Supplementary Material: Label Distribution Learning by Optimal Transport

Peng Zhao and **Zhi-Hua Zhou**
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210023, China
{zhaop, zhouzh}@lamda.nju.edu.cn

Abstract

This is the supplementary material for the paper *Label Distribution Learning by Optimal Transport* (Zhao and Zhou 2018), including proofs of the theorems and lemmas in the main paper.

Review of Optimal Transport Distance

In this part, we review some basic concepts and properties for optimal transport distance.

Definition 1. *(Transport Polytope) For two probability vectors r and c in the simplex Σ_d, we write $U(r, c)$ for the transport polytope of r and c, namely the polyhedral set of $d \times d$ matrices,*

$$U(r, c) := \{P \in \mathbb{R}_+^{d\times d} | P\mathbf{1}_d = r, P^{\mathrm{T}}\mathbf{1}_d = c\}. \quad (1)$$

Definition 2. *(Optimal Transport) Given a $d \times d$ cost matrix M, the total cost of mapping from r to c using a transport matrix (or coupling probability) P can be quantified as $\langle P, M\rangle$. The optimal transport (OT) problem is defined as,*

$$d_M(r, c) := \min_{P\in U(r,c)} \langle P, M\rangle. \quad (2)$$

Theorem 1. *(Optimal Transport Distance) d_M defined in* (2) *is a distance on Σ_d whenever M is a metric matrix.*

Theorem 1 is proved by gluing lemma, and a detailed proof could be found in Chapter 6 in the seminal book (Villani 2008).

Proof of Optimal Transport with a Pseudo-Metric Cost

In this part, we will prove that for optimal transport with a pseudo-metric cost matrix, it preserves the sub-additivity property, which plays a key role in measuring difference between prediction and groundtruth. Meanwhile, it is sufficient to make it a strict distance by multiplying d_M by $\mathbf{1}_{r\neq c}$.

The proof here is similar to proofs in papers (Cuturi 2013; Cuturi and Avis 2014), we provide a detailed proof as follows for self-containedness.

Theorem 2. *For a pseudo-metric M and probability distributions $r, c \in \Sigma_d$, the function $(r, c) \to \mathbf{1}_{r\neq c} d_M(r, c)$ satisfies all four distance axioms, i.e., non-negativity, symmetry, definiteness and sub-additivity (triangle inequality).*

Proof. Non-negativity is easy to prove: since the coupling matrix P and cost matrix M are nonnegative. Besides, by the symmetry of M, d_M is itself symmetric in its two arguments. Also, the definiteness is a direct result of the $\mathbf{1}_{r\neq c}$ term in function definition. The main point is to prove sub-additivity.

Let x, y, z be three elements in Σ_d. Let $P \in U(x, y)$ and $Q \in U(y, z)$ be two optimal solutions of the transport problems $d_M(x, y)$ and $d_M(y, z)$. Let T be a $d \times d \times d$ tensor,

$$T_{ijk} = \begin{cases} \frac{p_{ij}q_{jk}}{y_j} & \text{when } y_j \neq 0 \\ 0 & \text{when } y_j = 0 \end{cases}$$

Define $R \triangleq [r_{ik}]$, where $r_{ik} = \sum_{j=1}^{d} T_{ijk}$. Then, R is the coupling set of x and z, i.e., $R \in U(x, z)$. Indeed,

$$\begin{aligned} \sum_{i=1}^{d}\sum_{j=1}^{d} T_{ijk} &= \sum_{j=1}^{d}\sum_{i=1}^{d} \frac{p_{ij}q_{jk}}{y_j} = \sum_{j=1}^{d} \frac{q_{jk}}{y_j} \sum_{i=1}^{d} p_{ij} \\ &= \sum_{j=1}^{d} \frac{q_{jk}}{y_j} y_j = \sum_{j=1}^{d} q_{jk} = z_k \\ \sum_{k=1}^{d}\sum_{j=1}^{d} T_{ijk} &= \sum_{j=1}^{d}\sum_{k=1}^{d} \frac{p_{ij}q_{jk}}{y_j} = \sum_{j=1}^{d} \frac{p_{ij}}{y_j} \sum_{i=1}^{d} q_{jk} \\ &= \sum_{j=1}^{d} \frac{p_{ij}}{y_j} y_j = \sum_{j=1}^{d} p_{ij} = x_i \end{aligned}$$

