcal{T}), b\gamma^{**}(\mathcal{T} \times \mathcal{T})]$. Since $\gamma^*, \gamma^{**} \in {}^{0}(\lambda^*)$, we infer that $m \in \partial u(\lambda^*)$. This is a contradiction and the proof is complete. We note that since $\lambda_n^1 \leq \lambda^* \leq \lambda_n^2$, we have from the monotonicity in i) that

$$\gamma_{min}^n(\mathcal{T} \times \mathcal{T}) \leq \gamma(\mathcal{T} \times \mathcal{T}) \leq \gamma_{max}^n(\mathcal{T} \times \mathcal{T})$$

for every $\gamma \in {}^{0}(\lambda^*)$. By sending n to infinity, it follows that $\gamma^*(\mathcal{T} \times \mathcal{T}) \leq \gamma(\mathcal{T} \times \mathcal{T}) \leq \gamma^{**}(\mathcal{T} \times \mathcal{T})$ for every $\gamma \in {}^{0}(\lambda^*)$. That is, $\gamma^* = \gamma_{min}^{\lambda}$ and $\gamma^{**} = \gamma_{max}^{\lambda}$. ■

A.2 Proof for Lemma 3.2 in the main text

Proof. We first observe for any Borel set $A \subset \mathcal{T}$ that

$$\hat{\gamma}(A \times \{\hat{s}\}) = \hat{\gamma}(A \times \hat{\mathcal{T}}) - \hat{\gamma}(A \times \mathcal{T}) = \hat{\mu}(A) - \gamma(A \times \mathcal{T}) = \mu(A) - \gamma_1(A) = \int_A (1 - f_1)\mu(dx).$$

For the same reason, we have $\hat{\gamma}(\{\hat{s}\} \times B) = \int_B (1 - f_2)\nu(dx)$ for any set Borel set $B \subset \mathcal{T}$. Also,

$$\begin{aligned}
\hat{\gamma}(\{\hat{s}\} \times \{\hat{s}\}) &= \hat{\gamma}(\hat{\mathcal{T}} \times \{\hat{s}\}) - \hat{\gamma}(\mathcal{T} \times \{\hat{s}\}) \\
&= \hat{\gamma}(\hat{\mathcal{T}} \times \hat{\mathcal{T}}) - \hat{\gamma}(\hat{\mathcal{T}} \times \mathcal{T}) - \left[\hat{\gamma}(\mathcal{T} \times \hat{\mathcal{T}}) - \hat{\gamma}(\mathcal{T} \times \mathcal{T})\right] \\
&= \hat{\mu}(\hat{\mathcal{T}}) - \hat{\nu}(\mathcal{T}) - \hat{\mu}(\mathcal{T}) + \gamma(\mathcal{T} \times \mathcal{T}) = \gamma(\mathcal{T} \times \mathcal{T}).
\end{aligned}$$

Since the Equation (6) in the main text is obviously true for sets of the form $A \times B$ with $A, B \subset \mathcal{T}$ being Borel sets, we only need to verify it for sets of the following forms: $(A \cup \{\hat{s}\}) \times B$, $A \times (B \cup \{\hat{s}\})$, $(A \cup \{\hat{s}\}) \times (B \cup \{\hat{s}\})$ for Borel sets $A, B \subset \mathcal{T}$. We check it case by case as follows.

Case 1: Using the above observation, we have

$$\hat{\gamma}((A \cup \{\hat{s}\}) \times B) = \hat{\gamma}(A \times B) + \hat{\gamma}(\{\hat{s}\} \times B) = \gamma(A \times B) + \int_B (1 - f_2)\nu(dx).$$

Therefore, the Equation (6) in the main text holds in this case.

Case 2: the Equation (6) in the main text is also true for this case because

$$\hat{\gamma}(A \times (B \cup \{\hat{s}\})) = \hat{\gamma}(A \times B) + \hat{\gamma}(A \times \{\hat{s}\}) = \gamma(A \times B) + \int_A (1 - f_1)\mu(dx).$$

Case 3: the Equation (6) in the main text is true as well since

$$\begin{aligned}
\hat{\gamma}((A \cup \{\hat{s}\}) \times (B \cup \{\hat{s}\})) &= \hat{\gamma}(A \times B) + \hat{\gamma}(A \times \{\hat{s}\}) + \hat{\gamma}(\{\hat{s}\} \times B) + \hat{\gamma}(\{\hat{s}\} \times \{\hat{s}\}) \\
&= \gamma(A \times B) + \int_A (1 - f_1)\mu(dx) + \int_B (1 - f_2)\nu(dx) + \gamma(\mathcal{T} \times \mathcal{T}).
\end{aligned}$$

Now as the Equation (6) in the main text holds, we obviously have $\gamma(U \times \mathcal{T}) \leq \hat{\gamma}(U \times \mathcal{T}) \leq \hat{\gamma}(U \times \hat{\mathcal{T}}) = \hat{\mu}(U) = \mu(U)$ for any Borel set $U \subset \mathcal{T}$. Likewise, $\gamma(\mathcal{T} \times U) \leq \nu(U)$ for any Borel set $U \subset \mathcal{T}$. Therefore, $\gamma \in {}_{\leq}(\mu, \nu)$. ■

A.3 Proof for Proposition 3.3 in the main text

Proof. We first show that $\mathsf{KT}(\hat\mu,\hat\nu) \le \mathsf{ET}_{c,\lambda}(\mu,\nu)$.

For any $\gamma \in {}_{\le}(\mu,\nu)$, let $\hat\gamma$ be given by the Equation (6) in the main text. Then, $\hat\gamma \in (\hat\mu,\hat\nu)$ and

$$\mathsf{KT}(\hat\mu,\hat\nu) \le \int_{\hat{\mathcal{T}}\times\hat{\mathcal{T}}} \hat c(x,y)\hat\gamma(dx,dy) = b\int_{\mathcal{T}\times\mathcal{T}}[c(x,y)-\lambda]\gamma(dx,dy) \\ + \int_{\mathcal{T}} w_1[1-f_1(x)]\mu(dx) + \int_{\mathcal{T}} w_2[1-f_2(x)]\nu(dx).$$

It follows that $\mathsf{KT}(\hat\mu,\hat\nu) \le \mathsf{ET}_{c,\lambda}(\mu,\nu)$.

We next show that $\mathsf{KT}(\hat\mu,\hat\nu) \ge \mathsf{ET}_{c,\lambda}(\mu,\nu)$. To see this, for any $\hat\gamma \in (\hat\mu,\hat\nu)$ we let γ be the restriction of $\hat\gamma$ to $\mathcal{T}$. Then by Lemma 3.2 in the main text, we have $\gamma \in {}_{\le}(\mu,\nu)$ and the Equation (6) in the main text holds. Consequently,

$$\int_{\hat{\mathcal{T}}\times\hat{\mathcal{T}}} \hat c(x,y)\hat\gamma(dx,dy) = b\int_{\mathcal{T}\times\mathcal{T}}[c(x,y)-\lambda]\gamma(dx,dy) \\ + \int_{\mathcal{T}} w_1[1-f_1(x)]\mu(dx) + \int_{\mathcal{T}} w_2[1-f_2(x)]\nu(dx) \\ \ge \mathsf{ET}_{c,\lambda}(\mu,\nu).$$

By taking the infimum over $\hat\gamma$, we infer that $\mathsf{KT}(\hat\mu,\hat\nu) \ge \mathsf{ET}_{c,\lambda}(\mu,\nu)$.