Now, we proceed to prove the sub-additivity,

$$\begin{aligned} d_M(x, z) &= \min_{S\in U(x,z)} \langle S, M\rangle \\ &\leq \langle R, M\rangle = \sum_{i=1}^{d}\sum_{k=1}^{d} M_{ik} r_{ik} \\ &= \sum_{i=1}^{d}\sum_{k=1}^{d} M_{ik} \sum_{j=1}^{d} \frac{p_{ij}q_{jk}}{y_j} \end{aligned}$$

$$
\begin{aligned}
&\le \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d}(M_{ij}+M_{jk})\frac{p_{ij}q_{jk}}{y_j} \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d} M_{ij}p_{ij}\sum_{k=1}^{d}\frac{q_{jk}}{y_j} + \sum_{j=1}^{d}\sum_{k=1}^{d} M_{jk}q_{jk}\sum_{i=1}^{d}\frac{p_{ij}}{y_j} \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d} M_{ij}p_{ij} + \sum_{j=1}^{d}\sum_{k=1}^{d} M_{jk}q_{jk} \\
&= d_M(x,y) + d_M(y,z)
\end{aligned}
$$

where we can see that the sub-additivity of M plays an important role in the proof (in the second inequality). □

Proof of Risk Bounds Analysis

In this part, we provide proof of risk bound analysis. To simplify the presentation, we introduce the *Sinkhorn loss* as:

$$
\ell(h(\mathbf{x}),\mathbf{y}) := d_M^{\lambda}(h(\mathbf{x}),\mathbf{y}) = \langle P^{\lambda}, M\rangle \tag{3}
$$

where P^{λ} is obtained by Sinkhorn iteration defined in Sinkhorn relaxation for optimal transport.

Proposition 1. Loss function defined in (3) satisfies $0 \le \ell(h(\mathbf{x}),\mathbf{y}) \le \|M\|_{\infty}$, where $\|M\|_{\infty} = \max_{ij} M_{ij}$.

Proof. The loss is non-negative due to the non-negativity of coupling matrix and ground metric. Moreover,

$$
\ell(h(\mathbf{x}),\mathbf{y}) \le \langle P, M\rangle \le \|M\|_{\infty}\sum_{i=1}^{L}\sum_{j=1}^{L} P_{ij} = \|M\|_{\infty}.
$$

□

Based on Sinkhorn loss defined in(3), we introduce notations of corresponding risk and empirical risk, respectively.

$$
R(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(h(\mathbf{x}_i),\mathbf{y}_i), \tag{4}
$$

$$
\hat{R}(h) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}[\ell(h(\mathbf{x}),\mathbf{y})]. \tag{5}
$$

In the following, we will utilize the notion of Rademacher complexity (Bartlett and Mendelson 2002) to measure the hypothesis complexity and use it to bound the risk bounds.

Definition 3. *(Rademacher Complexity (Bartlett and Mendelson 2002)) Let $\mathcal{G}$ be a family of functions and a fixed sample of size m as $S = (\mathbf{z}_1,\cdots,\mathbf{z}_m)$. Then, the empirical Rademacher complexity of $\mathcal{G}$ with respect to the sample S is defined as:*

$$
\hat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(\mathbf{z}_i)\right]
$$

where $\boldsymbol{\sigma} = (\sigma_1,\cdots,\sigma_m)$, with σ_is independent uniform random variables taking values in $\{1,+1\}$. The random variables σ_is are called Rademacher variables.

Besides, the Rademacher complexity of $\mathcal{G}$ is the expectation of the empirical Rademacher complexity over all samples of size m drawn according to $\mathcal{D}$:

$$
\mathfrak{R}_m(\mathcal{G}) = \mathbb{E}_{S\sim\mathcal{D}^m}[\hat{\mathfrak{R}}_S(\mathcal{G})] \tag{6}
$$

Then, we are able to establish a generalization bound based on Rademacher complexity.

Theorem 3. *(Mohri, Rostamizadeh, and Talwalkar 2012) Let $\mathcal{L}$ be the family of loss function associated to $\mathcal{H}$, i.e., $\mathcal{L} = \{\ell(h(\mathbf{x},\mathbf{y}), h\in\mathcal{H}\}$. Then, for any $\delta > 0$, with probability at least $1-\delta$, each of the following holds for all $h\in\mathcal{H}$:*

$$
R(h) \le \hat{R}(h) + 2\mathfrak{R}_m(\mathcal{L}) + \|M\|_{\infty}\sqrt{\frac{\log(1/\delta)}{2m}}, \tag{7}
$$

where $\mathfrak{R}_m(\mathcal{L})$ is Rademacher complexity of loss function class $\mathcal{L}$ associated to $\mathcal{H}$.

Proof of Theorem 3

The proof of Theorem 3 is standard, and we provide in here to make the supplementary self-contained. To prove this theorem, we need following concentration inequality.

Support-Theorem 1. *(McDiarmids inequality)*
Let $X_1, X_2,\cdots,X_m \in \mathcal{X}^m$ be a set of $m\ge 1$ independent random variables and assume that there exist $c_1,\cdots,c_m > 0$ such that $f:\mathcal{X}^m\to\mathbb{R}$ satisfies the following conditions:

$$
|f(x_1,\cdots,x_i,\cdots,x_m) - f(x_1,\cdots,x_i',\cdots,x_m)| \le c_i,
$$

for all $i\in[1,m]$ and any point $x_1,\cdots,x_i,\cdots,x_m,x_i'\in\mathcal{X}$. Let $f(S)$ denote $f(X_1,X_2,\cdots,X_m)$, then, for all $\epsilon>0$, the following inequalities hold:

$$
\begin{aligned}
\Pr[f(S)-\mathbb{E}[f(S)]\ge\epsilon] &\le \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m}c_i^2}\right) \\
\Pr[f(S)-\mathbb{E}[f(S)]\le-\epsilon] &\le \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m}c_i^2}\right)
\end{aligned} \tag{8}
$$

Based on Proposition 1 and Support-Theorem 1, we provide the detailed proof of Theorem 3 as follows,

Proof. The proof is similar to the proof of Theorem 3.1 in (Mohri, Rostamizadeh, and Talwalkar 2012). For any sample $S = (\mathbf{z}_1,\cdots,\mathbf{z}_m), \mathbf{z}_i = (\mathbf{x}_i,\mathbf{y}_i)$ and any $\ell\in\mathcal{L}$, we denote $\hat{\mathbb{E}}_S[\ell]$ the empirical average of ℓ over S : $\hat{\mathbb{E}}_S[\ell] = \frac{1}{m}\sum_{i=1}^{m}\ell(z_i)$. Now we define the function Φ as follows,

$$
\Phi(S) = \sup_{\ell\in\mathcal{L}}\mathbb{E}[\ell] - \hat{\mathbb{E}}_S[\ell].
$$

Let S and S' be two samples differing by exactly one point, say $\mathbf{z}_m$ in S and $\mathbf{z}_m'$ in S'. Then, since the difference of suprema does not exceed the supremum of the difference, we have

$$
\begin{aligned}
\Phi(S') - \Phi(S) &\le \sup_{\ell\in\mathcal{L}}\hat{\mathbb{E}}_S'[\ell] - \hat{\mathbb{E}}_S[\ell] \\
&= \sup_{\ell\in\mathcal{L}}\frac{\ell(\mathbf{z}_m)-\ell(\mathbf{z}_m')}{m} \\
&\le \|M\|_{\infty}/m
\end{aligned}
$$

Similarly, we can obtain $\Phi(S)-\Phi(S') \le \|M\|_{\infty}/m$, thus $|\Phi(S)-\Phi(S') \le \|M\|_{\infty}/m$. Then, by McDiarmids inequality, for any $\delta > 0$, with probability at least $1 - \delta/2$, the following holds:

$$\Phi(S) \le \mathbb{E}_S[\Phi(S)] + \|M\|_\infty \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