Thus we obtain

$$\mathsf{KT}(\hat\mu,\hat\nu) = \mathsf{ET}_{c,\lambda}(\mu,\nu).$$

The relation about the optimal solutions also follows from the above arguments. ■

A.4 Proof for Theorem 3.4 in the main text

Proof. From Proposition 3.3 in the main text and the dual formulation for $\mathsf{KT}(\hat\mu,\hat\nu)$ proved in [3, Corollary 2.6], we have

$$\mathsf{ET}_{c,\lambda}(\mu,\nu) = \sup_{\substack{\hat u\in L^1(\hat\mu),\, \hat v\in L^1(\hat\nu)\\ \hat u(x)+\hat v(y)\le \hat c(x,y)}} \int_{\hat{\mathcal{T}}} \hat u(x)\hat\mu(dx) + \int_{\hat{\mathcal{T}}} \hat v(x)\hat\nu(dx) =: I.$$

Therefore, it is enough to prove that $I = J$ where

$$J := \sup_{(u,v)\in\mathsf{K}} \Big[\int_{\mathcal{T}} u(x)\mu(dx) + \int_{\mathcal{T}} v(x)\nu(dx)\Big].$$

For (u,v) satisfying $u\le w_1$, $v\le w_2$ and $u(x)+v(y)\le b[c(x,y)-\lambda]$, we extend it to $\hat{\mathcal{T}}$ by taking $\hat u(\hat s) = 0$ and $\hat v(\hat s) = 0$. Then, it is clear that $\hat u(x)+\hat v(y)\le \hat c(x,y)$ for $x,y\in\hat{\mathcal{T}}$, and

$$I \ge \int_{\hat{\mathcal{T}}} \hat u(x)\hat\mu(dx) + \int_{\hat{\mathcal{T}}} \hat v(x)\hat\nu(dx) = \int_{\mathcal{T}} u(x)\mu(dx) + \int_{\mathcal{T}} v(x)\nu(dx).$$

It follows that $I \ge J$. In order to prove the converse, let $(\hat u,\hat v)$ be a maximizer for I. Then, by considering $(\hat u-\hat u(\hat s), \hat v+\hat u(\hat s))$, we can assume that $\hat u(\hat s) = 0$. Also, if we let $v(y) := \inf_{x\in\hat{\mathcal{T}}}[\hat c(x,y)-\hat u(x)]$, then $(\hat u, v)$ is still in the admissible class for I and $\hat v(y)\le v(y)$. This implies that $(\hat u,v)$ is also a maximizer for I. For these reasons, we can assume w.l.g. that the maximizer $(\hat u,\hat v)$ has the following additional properties: $\hat u(\hat s) = 0$ and

$$\hat v(y) = \inf_{x\in\hat{\mathcal{T}}}[\hat c(x,y)-\hat u(x)] \quad \forall y\in\hat{\mathcal{T}}.$$

In particular, $\hat v(\hat s) = \inf_{x\in\hat{\mathcal{T}}}[\hat c(x,\hat s)-\hat u(x)]$. For convenience, define $w_1(\hat s) = 0$ and consider the following two possibilities.

Case 1: $\inf_{x\in\hat{\mathcal{T}}}[w_1(x)-\hat u(x)] \geq 0$. Then, since $\hat c(\hat s,\hat s)-\hat u(\hat s) = 0$ and $\inf_{x\in\mathcal{T}}[\hat c(x,\hat s)-\hat u(x)] = \inf_{x\in\mathcal{T}}[w_1(x)-\hat u(x)] \geq 0$, we have $\hat v(\hat s) = 0$. Also, $\hat v(y) \leq \hat c(\hat s, y) - \hat u(\hat s) \leq w_2(y)$ for all $y \in \hat{\mathcal{T}}$. For each $y \in \mathcal{T}$, by using the facts $\hat u \leq w_1$ and $\hat c(\hat s, y) - w_1(\hat s) = w_2(y) \geq 0$ we get

$$\hat v(y) \geq \inf_{x\in\hat{\mathcal{T}}}[\hat c(x,y) - w_1(x)] = \inf_{x\in\mathcal{T}}\{b[c(x,y)-\lambda] - w_1(x)\} = -b\lambda + \inf_{x\in\mathcal{T}}[b\, c(x,y) - w_1(x)].$$

Thus $(\hat u, \hat v) \in \mathsf{K}$ and

$$\begin{aligned} I = \int_{\hat{\mathcal{T}}} \hat u(x)\hat\mu(dx) + \int_{\hat{\mathcal{T}}} \hat v(x)\hat\nu(dx) &= \int_{\mathcal{T}} \hat u(x)\hat\mu(dx) + \int_{\mathcal{T}} \hat v(x)\hat\nu(dx) + \hat v(\hat s)\mu(\mathcal{T}) \\ &= \int_{\mathcal{T}} \hat u(x)\mu(dx) + \int_{\mathcal{T}} \hat v(x)\nu(dx) \leq J. \end{aligned}$$

Case 2: $\inf_{x\in\hat{\mathcal{T}}}[w_1(x) - \hat u(x)] < 0$. Then, by arguing as in Case 1, we have $\hat v(\hat s) = \inf_{x\in\mathcal{T}}[w_1(x) - \hat u(x)] < 0$ and

$$I = \int_{\mathcal{T}} \hat v(x)\nu(dx) + \int_{\mathcal{T}} \hat u(x)\mu(dx) + \mu(\mathcal{T}) \inf_{\mathcal{T}}[w_1 - \hat u]. \tag{5}$$

Let $\tilde u(x) := \min\{\hat u(x), w_1(x)\}$. Then, it is obvious that $\tilde u(x) + \hat v(y) \leq \hat c(x,y)$ and $\tilde u(\hat s) = 0$. Since $\inf_{x\in\mathcal{T}}[w_1(x) - \hat u(x)] < 0$, there exists $x_0 \in \mathcal{T}$ such that $w_1(x_0) < \hat u(x_0)$. Thus, $\tilde u(x_0) = w_1(x_0)$ and hence $\inf_{\mathcal{T}}[w_1 - \tilde u] \leq 0$. As $\tilde u \leq w_1$, we infer further that $\inf_{\mathcal{T}}[w_1 - \tilde u] = 0$. We also have

$$\begin{aligned} &\int_{\mathcal{T}} \hat u(x)\mu(dx) + \mu(\mathcal{T}) \inf_{\mathcal{T}}[w_1 - \hat u] \\ &= \int_{\mathcal{T}} \tilde u(x)\mu(dx) + \int_{\mathcal{T}:\hat u > w_1} [\hat u(x) - w_1(x)]\mu(dx) + \mu(\mathcal{T}) \inf_{\mathcal{T}}[w_1 - \hat u] \leq \int_{\mathcal{T}} \tilde u(x)\mu(dx). \end{aligned}$$

This together with (5) gives

$$I \leq \int_{\mathcal{T}} \tilde u(x)\mu(dx) + \int_{\mathcal{T}} \hat v(x)\nu(dx).$$

Now let $\tilde v(y) = \inf_{x\in\hat{\mathcal{T}}}[\hat c(x,y) - \tilde u(x)]$ for $y \in \mathcal{T}$. Then, $\hat v(y) \leq \tilde v(y) \leq \hat c(\hat s, y) - \tilde u(\hat s) = w_2(y)$ for $y \in \mathcal{T}$. For each $y \in \mathcal{T}$, by using the facts $\tilde u \leq w_1$ and $\hat c(\hat s, y) - w_1(\hat s) = w_2(y) \geq 0$ we also get

$$\tilde v(y) \geq \inf_{x\in\hat{\mathcal{T}}}[\hat c(x,y) - w_1(x)] = \inf_{x\in\mathcal{T}}\{b[c(x,y)-\lambda] - w_1(x)\} = -b\lambda + \inf_{x\in\mathcal{T}}[b\, c(x,y) - w_1(x)].$$

It follows that $(\tilde u, \tilde v) \in \mathsf{K}$ and

$$I \leq \int_{\mathcal{T}} \tilde u(x)\mu(dx) + \int_{\mathcal{T}} \tilde v(x)\nu(dx) \leq J.$$

Thus we conclude that $I = J$ and the theorem follows. ■

A.5 Proof for Corollary 3.5 in the main text

Proof. Notice that as w_i $(i = 1, 2)$ is b-Lipschitz, we have for every $x \in \mathcal{T}$ that

$$-\, w_i(x) \leq \inf_{y\in\mathcal{T}} \big[b\, d_{\mathcal{T}}(x,y) - w_i(y)\big]. \tag{6}$$

For each $(u, v) \in \mathsf{K}$, let

$$\begin{aligned} v^*(x) &:= \inf_{y\in\mathcal{T}} \big\{b[d_{\mathcal{T}}(x,y) - \lambda] - v(y)\big\} = -b\lambda + \inf_{y\in\mathcal{T}} \big[b\, d_{\mathcal{T}}(x,y) - v(y)\big] \geq u(x), \\ v^{**}(y) &:= \inf_{x\in\mathcal{T}} \big\{b[d_{\mathcal{T}}(x,y) - \lambda] - v^*(x)\big\} = -b\lambda + \inf_{x\in\mathcal{T}} \big[b\, d_{\mathcal{T}}(x,y) - v^*(x)\big] \geq v(y). \end{aligned}$$