Now we will proceed to bound $\mathbb{E}_S[\Phi(S)]$ as follows,

$$\begin{aligned}
&\mathbb{E}_S[\Phi(S)] \\
&= \mathbb{E}_S[\sup_{\ell \in \mathcal{L}} \mathbb{E}[\ell] - \hat{\mathbb{E}}_S[\ell]] \\
&= \mathbb{E}_S[\sup_{\ell \in \mathcal{L}} \mathbb{E}_{S'}[\hat{\mathbb{E}}_{S'}[\ell] - \hat{\mathbb{E}}_S[\ell]]] \\
&\le \mathbb{E}_{S,S'}[\sup_{\ell \in \mathcal{L}} \hat{\mathbb{E}}_{S'}[\ell] - \hat{\mathbb{E}}_S[\ell]] \\
&= \mathbb{E}_{S,S'}\left[\sup_{\ell \in \mathcal{L}} \frac{1}{m}\sum_{i=1}^m (\ell(\mathbf{z}_i') - \ell(\mathbf{z}_i))\right] \\
&= \mathbb{E}_{\sigma,S'}\left[\sup_{\ell \in \mathcal{L}} \frac{1}{m}\sum_{i=1}^m \sigma_i \ell(\mathbf{z}_i')\right] + \mathbb{E}_{\sigma,S}\left[\sup_{\ell \in \mathcal{L}} \frac{1}{m}\sum_{i=1}^m -\sigma_i \ell(\mathbf{z}_i)\right] \\
&= 2\mathbb{E}_{\sigma,S}\left[\sup_{\ell \in \mathcal{L}} \frac{1}{m}\sum_{i=1}^m \sigma_i \ell(\mathbf{z}_i)\right] = 2\mathfrak{R}_m(\mathcal{L})
\end{aligned} \tag{9}$$

Thus, we have

$$R(h) \le \hat{R}(h) + 2\mathfrak{R}_m(\mathcal{L}) + \|M\|_\infty \sqrt{\frac{\log(1/\delta)}{2m}}.$$

□

Proof of Useful Properties of Sinkhorn Distance

Lemma 1. *For any double stochastic matrix $S \in \mathbb{R}_+^{d\times d}$, its entropy $H(S)$ satisfies $H(S) \le 2\log d$.*

Proof. Since the entropy function is concave,

$$\begin{aligned}
H(S) &= -\sum_{i=1}^d \sum_{j=1}^d S_{ij} \log S_{ij} \\
&\le -d^2 \cdot \frac{\mathtt{S}}{d^2} \log \frac{\mathtt{S}}{d^2} = 2\log d
\end{aligned} \tag{10}$$

where the last equation holds due to S is a double stochastic matrix such that $\mathtt{S} = \sum_{i=1}^d \sum_{j=1}^d S_{ij} = 1$. □

Lemma 2. *For two probability distributions $r, c \in \Sigma_d$, Sinkhorn distance $d_M^\lambda(r,c)$ and optimal transport distance $d_M(r,c)$ satisfy the following relationship,*

$$d_M(r,c) \le d_M^\lambda(r,c) \le d_M(r,c) + \frac{2}{\lambda}\log d \tag{11}$$

Proof. Let P^* and P^λ be corresponding coupling matrix of $d_M(r,c)$ and $d_M^\lambda(r,c)$, i.e.,

$$\begin{aligned}
P^* &= \underset{P \in U(r,c)}{\arg\min} \langle P, M \rangle, \\
P^\lambda &= \underset{P \in U(r,c)}{\arg\min} \langle P, M \rangle - \frac{1}{\lambda} H(P).
\end{aligned} \tag{12}$$

Then the left inequality is obvious since P^* is the optimal solution of optimal transport distance. Moreover, due to the optimality of P^λ for Sinkhorn distance, we have

$$\langle P^\lambda, M \rangle - \frac{1}{\lambda} H(P^\lambda) \le \langle P^*, M \rangle - \frac{1}{\lambda} H(P^*)$$

Therefore, we have

$$\begin{aligned}
d_M^\lambda(r,c) &\le d_M(r,c) + \frac{1}{\lambda}[H(P^\lambda) - H(P^*)] \\
&\le d_M(r,c) + \frac{2}{\lambda}\log d
\end{aligned} \tag{13}$$

The last inequality holds due to Lemma 1 and the non-negativity. □

In order to establish the relationship between Rademacher complexity of Sinkhorn distance loss and function space, we need introduce another loss definition based on original optimal transport distance as

$$\ell_{OT}(h(\mathbf{x}), \mathbf{y}) := d_M(h(\mathbf{x}), \mathbf{y}) = \langle P, M \rangle \tag{14}$$

Then, based on Lemma 2, we know that

$$\ell_{OT}(h(\mathbf{x}), \mathbf{y}) \le \ell(h(\mathbf{x}), \mathbf{y}) \le \ell_{OT}(h(\mathbf{x}), \mathbf{y}) + \frac{2\log L}{\lambda}$$

holds for any instance $(\mathbf{x}, \mathbf{y})$. Now we can relate the Rademacher complexity associated with these two losses as stated in Theorem 4.

Theorem 4. *Let $\mathcal{L}$ and $\mathcal{L}_{OT}$ correspond the family of loss functions ℓ and ℓ_{OT} associated to function space $\mathcal{H}$. Then the empirical Rademacher complexities of $\mathcal{L}$ and $\mathcal{L}_{OT}$ satisfy,*

$$\mathfrak{R}_m(\mathcal{L}) \le \mathfrak{R}_m(\mathcal{L}_{OT}) + \frac{\log L}{\lambda}. \tag{15}$$

Proof. Let the Rademacher random variables sequence be $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_m)^{\mathrm{T}}$, with σ_is independent uniform random variables taking values in $\{1, +1\}$, then we have

$$\sigma_i \ell(h(\mathbf{x}_i), \mathbf{y}_i) \le \sigma_i \ell_{OT}(h(\mathbf{x}_i), \mathbf{y}_i) + \mathbf{1}_{\sigma_i = 1} \frac{2\log L}{\lambda}$$

Thus, with index summing to m, we have

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{L}) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i \ell(h(\mathbf{x}_i), \mathbf{y}_i)\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sum_{i=1}^m \sigma_i \ell(h^*(\mathbf{x}_i), \mathbf{y}_i)\right] \\
&\le \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sum_{i=1}^m \sigma_i \ell_{OT}(h^*(\mathbf{x}_i), \mathbf{y}_i)\right] \\
&\quad + \frac{2\log L}{\lambda}\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\sigma_i = 1}\right] \\
&\le \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i \ell_{OT}(h(\mathbf{x}_i), \mathbf{y}_i)\right] + \frac{\log L}{\lambda} \\
&= \hat{\mathfrak{R}}_S(\mathcal{L}_{OT}) + \frac{\log L}{\lambda}
\end{aligned}$$

In the second last step, we use the fact that

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_{\sigma_i=1}\right]=\frac{1}{2^m}\sum_{k=0}^{m}\frac{1}{m}\binom{m}{k}k$$
$$=\frac{1}{2^m}\sum_{k=1}^{m}\binom{m-1}{k-1}=\frac{1}{2}.$$

Taking the expectation w.r.t. sample set S, we can immediately obtain,

$$\mathfrak{R}_m(\mathcal{L})\leq\mathfrak{R}_m(\mathcal{L}_{OT})+\frac{\log L}{\lambda}.$$

□

Now, we can provide the risk bound for ERM based on the Sinkhorn loss.

Theorem 5. *Let $\mathcal{H}$ be the family of hypothesis set, and denote the hypothesis returned by LALOT as $\hat{h}$. Then, for any $\delta>0$, with probability at least $1-\delta$,*

$$R(\hat{h})\leq\inf_{h\in\mathcal{H}}R(h)+\frac{4\log L}{\lambda}+\|M\|_\infty\left(16L\mathfrak{R}_m(\mathcal{H})+\sqrt{\frac{2\log\frac{1}{\delta}}{m}}\right),$$

where $\mathfrak{R}_m(\mathcal{H})$ is Rademacher complexity of hypothesis class $\mathcal{H}$, and $\|M\|_\infty=\max_{ij}M_{ij}$.

Proof of Theorem 5

The proof of Theorem 5 relies on the *Rademacher Vector Contraction Inequality* (Maurer 2016).

Support-Theorem 2 (Rademacher Vector Contraction Inequality (Maurer 2016))**.** *Let $\mathcal{F}$ be a class of real functions, and $\mathcal{H}\subset\mathcal{F}=\mathcal{F}_1\times\cdots\times\mathcal{F}_L$ be a L-valued function class. If $\Phi:\mathbb{R}^L\to\mathbb{R}$ is a G-Lipschitz continuous function and $\Phi(0)=0$, then $\hat{\mathfrak{R}}_S(\Phi\circ\mathcal{H})\leq\sqrt{2}G\sum_{i=1}^{L}\hat{\mathfrak{R}}_S(\mathcal{F}_i)$.*

Remark. The support-Theorem 2 here is tighter than the typically used *Generalized Talagrand's Comparison Inequality* (Ledoux and Talagrand 2013) in this scenario.