By using $-b\lambda + \inf_{x\in\mathcal{T}}[b\, d_{\mathcal{T}}(x,y) - w_1(x)] \le v(y) \le w_2(y)$ and (6), we obtain for every $x \in \mathcal{T}$ that

$$v^*(x) \le -b\lambda - v(x) \le -\inf_{y\in\mathcal{T}}[b\, d_{\mathcal{T}}(x,y) - w_1(y)] \le w_1(x),$$
$$v^*(x) \ge -b\lambda + \inf_{y\in\mathcal{T}}\left[b\, d_{\mathcal{T}}(x,y) - w_2(y)\right] \ge -b\lambda - w_2(x).$$

We also have v^* is b-Lipschitz, i.e., $|v^*(x_1) - v^*(x_2)| \le b\, d_{\mathcal{T}}(x_1,x_2)$. Indeed, let $x_1, x_2 \in \mathcal{T}$. Then for any $\varepsilon > 0$, there exists $y_1 \in \mathcal{T}$ such that $b\, d_{\mathcal{T}}(x_1,y_1) - v(y_1) < v^*(x_1) + b\lambda + \varepsilon$. It follows that

$$v^*(x_2) - v^*(x_1) \le b\, d_{\mathcal{T}}(x_2,y_1) - v(y_1) + \varepsilon - [b\, d_{\mathcal{T}}(x_1,y_1) - v(y_1)] \le b\, d_{\mathcal{T}}(x_1,x_2) + \varepsilon.$$

Since this holds for every $\varepsilon > 0$, we get $v^*(x_2) - v^*(x_1) \le b\, d_{\mathcal{T}}(x_1,x_2)$. By interchanging the role of x_1 and x_2, we also obtain $v^*(x_1) - v^*(x_2) \le b\, d_{\mathcal{T}}(x_1,x_2)$. Thus, $|v^*(x_1) - v^*(x_2)| \le b\, d_{\mathcal{T}}(x_1,x_2)$. Hence, we have shown that $v^* \in \mathsf{L}'$ with

$$\mathsf{L}' := \Big\{f \in C(\mathcal{T}) :\, -b\lambda - w_2 \le f \le w_1,\ |f(x) - f(y)| \le b\, d_{\mathcal{T}}(x,y)\Big\}.$$

We next claim $v^{**} = -b\lambda - v^*$. For this, it is clear from the definition that $v^{**}(y) \le -b\lambda - v^*(y)$. On the other hand, from the Lipschitz property of v^* we obtain

$$-v^*(y) \le b\, d_{\mathcal{T}}(x,y) - v^*(x) \quad \forall x \in \mathcal{T},$$

which gives $-b\lambda - v^*(y) \le v^{**}(y)$. Thus, we conclude that $v^{**} = -b\lambda - v^*$ as claimed.

From these, we obtain that

$$\begin{aligned}\int_{\mathcal{T}} u(x)\mu(dx) + \int_{\mathcal{T}} v(x)\nu(dx) &\le \int_{\mathcal{T}} v^*(x)\mu(dx) + \int_{\mathcal{T}} v^{**}(x)\nu(dx)\\ &= \int_{\mathcal{T}} v^*(x)\mu(dx) - \int_{\mathcal{T}} v^*(x)\nu(dx) - b\lambda\nu(\mathcal{T})\\ &\le -b\lambda\nu(\mathcal{T}) + \sup\left\{\int_{\mathcal{T}} f(\mu-\nu) :\, f \in \mathsf{L}'\right\}.\end{aligned}$$

This together with Theorem 3.4 in the main text implies that $\mathsf{ET}_\lambda(\mu,\nu) \le -b\lambda\nu(\mathcal{T}) + \sup\left\{\int_{\mathcal{T}} f(\mu-\nu) :\, f \in \mathsf{L}'\right\}$. To prove the converse, let $f \in \mathsf{L}'$. Define $u := f$ and $v := -b\lambda - f$. Then, we have $u(x) \le w_1(x)$, $v(x) \le -b\lambda - [-b\lambda - w_2(x)] = w_2(x)$, and

$$v(x) \ge -b\lambda - w_1(x) \ge -b\lambda + \inf_{y\in\mathcal{T}}[b\, d_{\mathcal{T}}(x,y) - w_1(y)].$$

Also, the Lipschitz property of f gives

$$u(x) + v(y) = -b\lambda + f(x) - f(y) \le b[d_{\mathcal{T}}(x,y) - \lambda] \quad \forall x, y \in \mathcal{T}.$$

Thus $(u,v) \in \mathsf{K}$, and hence we obtain from Theorem 3.4 in the main text that

$$-b\lambda\nu(\mathcal{T}) + \int_{\mathcal{T}} f(\mu-\nu) = \int_{\mathcal{T}} u(x)\mu(dx) + \int_{\mathcal{T}} v(x)\nu(dx) \le \mathsf{ET}_\lambda(\mu,\nu).$$

As this holds for every $f \in \mathsf{L}'$, we get

$$-b\lambda\nu(\mathcal{T}) + \sup\left\{\int_{\mathcal{T}} f(\mu-\nu) :\, f \in \mathsf{L}'\right\} \le \mathsf{ET}_\lambda(\mu,\nu).$$

Thus, we have shown that

$$\mathsf{ET}_\lambda(\mu,\nu) = -b\lambda\nu(\mathcal{T}) + \sup\left\{\int_{\mathcal{T}} f(\mu-\nu) :\, f \in \mathsf{L}'\right\}. \tag{7}$$

Now consider $f = \tilde{f} - \frac{b\lambda}{2}$. Then, $f \in \mathsf{L}'$ if and only if $\tilde{f} \in \mathsf{L}$. Moreover,

$$\int_{\mathcal{T}} f(\mu-\nu) = -\frac{b\lambda}{2}\left[\mu(\mathcal{T}) - \nu(\mathcal{T})\right] + \int_{\mathcal{T}} \tilde{f}(\mu-\nu).$$

Therefore, the conclusion of the corollary follows from (7). ■

A.6 Proof for Proposition 3.7 in the main text

In order to prove Proposition 3.7 in the main text, we need the following auxiliary result.

Lemma A.1. *Assume that $w_1 > 0$ and $w_2 > 0$. Then, $d(\mu,\nu) = 0$ implies that $\mu = \nu$.*

Proof. Assume that $d(\mu,\nu) = 0$. Let γ^0 be an optimal plan for $\mathsf{ET}_\lambda(\mu,\nu)$, and set $m := \gamma^0(\mathcal{T} \times \mathcal{T})$. Then, $m \le \min\{\mu(\mathcal{T}), \nu(\mathcal{T})\}$, and hence we obtain from Problem (3) in the main text that

$$\int_{\mathcal{T}} w_1[1 - f_1(x)]\mu(dx) + \int_{\mathcal{T}} w_2[1 - f_2(x)]\nu(dx) + b \int_{\mathcal{T}\times\mathcal{T}} d_{\mathcal{T}}(x,y)\gamma^0(dx,dy)$$
$$= \mathsf{ET}_\lambda(\mu,\nu) + \lambda b m \le \mathsf{ET}_\lambda(\mu,\nu) + \frac{b\lambda}{2}\big[\mu(\mathcal{T}) + \nu(\mathcal{T})\big] = d(\mu,\nu) = 0.$$

Thus,

$$\int_{\mathcal{T}} w_1[1 - f_1(x)]\mu(dx) = \int_{\mathcal{T}} w_2[1 - f_2(x)]\nu(dx) = \int_{\mathcal{T}\times\mathcal{T}} d_{\mathcal{T}}(x,y)\gamma^0(dx,dy) = 0.$$

Since w_1 and w_2 are positive, it follows in particular that $f_1 = 1$ μ-a.e. and $f_2 = 1$ ν-a.e. That is, $\gamma_1^0 = \mu$ and $\gamma_2^0 = \nu$. Moreover, the above last identity implies that γ^0 is supported on the diagonal $(y = x)$. Therefore, for any continuous function φ on $\mathcal{T}$ we have

$$\int_{\mathcal{T}} \varphi(x)\mu(dx) = \int_{\mathcal{T}\times\mathcal{T}} \varphi(x)\gamma^0(dx,dy) = \int_{\mathcal{T}\times\mathcal{T}} \varphi(y)\gamma^0(dx,dy) = \int_{\mathcal{T}} \varphi(y)\nu(dy).$$

We thus conclude that $\mu = \nu$. ■

Proof. [Of Proposition 3.7 in the main text]

i) This follows immediately from Corollary 3.5 in the main text.

ii) By Corollary 3.5 in the main text, it is clear that $d(\mu,\nu) \ge 0$ and $d(\mu,\mu) = 0$. Also, if $d(\mu,\nu) = 0$, then by Lemma A.1, we have $\mu = \nu$. It is obvious that d satisfies the triangle inequality.

iii) Due to the assumption $w_1 = w_2$ we have $f \in \mathsf{L}$ if and only if $-f \in \mathsf{L}$. It follows that $d(\mu,\nu) = d(\nu,\mu)$. This together with ii) implies that $(\mathcal{M}(\mathcal{T}), d)$ is a metric space. Its completeness follows from [13, Proposition 4]. As a complete metric space, it is well known that $(\mathcal{M}(\mathcal{T}), d)$ is a geodesic space if and only if for every $\mu,\nu \in \mathcal{M}(\mathcal{T})$ there exists $\sigma \in \mathcal{M}(\mathcal{T})$ such that

$$d(\mu,\sigma) = d(\nu,\sigma) = \frac{1}{2}d(\mu,\nu).$$

To verify the latter, take $\sigma := \frac{\mu+\nu}{2}$. Then using Corollary 3.5 in the main text, we obtain

$$d(\mu,\sigma) = \frac{1}{2}\sup_{f\in\mathsf{L}} \int_{\mathcal{T}} f(\mu - \nu) = \frac{1}{2}d(\mu,\nu)$$

and

$$d(\nu,\sigma) = \frac{1}{2}\sup_{f\in\mathsf{L}} \int_{\mathcal{T}} f(\nu - \mu) = \frac{1}{2}d(\nu,\mu) = \frac{1}{2}d(\mu,\nu).$$

■

A.7 Proof for Proposition 3.8 in the main text

Proof. Observe that

$$\widetilde{\mathsf{ET}}^{\alpha}_{\lambda}(\mu,\nu) = -\frac{b\lambda}{2}\big[\mu(\mathcal{T}) + \nu(\mathcal{T})\big]$$
$$+ \sup\Big\{s[\mu(\mathcal{T}) - \nu(\mathcal{T})] : s \in \big[-\frac{b\lambda}{2} - w_2(r) + \alpha, w_1(r) + \frac{b\lambda}{2} - \alpha\big]\Big\}$$
$$+ \sup\Big\{\int_{\mathcal{T}} \Big[\int_{[r,x]} g(y)\omega(dy)\Big](\mu - \nu)(dx) : \|g\|_{L^1(\mathcal{T})} \le b\Big\}.$$

The first supremum equals to $[w_1(r) + \frac{b\lambda}{2} - \alpha][\mu(\mathcal{T}) - \nu(\mathcal{T})]$ if $\mu(\mathcal{T}) \geq \nu(\mathcal{T})$ and equals to $-[w_2(r) + \frac{b\lambda}{2} - \alpha][\mu(\mathcal{T}) - \nu(\mathcal{T})]$ if $\mu(\mathcal{T}) < \nu(\mathcal{T})$. On the other hand, by the same arguments as in [4, p.575-576], we see that the second supremum equals to $\int_{\mathcal{T}} |\mu(\ \ (x)) - \nu(\ \ (x))|\, \omega(dx)$. Putting them together, we obtain the desired formula for $\widetilde{\mathsf{ET}}^{\alpha}_{\lambda}(\mu, \nu)$. ■

A.8 Proof for Proposition 3.9 in the main text

Proof. The inequality $\mathsf{ET}_\lambda(\mu, \nu) \leq \widetilde{\mathsf{ET}}^{0}_{\lambda}(\mu, \nu)$ holds due to $\mathsf{L} \subset \mathsf{L}_0$ and Corollary 3.5 in the main text. Next, let

$$2bL(\mathcal{T}) \leq \alpha \leq \frac{1}{2}[b\lambda + w_1(r) + w_2(r)].$$

Then, thanks to Corollary 3.5 in the main text, the stated lower bound will follow if $\mathsf{L}_\alpha \subset \mathsf{L}$. This is achieved if we can show that any $f \in \mathsf{L}_\alpha$ satisfies $-w_2 - \frac{b\lambda}{2} \leq f \leq w_1 + \frac{b\lambda}{2}$. Indeed, for such function f we have

$$f(x) = s + \int_{[r,x]} g(y)\omega(dy),$$

with $s \in \big[- w_2(r) - \frac{b\lambda}{2} + \alpha, w_1(r) + \frac{b\lambda}{2} - \alpha\big]$ and $\|g\|_{L^1(\mathcal{T})} \leq b$. This together the b-Lipschitz property of w_1, w_2 gives for every $x \in \mathcal{T}$ that

$$f(x) \leq s + \|g\|_{L^1(\mathcal{T})}\omega([r, x]) \leq w_1(r) + \frac{b\lambda}{2} - \alpha + bL(\mathcal{T}) \leq w_1(x) + \frac{b\lambda}{2} - \alpha + 2bL(\mathcal{T}) \leq w_1(x) + \frac{b\lambda}{2}$$

and

$$\begin{aligned} f(x) &\geq s - \|g\|_{L^1(\mathcal{T})}\omega([r, x]) \geq -w_2(r) - \frac{b\lambda}{2} + \alpha - bL(\mathcal{T}) \\ &\geq -w_2(x) - \frac{b\lambda}{2} + \alpha - 2bL(\mathcal{T}) \geq -w_2(x) - \frac{b\lambda}{2}. \end{aligned}$$

It follows that $f \in \mathsf{L}$. Thus, $\mathsf{L}_\alpha \subset \mathsf{L}$ and we obtain

$$\widetilde{\mathsf{ET}}^{\alpha}_{\lambda}(\mu, \nu) \leq \mathsf{ET}_\lambda(\mu, \nu).$$

■

A.9 Proof of Proposition 3.10 in the main text

We begin with the following auxiliary result.

Lemma A.2. *Let $\mu, \nu \in \mathcal{M}(\mathcal{T})$. Then, $\mu = \nu$ if and only if $\mu(\ \ (x)) = \nu(\ \ (x))$ for every x in $\mathcal{T}$.*

Proof. It is obvious that $\mu = \nu$ implies that $\mu(\ \ (x)) = \nu(\ \ (x))$ for every x in $\mathcal{T}$. Now assume that $\mu(\ \ (x)) = \nu(\ \ (x))$ for every x in $\mathcal{T}$. We first claim that $\mu(\{a\}) = \nu(\{a\})$ for any $a \in \mathcal{T}$. Indeed, if a is not a node then we have $\ \ (a) \setminus \ \ (a_n) \downarrow \{a\}$, where $\{a_n\}_{n=1}^{\infty}$ is a sequence of distinct points on the same edge as a and converges to a from below. Hence,

$$\mu(\{a\}) = \lim_{n\to\infty} \big[\mu(\ \ (a)) - \mu(\ \ (a_n))\big] = \lim_{n\to\infty} \big[\nu(\ \ (a)) - \nu(\ \ (a_n))\big] = \nu(\{a\}).$$

In case a is a common node for edges $e_1, ..., e_k$, then we have $\ \ (a) \setminus \cup_{i=1}^{k}\ \ (a_n^i) \downarrow \{a\}$, where $\{a_n^i\}_{n=1}^{\infty}$ is a sequence of distinct points on edge e_i that converges to a from below. Then, we obtain

$$\mu(\{a\}) = \lim_{n\to\infty} \Big[\mu(\ \ (a)) - \sum_{i=1}^{k} \mu(\ \ (a_n^i))\Big] = \lim_{n\to\infty} \Big[\nu(\ \ (a)) - \sum_{i=1}^{k} \nu(\ \ (a_n^i))\Big] = \nu(\{a\}).$$

Thus, the claim is proved. On the other hand, for any points x, y belonging to the same edge

$$\mu([x, y)) = \mu(\ \ (x)) - \mu(\ \ (y)) = \nu(\ \ (x)) - \nu(\ \ (y)) = \nu([x, y)).$$

Thus, by combining them, we infer further that $\mu([x, y]) = \nu([x, y])$ for any $x, y \in \mathcal{T}$. It follows that $\mu = \nu$, and the proof is complete. ■

Proof. [Of Proposition 3.10 in the main text] We note first that the quantity d_α depends only on the values of the weights at the root r of the tree. This comes from the fact that only $w_1(r)$ and $w_2(r)$ are used in the definition of L_α. The proofs of i) and iii) are exactly the same as that of Proposition 3.7 in the main text.