Besides, as a well-known conclusion that optimal transport distance is controlled by total variation (Villani 2008).

Support-Theorem 3. *(Theorem 6.15 in (Villani 2008)) Let μ and ν be two probability measures on a Polish space (X,d). Let $p\in[1,\infty)$ and $x_0\in\mathcal{X}$. Then*

$$W_p(\mu,\nu)\leq 2^{\frac{1}{q}}\left(\int d(x_0,x)^p d|\mu-\nu|(x)\right)^{\frac{1}{p}},\quad \frac{1}{p}+\frac{1}{q}=1.$$

Corollary 1. When p = 1, if the diameter of $\mathcal{X}$ is bounded by D, this bound implies $W_1(\mu,\nu)\leq D\|\mu-\nu\|_{TV}$.

Corollary 2. For the optimal transport loss defined on cost matrix M, we have

$$\ell_{OT}(\hat{\mathbf{y}},\mathbf{y})\leq\|M\|_\infty\|\hat{\mathbf{y}}-\mathbf{y}\|_1 \tag{16}$$

However, the Support-Theorem 2 cannot be directly applied on optimal transport distance, because for a probability distribution, $\mathbf{0}$ is not a valid input. Thus, similar to the processing method in (Frogner et al. 2015), we get rid of this by adding a softmax layer before obtaining the final results,

$$\mathcal{H}=\{\mathfrak{s}\circ h^0:h^0\in\mathcal{H}^0\}$$

where $\mathcal{H}^0$ is a function class that maps into $\mathbb{R}^L$, and $\mathfrak{s}$ is the softmax function defined as $\mathfrak{s}(\circ)=(\mathfrak{s}_1(\circ),\cdots,\mathfrak{s}_L(\circ))$, with

$$\mathfrak{s}_k(\circ)=\frac{e^{\circ_k}}{\sum_{j=1}^{L}e^{\circ_j}},\quad k=1,\cdots,L$$

Now, we could provide the Lipschitz condition for optimal transport distance loss.

Support-Theorem 4. *Let the map $\mathfrak{l}:\mathbb{R}^L\times\mathbb{R}^L$ defined by $\mathfrak{l}(\mathbf{y},\mathbf{y}')=\ell_{OT}(\mathfrak{s}(\mathbf{y}),\mathfrak{s}(\mathbf{y}'))$, then we have*

$$|\mathfrak{l}(\mathbf{y},\hat{\mathbf{y}})-\mathfrak{l}(\mathbf{y}',\hat{\mathbf{y}}')|\leq 2\sqrt{2}\|M\|_\infty\|(\mathbf{y},\hat{\mathbf{y}})-(\mathbf{y}',\hat{\mathbf{y}}')\|_2.$$

Besides, $\mathfrak{l}(\mathbf{0},\mathbf{0})=0$.

Proof.

$$\begin{aligned}
&|\mathfrak{l}(\mathbf{y},\hat{\mathbf{y}})-\mathfrak{l}(\mathbf{y}',\hat{\mathbf{y}}')|\\
&=|\mathfrak{l}(\mathbf{y},\hat{\mathbf{y}})-\mathfrak{l}(\mathbf{y}',\hat{\mathbf{y}})+\mathfrak{l}(\mathbf{y}',\hat{\mathbf{y}})-\mathfrak{l}(\mathbf{y}',\hat{\mathbf{y}}')|\\
&\leq|\mathfrak{l}(\mathbf{y},\hat{\mathbf{y}})-\mathfrak{l}(\mathbf{y}',\hat{\mathbf{y}})|+|\mathfrak{l}(\mathbf{y}',\hat{\mathbf{y}})-\mathfrak{l}(\mathbf{y}',\hat{\mathbf{y}}')|\\
&\leq\mathfrak{l}(\mathbf{y},\mathbf{y}')+\mathfrak{l}(\hat{\mathbf{y}},\hat{\mathbf{y}}') && \text{(17a)}\\
&\leq\|M\|_\infty(\|\mathfrak{s}(\mathbf{y})-\mathfrak{s}(\mathbf{y}')\|_1+\|\mathfrak{s}(\hat{\mathbf{y}})-\mathfrak{s}(\hat{\mathbf{y}}')\|_1) && \text{(17b)}\\
&\leq 2\|M\|_\infty(\|\mathbf{y}-\mathbf{y}'\|_2+\|\hat{\mathbf{y}}-\hat{\mathbf{y}}'\|_2) && \text{(17c)}\\
&\leq 2\sqrt{2}\|M\|_\infty\left(\|\mathbf{y}-\mathbf{y}'\|_2^2+\|\hat{\mathbf{y}}-\hat{\mathbf{y}}'\|_2^2\right)^{1/2} && \text{(17d)}\\
&=2\sqrt{2}\|M\|_\infty\|(\mathbf{y},\hat{\mathbf{y}})-(\mathbf{y}',\hat{\mathbf{y}}')\|_2
\end{aligned}$$