For ii), it follows from the fact

$$d_\alpha(\mu,\nu) = \sup\left\{\int_{\mathcal{T}} f(\mu-\nu) : f \in \mathsf{L}_\alpha\right\}$$

that $d_\alpha(\mu,\nu) \geq 0$, $d_\alpha(\mu,\mu) = 0$, and d_α satisfies the triangle inequality. Also, if $d_\alpha(\mu,\nu) = 0$, then by Proposition 3.8 in the main text, we get

$$\left[w_i(r) + \frac{b\lambda}{2} - \alpha\right]|\mu(\mathcal{T}) - \nu(\mathcal{T})| + \int_{\mathcal{T}} |\mu(\ (x)) - \nu(\ (x))|\,\omega(dx) = 0.$$

As $\left[w_i(r) + \frac{b\lambda}{2} - \alpha\right] > 0$ by the assumption, we must have $\mu(\mathcal{T}) = \nu(\mathcal{T})$ and $\int_{\mathcal{T}} |\mu(\ (x)) - \nu(\ (x))|\,\omega(dx) = 0$. Therefore, $\mu(\ (x)) = \nu(\ (x))$ for every $x \in \mathcal{T}$. By using Lemma A.2, we then conclude that $\mu = \nu$.

Alternatively, we can argue as follows. Assume that $d_\alpha(\mu,\nu) = 0$. Since

$$\mathsf{L}_\alpha \supset \tilde{\mathsf{L}} := \left\{f : -w_2(r) - \frac{b\lambda}{2} + \alpha \leq f(x) \leq w_1(r) + \frac{b\lambda}{2} - \alpha,\ \|f\|_{Lip(\mathcal{T})} \leq b\right\},$$

we have

$$0 \leq \sup_{f\in\tilde{\mathsf{L}}} \int_{\mathcal{T}} f(\mu-\nu) \leq d_\alpha(\mu,\nu) = 0.$$

Thus, $\sup_{f\in\tilde{\mathsf{L}}}\int_{\mathcal{T}} f(\mu-\nu) = 0$. Then, by applying Corollary 3.5 in the main text and Lemma A.1 for constant weights $\bar{w}_1 := w_1(r) + \frac{b\lambda}{2} - \alpha > 0$ and $\bar{w}_2 := w_2(r) + \frac{b\lambda}{2} - \alpha > 0$, we obtain that $\mu = \nu$. ■

A.10 Proof for Proposition 3.11 in the main text

Proof. Let $\mathbf{f}(x_i,x_j) = \bar{a}(x_i + x_j)$ for $\bar{a}, x_i, x_j \in \mathsf{R}$. We first prove that $\mathbf{f}$ is negative definite.

For all $n \geq 2$, for $c_1, c_2, \ldots, c_n$ such that $\sum_{i=1}^n c_i = 0$. Given $x_1, x_2, \ldots, x_n \in \mathsf{R}$, we have

$$\sum_{i,j} c_i c_j \mathbf{f}(x_i,x_j) = \sum_{i,j} c_i c_j \bar{a} x_i + \sum_{i,j} c_i c_j \bar{a} x_j \leq 0.$$

Therefore, $\mathbf{f}$ is negative definite.

From Proposition 3.8 in the main text, we have

$$\widetilde{\mathsf{ET}}_\lambda^\alpha(\mu,\nu) = -\frac{b\lambda}{2}\left[\mu(\mathcal{T}) + \nu(\mathcal{T})\right] + \left[w_i(r) + \frac{b\lambda}{2} - \alpha\right]|\mu(\mathcal{T}) - \nu(\mathcal{T})| + \int_{\mathcal{T}} |\mu(\ (x)) - \nu(\ (x))|\,\omega(dx).$$

The first term is negative definite since $\mathbf{f}$ is negative definite. Additionally, the second and third terms are equivalent to the weighted ℓ_1 distance with nonnegative weights (i.e., $\alpha \leq w_i(r) + \frac{b\lambda}{2}$ and lengths of edges in tree $\mathcal{T}$ are nonnegative). Therefore, the second and third terms are also negative definite. Hence, $\widetilde{\mathsf{ET}}_\lambda^\alpha$ is negative definite.

From Proposition 3.10 in the main text, we have

$$d_\alpha(\mu,\nu) = \widetilde{\mathsf{ET}}_\lambda^\alpha(\mu,\nu) + \frac{b\lambda}{2}\left[\mu(\mathcal{T}) + \nu(\mathcal{T})\right].$$

Both terms are negative definite. Therefore, d_α is also negative definite.

■

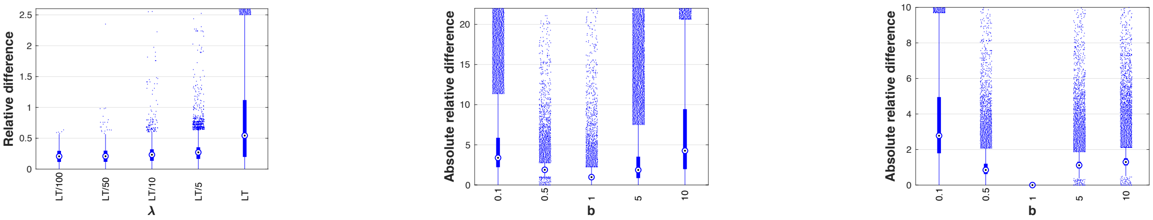
B Further Experimental Results

In this section, we illustrate further experimental results.

B.1 Further Results on the Efficient Approximation of $\widetilde{\mathrm{ET}}_{\lambda}^{0}$ for ET_{λ}

In this section, we consider some further setups.

Change λ. In Figure 1a, we use the same setup as in Figure 2 in the main text, but set the Lipschitz $a_1 = \frac{b}{2} = 0.5$ for w_1, w_2. It shows that when λ is increased, $\widetilde{\mathrm{ET}}_{\lambda}^{0}$ is farther to ET_{λ}.



(a) (b) (c)

Figure 1: In (a), an illustration about the relative difference between $\widetilde{\mathrm{ET}}_{\lambda}^{0}$ and ET_{λ} w.r.t. λ. LT is the longest path from a root to a node in tree $\mathcal{T}$ ($\mathrm{LT} := L_{\mathcal{T}}$). Lipchitz for functions w_1, w_2 is $a_1 = 0.5$ (where $b = 1$). In (b, c), an illustration about the absolute relative difference between $\widetilde{\mathrm{ET}}_{\lambda}^{0}$ and ET_{λ}, i.e., $(\widetilde{\mathrm{ET}}_{\lambda}^{0} - \mathrm{ET}_{\lambda}) / |\mathrm{ET}_{\lambda}|$, w.r.t. b. For (b), the weight functions w_1, w_2 are set constants ($a_1 = 0$, or $w_1 = w_2 = a_0$) while for (c), the weight functions w_1, w_2 are set with largest Lipchitz ($a_1 = b$).

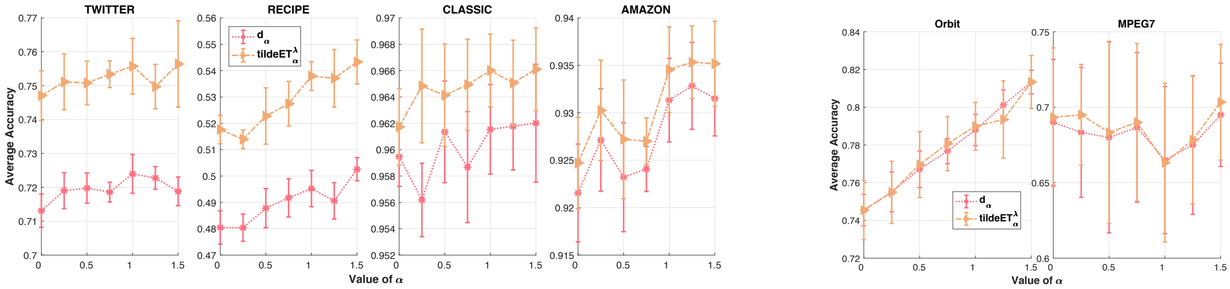
Change b. We consider 2 following cases:

- **For constant functions w_1, w_2 (with $a_1 = 0$).** We use the same setup as in Figure 1 in the main text, but with constant functions for w_1, w_2 (i.e., $a_1 = 0$, or $w_1 = w_2 = a_0$), and change b. We set $\lambda = a_0 = 1$. In Figure 1b, we illustrate that when the regularization b between entropy and partial matching is farther to 1 (one of the two terms is more weighted, see Equation (2) in the main text), $\widetilde{\mathrm{ET}}_{\lambda}^{0}$ is farther to ET_{λ}.

- **For functions w_1, w_2 with largest Lipschitz $a_1 = b$.** We use the same setup as in Figure 1b, but with $a_1 = b$. Figure 1c shows similar results as in Figure 1b for $a_1 = 0$. For the largest Lipchitz for functions w_1, w_2 (i.e., $a_1 = b$), but for $b = a_0 = 1$, $\widetilde{\mathrm{ET}}$ is almost identical to KT, but they are different when when the regularization b between entropy and partial matching is farther to 1 (one of the two terms is more weighted, see Equation (2) in the main text).

B.2 Further Results w.r.t. α

We illustrate further SVM results of d_{α} and $\widetilde{\mathrm{ET}}_{\lambda}^{\alpha}$ w.r.t. value of α in TWITTER, RECIPE, CLASSIC, AMAZON datasets in Figure 2a, and in Orbit, MPEG7 datasets in Figure 2b. The value of α may affect performances of d_{α} and $\widetilde{\mathrm{ET}}_{\lambda}^{\alpha}$ in some datasets (e.g., RECIPE, AMAZON datasets for document classification, and Orbit dataset in TDA), but may not sensitive in some other datasets (e.g., TWITTER, CLASSIC datasets for document classification, and MPEG7 dataset in TDA). Therefore, although $\alpha = 0$ gives $\widetilde{\mathrm{ET}}_{\lambda}^{\alpha}$ good property as in Proposition 3.9 in the main text (upper bound for ET_{λ}), there is a possibility to choose suitable value for α (e.g., via cross validation) to improve performances of d_{α} and $\widetilde{\mathrm{ET}}_{\lambda}^{\alpha}$ for some certain datasets.



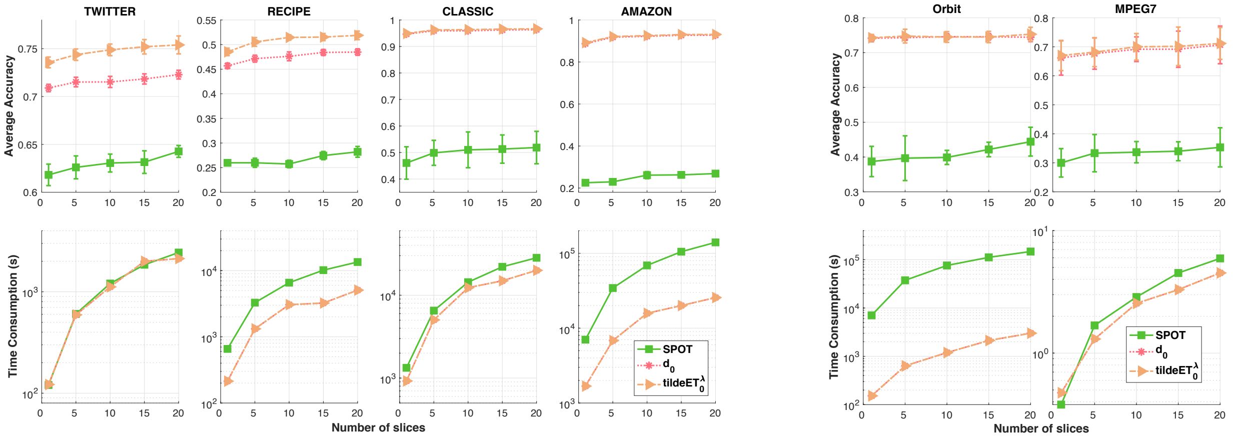
(a) In TWITTER, RECIPE, CLASSIC, AMAZON datasets.

(b) In Orbit, MPEG7 datasets.

Figure 2: SVM results of d_{α} and $\widetilde{\mathrm{ET}}_{\lambda}^{\alpha}$ w.r.t. value of α with 10 tree slices.

B.3 Further Results w.r.t. the Number of (Tree) Slices

Similar as Figure 6 in the main text, we illustrate further SVM results and time consumption for corresponding kernel matrices for document classification (e.g., `TWITTER, RECIPE, CLASSIC, AMAZON` datasets) and TDA (`Orbit, MPEG7` datasets in Figure 3a and Figure 3b respectively. For a trade-off between performances and time consumption, one can choose about $n_s = 10$ slices in applications.



(a) In `TWITTER, RECIPE, CLASSIC, AMAZON` datasets.

(b) In `Orbit, MPEG7` datasets.

Figure 3: SVM results and time consumption for corresponding kernel matrices w.r.t. the number of (tree) slices.

B.4 Further Results w.r.t. Parameters of Tree Metric Sampling

Document classification.

- In Figure 4a, Figure 4b, Figure 4c, Figure 4d, we illustrate further SVM results and time consumption for corresponding kernel matrices of d_0 in `TWITTER, RECIPE, CLASSIC, AMAZON` datasets respectively w.r.t. different parameters for clustering-based tree metric sampling such as the predefined tree deepest level $H_{\mathcal{T}}$, and number of tree branches κ which is the number of clusters in the farthest-point clustering.
- In Figure 5a, Figure 5b, Figure 5c, Figure 5d, we illustrate further SVM results and time consumption for corresponding kernel matrices of $\widetilde{\mathrm{ET}}_{\lambda}^{0}$ in `TWITTER, RECIPE, CLASSIC, AMAZON` datasets respectively w.r.t. different parameters for clustering-based tree metric sampling such as the predefined tree deepest level $H_{\mathcal{T}}$, and number of tree branches κ which is the number of clusters in the farthest-point clustering.

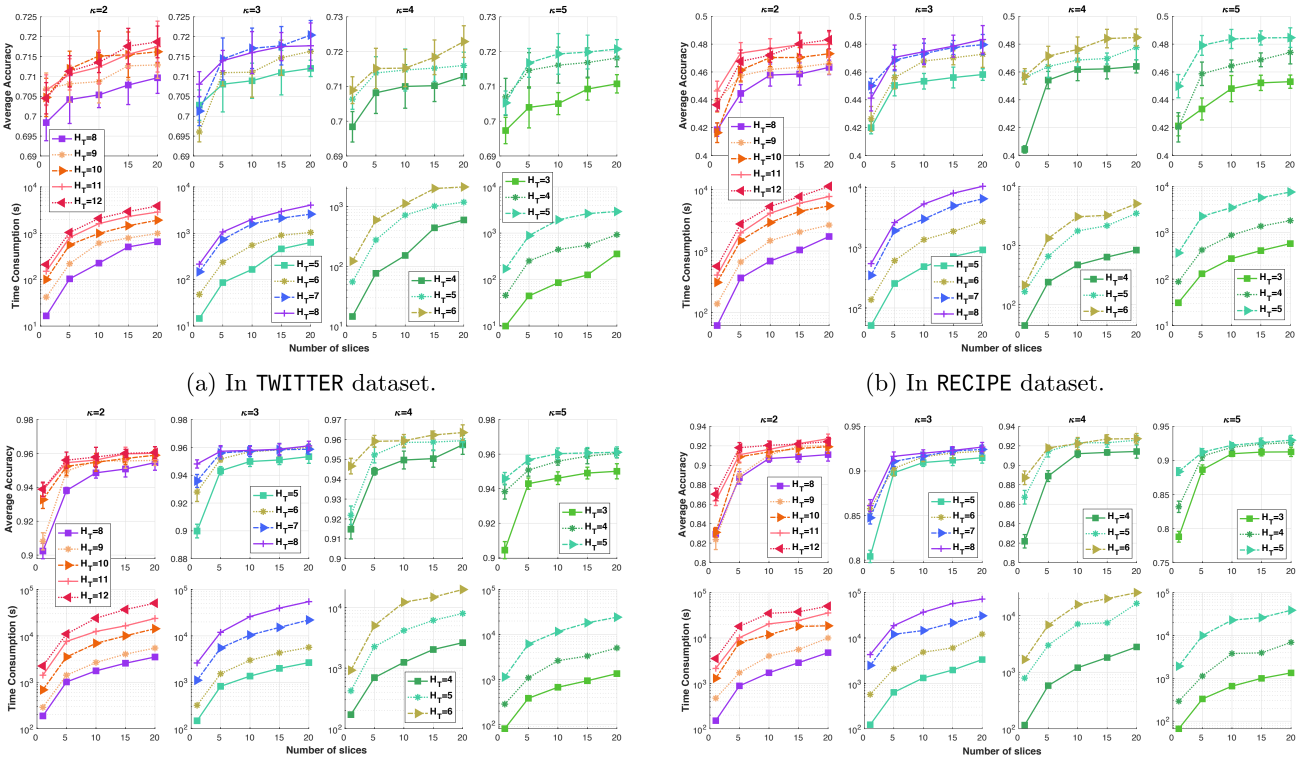
TDA.