here, (17a) holds due to the sub-additivity of optimal transport loss, and (17b) can be directly obtained by Corollary 2. (17b) can be proved by mean value theorem, and a detailed proof can be found in (Frogner et al. 2015). (17d) immediately follows by Cauchy-Schwarz inequality. □

Based on above Support-Theorem 2, 3 and 4, we will proceed the proof of Theorem 5.

Proof. From Support-Theorem 4, we know that $\mathfrak{l}$ defined there is a $2\sqrt{2}\|M\|_\infty$-Lipschitz function. Thus, we could apply Support-Theorem 2. It holds

$$\hat{\mathfrak{R}}_S(\mathcal{L}_{OT})\leq\sqrt{2}\cdot 2\sqrt{2}\|M\|_\infty L\mathfrak{R}_S(\mathcal{H}) \tag{18}$$

Thus,

$$\begin{aligned}\mathfrak{R}_m(\mathcal{L})&\leq\mathfrak{R}_m(\mathcal{L}_{OT})+\frac{\log L}{\lambda}\\&\leq 4L\|M\|_\infty\mathfrak{R}_S(\mathcal{H})+\frac{\log L}{\lambda}\end{aligned} \tag{19}$$

The conclusion in Theorem 5 follows immediately by plugging (19) back to Theorem 3.

By plugging (19) back to Theorem 3, a generalization bound is immediately obtained so far, namely, for any hypothesis h in $\mathcal{H}$,

$$R(h)-\hat{R}(h)\leq\frac{2\log L}{\lambda}+\|M\|_\infty\left(8L\mathfrak{R}_m(\mathcal{H})+\sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right) \tag{20}$$

To make the risk bound succeed, we only need to bound it by uniform generalization error as follows,

$$
\begin{aligned}
& R(\hat{h}) - \inf_{h\in\mathcal{H}} R(h) \\
= & R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(\hat{h}) - \inf_{h\in\mathcal{H}} R(h) \\
\leq & R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(h^\star) - R(h^\star) \\
\leq & 2 \sup_{h\in\mathcal{H}} |R(h) - \hat{R}(h)|
\end{aligned}
\tag{21}
$$

where $h^\star = \arg\inf_{h\in\mathcal{H}} R(h)$, and the first inequality holds due to the fact that $\hat{h}$ is the minimizer of empirical risk. Combine (20) and (21), the proposition in Theorem 5 can be immediately obtained. □

References

Bartlett, P. L., and Mendelson, S. 2002. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* 3:463–482.

Cuturi, M., and Avis, D. 2014. Ground metric learning. *Journal of Machine Learning Research* 15(1):533–564.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, 2292–2300.

Frogner, C.; Zhang, C.; Mobahi, H.; Araya-Polo, M.; and Poggio, T. A. 2015. Learning with a Wasserstein loss. In *NIPS*, 2053–2061.

Ledoux, M., and Talagrand, M. 2013. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Maurer, A. 2016. A vector-contraction inequality for rademacher complexities. In *Proceedings of International Conference on Algorithmic Learning Theory (ALT)*, 3–17.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2012. *Foundations of machine learning*. MIT press.

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

Zhao, P., and Zhou, Z.-H. 2018. Label distribution learning by optimal transport. In *AAAI*, 4506–4513.