- In Figure 6a, we illustrate further SVM results and time consumption for corresponding kernel matrices of d_0 in `Orbit, MPEG7` datasets w.r.t. different parameters for partition-based tree metric sampling such as the predefined tree deepest level $H_{\mathcal{T}}$.
- In Figure 6b, we illustrate further SVM results and time consumption for corresponding kernel matrices of $\widetilde{\mathrm{ET}}_{\lambda}^{0}$ in `Orbit, MPEG7` datasets w.r.t. different parameters for partition-based tree metric sampling such as the predefined tree deepest level $H_{\mathcal{T}}$.

Similar as in [12] (tree metric sampling for tree-sliced-Wasserstein in applications), we also observed that the default parameters (e.g., the predefined deepest level $H_{\mathcal{T}} = 6$, and the tree branches $\kappa = 4$—the number of clusters in the farthest-point clustering) is a reasonable choice to trade-off about performances and time consumption. With these default parameters, sampled trees contains about 4000 nodes.

C Further Details and Discussions

In this section, we give further details about experiments, some brief reviews about important aspects used in our work and discuss other relations to other work.



(a) In TWITTER dataset.

(b) In RECIPE dataset.

(c) In CLASSIC dataset.

(d) In AMAZON dataset.

Figure 4: SVM results and time consumption for corresponding kernel matrices of d_0 w.r.t. different parameters for clustering-based tree metric sampling (predefined tree deepest level $H_{\mathcal{T}}$, and number of tree branches κ—the number of clusters in the farthest-point clustering.).

C.1 More Details about Experiments

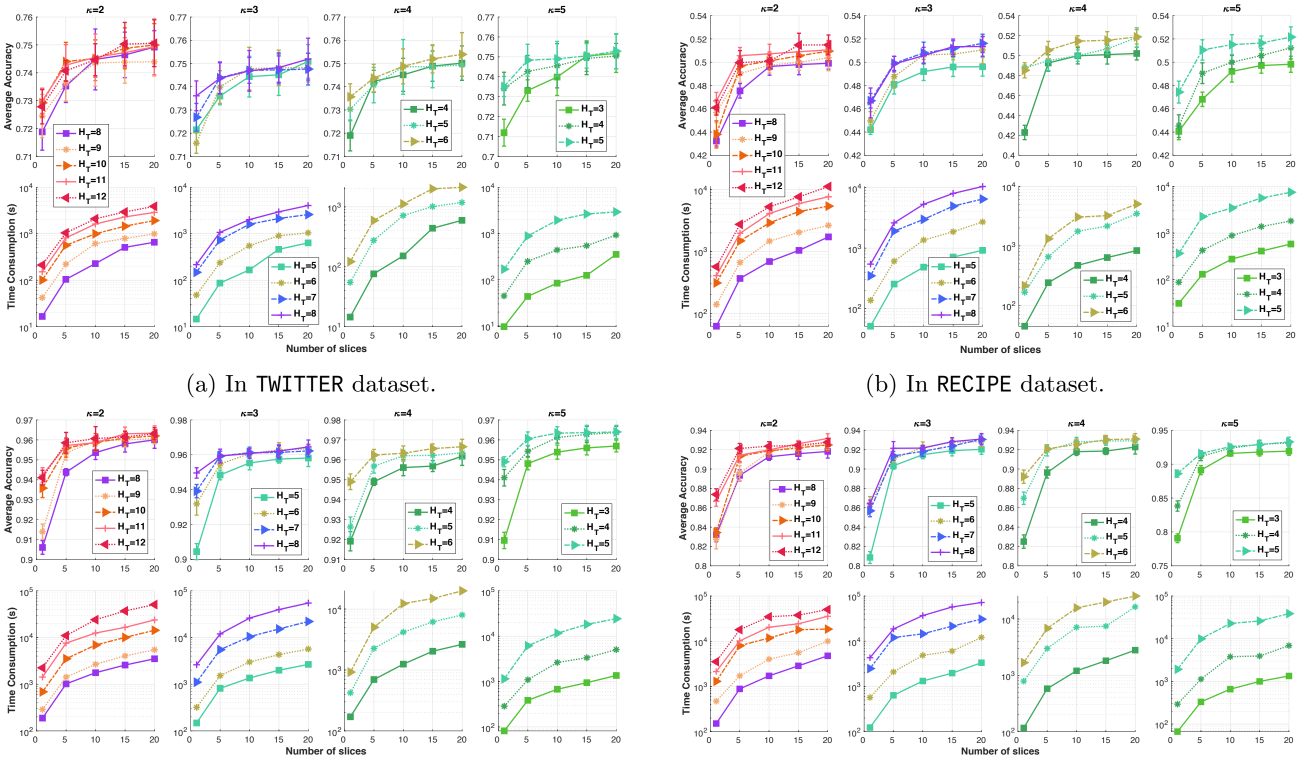
In this section, we give further details about softwares, datasets and experimental setups.

For softwares.

- For experiments in topological data analysis, we used DIPHA toolbox, available at `https://github.com/DIPHA/dipha`, to extract persistence diagrams.
- For the standard complete optimal transport (OT) problem (e.g., KT in our work which we used to compute the corresponding ET_λ), we used a fast OT implementation, available at `https://github.com/gpeyre/2017-ot-beginners/tree/master/matlab/mexEMD`. It is about 4 times faster than the popular mex-file with Rubner's implementation in C, available at `http://robotics.stanford.edu/~rubner/emd/default.htm`.
- For tree metric sampling, we used the MATLAB implementation, available at `https://github.com/lttam/TreeWasserstein`. We directly used this code for clustering-based tree metric sampling, and adapted it into its special case partition-based tree metric sampling.
- For Sinkhorn-based approach for unbalanced OT (Sinkhorn-UOT), we used the MATLAB implementation, available at `https://github.com/gpeyre/2017-MCOM-unbalanced-ot`.
- For sliced partial optimal transport (SPOT), we adapt the C++ implementation, available at `https://github.com/nbonneel/spot`, into MATLAB.

For datasets.

- For document datasets (e.g., TWITTER, RECIPE, CLASSIC, AMAZON), they are available at `https://github.com/mkusner/wmd`.



(a) In TWITTER dataset.

(b) In RECIPE dataset.

(c) In CLASSIC dataset.

(d) In AMAZON dataset.

Figure 5: SVM results and time consumption for corresponding kernel matrices of $\widetilde{\mathrm{ET}}^{0}_{\lambda}$ w.r.t. different parameters for clustering-based tree metric sampling (predefined tree deepest level $H_{\mathcal{T}}$, and number of tree branches κ—the number of clusters in the farthest-point clustering.).

- For Orbit dataset, we follow the procedure, detailed in [1] to generate the dataset.
- For MPEG7 dataset, it is available at http://www.imageprocessingplace.com/downloads_V3/root_downloads/image_databases/MPEG7_CE-Shape-1_Part_B.zip, then we follow [12] to extract the 10-class subset of the dataset.
- For granular packing system and SiO_2 datasets, one may access to them by contacting the corresponding authors.

For experimental setups. We further clarify some details about experimental setup.

As mentioned in the main text, for d_0 and $\widetilde{\mathrm{ET}}^{0}_{\lambda}$, we choose the weight functions for w_1, w_2 as

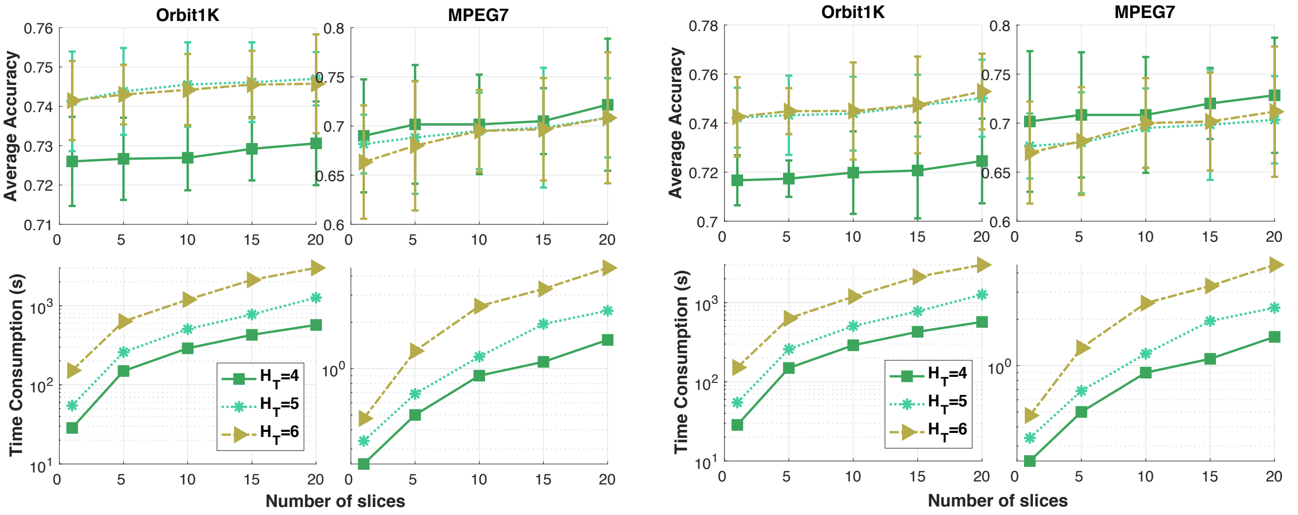
$$w_1(x) = w_2(x) = a_1 d_{\mathcal{T}}(r, x) + a_0,$$

where r is the root of tree $\mathcal{T}$, we set $\lambda = b = 1$, $a_0 = 1$. Following §5.1 in the main text, we set $a_1 = b = 1$. As in §3.2 in the main text, $\alpha \in \left[0, \frac{1}{2}\left(b\lambda + w_1(r) + w_2(r)\right)\right]$. Thus, $\alpha \in [0, \frac{3}{2}]$ in our experiments (see more experiment results with different values of α in §B.2). We used $n_s = 10$ (tree) slices for d_0, $\widetilde{\mathrm{ET}}^{0}_{\lambda}$ and SPOT. For tree metric sampling, we used the default hyperparameters, the predefined tree deepest level $H_{\mathcal{T}} = 6$, and the tree branches $\kappa = 4$—the number of clusters used in the farthest-point clustering.

C.2 Some Brief Reviews

In this section, we give some brief reviews (or more referred details) about some important aspects in our work.

For kernels. We review some important definitions (e.g., positive/negative definite kernels [2]) and theorems (e.g., Theorem 3.2.2 in [2]) about kernels used in our work.



(a) For d_0.

(b) For $\widetilde{\mathrm{ET}}^0_\lambda$.

Figure 6: SVM results and time consumption for corresponding kernel matrices in `Orbit, MPEG7` datasets w.r.t. different parameters for partition-based tree metric sampling (predefined tree deepest level $H_\mathcal{T}$).

- **Positive definite kernels [2, p.66–67].** A kernel function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathsf{R}$ is positive definite if $\forall n \in \mathsf{N}^*, \forall x_1, x_2, ..., x_n \in \mathcal{X}$, we have

$$\sum_{i,j} c_i c_j k(x_i, x_j) \geq 0, \qquad \forall c_i \in \mathsf{R}.$$

- **Negative definite kernels [2, p.66–67].** A kernel function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathsf{R}$ is negative definite if $\forall n \geq 2, \forall x_1, x_2, ..., x_n \in \mathcal{X}$, we have

$$\sum_{i,j} c_i c_j k(x_i, x_j) \leq 0, \qquad \forall c_i \in \mathsf{R} \text{ s.t. } \sum_i c_i = 0.$$

- **Theorem 3.2.2 in [2, p.74] for kernels.** If κ is a *negative definite* kernel, then $\forall t > 0$, kernel

$$k_t(x, z) := \exp\left(-t\kappa(x, z)\right)$$

is positive definite.

For tree metric sampling. The tree metric sampling is described in details in [12][S4]. Le et al. [12] also reviewed the details of the farthest-point clustering in §4.2 in the supplementary, discussed about thee quantization/clustering sensitivity problems of tree metric sampling in §5 in the supplementary. Tree metric sampling is also leveraged in other advanced OT problems, e.g., tree-Wasserstein barycenter [10], and a variant of Gromov-Wasserstein (i.e., alignment problems for probability measures having supports in different spaces) [9].

For persistence diagrams and related mathematical definitions in topological data analysis. We refer the reader into [8, §2] for a review about mathematical framework for persistence diagrams (e.g., persistence diagrams, filtrations, persistent homology).

C.3 Discussions about Other Relations to Other Work

We note that ultrametric (i.e., non-Archimedean metric, or isosceles metric) and its special case—binary metric are tree metrics [12]. Additionally, a metric for points in a line (e.g., in 1-dimensional projections for supports in SPOT, or SW), or in 1-dimensional manifold (e.g., in 1-dimensional manifold projections for supports in generalized SW [7]) is also a tree metric since we have a corresponding tree as a chain of these points.

We also list some other studies related to OT problem with tree metrics as follows: (i) Kloeckner [6] derived geometric properties of OT space for measures on an ultrametric space, (ii) Sommerfeld and Munk [14] studied statistical inferences for OT on finite spaces including tree metrics.

We note that we consider the ***discrete*** measures in our work (e.g., empirical measures). The closed-form formulation of our regularized entropy partial transport (EPT) $\widetilde{\mathrm{ET}}_{\lambda}^{\alpha}$ in Equation (8) in the main text is for *general* discrete nonnegative measures having different masses. To our knowledge, the proposed regularized EPT (i.e., $\widetilde{\mathrm{ET}}_{\lambda}^{\alpha}$ in Equation (8) in the main text) is the first approach that yields a closed-form solution among available variants of unbalanced OT for discrete measures. In the context of unbalanced OT for ***continuous*** measures (e.g., probability measures are scaled by positive constants), Janati et al. [5] recently showed that entropic optimal transport for unbalanced *Gaussian* measures (i.e., Gaussian measures are scaled by different positive constants) has a closed-form solution.

References

[1] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(1):218–252, 2017.

[2] C. Berg, J. P. R. Christensen, and P. Ressel, editors. *Harmonic analysis on semigroups*. Springer-Verglag, New York, 1984.

[3] Luis A Caffarelli and Robert J McCann. Free boundaries in optimal transport and monge-ampere obstacle problems. *Annals of mathematics*, pages 673–730, 2010.

[4] Steven N Evans and Frederick A Matsen. The phylogenetic kantorovich–rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):569–592, 2012.

[5] Hicham Janati, Boris Muzellec, Gabriel Peyré, and Marco Cuturi. Entropic optimal transport between (unbalanced) gaussian measures has a closed form. In *Advances in neural information processing systems*, 2020.

[6] Benoît R Kloeckner. A geometric study of Wasserstein spaces: ultrametrics. *Mathematika*, 61(1):162–178, 2015.

[7] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced wasserstein distances. In *Advances in Neural Information Processing Systems*, pages 261–272, 2019.

[8] Genki Kusano, Kenji Fukumizu, and Yasuaki Hiraoka. Kernel method for persistence diagrams via kernel embedding and weight factor. *The Journal of Machine Learning Research*, 18(1):6947–6987, 2017.

[9] Tam Le, Nhat Ho, and Makoto Yamada. Flow-based alignment approaches for probability measures in different spaces. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2021.

[10] Tam Le, Viet Huynh, Nhat Ho, Dinh Phung, and Makoto Yamada. On scalable variant of wasserstein barycenter. *arXiv preprint arXiv:1910.04483*, 2019.

[11] Tam Le and Truyen Nguyen. Entropy partial transport with tree metrics: Theory and practice. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2021.

[12] Tam Le, Makoto Yamada, Kenji Fukumizu, and Marco Cuturi. Tree-sliced variants of Wasserstein distances. In *Advances in neural information processing systems*, pages 12283–12294, 2019.

[13] Benedetto Piccoli and Francesco Rossi. Generalized wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211(1):335–358, 2014.

[14] Max Sommerfeld and A. Munk. Inference for empirical wasserstein distances on finite spaces. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 80:219–238, 2016.