Representation Subspace Distance for Domain Adaptation Regression

Xinyang Chen [1] **Sinan Wang** [1] **Jianmin Wang** [1] **Mingsheng Long** [1]

A. Appendix

A.1. Proof of Theorem 1

Proof. We prove the three properties of general metrics in order.

(1) $\mathsf{dis}_{\mathrm{RSD}}^{\mathcal{S}\leftrightarrow\mathcal{T}} \geq 0$, and $\mathsf{dis}_{\mathrm{RSD}}^{\mathcal{S}\leftrightarrow\mathcal{T}} = 0$ if and only if $\mathcal{S} = \mathcal{T}$.

It is shown in (Golub & Van Loan, 1996): $\theta_i \in (0, \pi/2), \sin\theta_i \in [0,1)$ Hence $\mathsf{dis}_{\mathrm{RSD}}^{\mathcal{S}\leftrightarrow\mathcal{T}} \geq 0$ is proved. It is also shown in (Golub & Van Loan, 1996): $\forall \theta_i$, $\mathcal{S} = \mathcal{T}$ if and only if $\theta_i = 0$. Thus $\mathsf{dis}_{\mathcal{S}\leftrightarrow\mathcal{T}} = 0$ is satisfied.

(2) $\mathsf{dis}_{\mathrm{RSD}}^{\mathcal{S}\leftrightarrow\mathcal{T}} = \mathsf{dis}_{\mathrm{RSD}}^{\mathcal{T}\leftrightarrow\mathcal{S}}$. (symmetric)

We can easily prove it in the definition:

$$
\begin{aligned}
\theta_1^{\mathcal{S}\leftrightarrow\mathcal{T}} &= \min_{\mathbf{u}_1^s \in \mathcal{S}, \mathbf{u}_1^t \in \mathcal{T}} \arccos\left(\frac{(\mathbf{u}_1^s)^\mathsf{T}\mathbf{u}_1^t}{\|\mathbf{u}_1^s\|\,\|\mathbf{u}_1^t\|}\right), \\
\theta_2^{\mathcal{S}\leftrightarrow\mathcal{T}} &= \min_{\substack{\mathbf{u}_2^s \in \mathcal{S}, \mathbf{u}_2^t \in \mathcal{T} \\ \mathbf{u}_2^s \perp \mathbf{u}_1^s, \mathbf{u}_2^t \perp \mathbf{u}_1^t}} \arccos\left(\frac{(\mathbf{u}_2^s)^\mathsf{T}\mathbf{u}_2^t}{\|\mathbf{u}_2^s\|\,\|\mathbf{u}_2^t\|}\right), \\
&\;\;\vdots \\
\theta_b^{\mathcal{S}\leftrightarrow\mathcal{T}} &= \min_{\substack{\mathbf{u}_b^s \in \mathcal{S}, \mathbf{u}_b^t \in \mathcal{T} \\ \mathbf{u}_b^s \perp \mathbf{u}_1^s, \ldots, \mathbf{u}_{b-1}^s \\ \mathbf{u}_b^t \perp \mathbf{u}_1^t, \ldots, \mathbf{u}_{b-1}^t}} \arccos\left(\frac{(\mathbf{u}_b^s)^\mathsf{T}\mathbf{u}_b^t}{\|\mathbf{u}_b^s\|\,\|\mathbf{u}_b^t\|}\right),
\end{aligned}
\tag{1}
$$

cos function is symmetric, thus $\theta_i^{\mathcal{S}\leftrightarrow\mathcal{T}} = \theta_i^{\mathcal{T}\leftrightarrow\mathcal{S}}$, for $i = 1, ..., b$. It is observed that $\mathsf{dis}_{\mathrm{RSD}}^{\mathcal{S}\leftrightarrow\mathcal{T}}$ is symmetric.

(3) $\mathsf{dis}_{\mathrm{RSD}}^{\mathcal{S}\leftrightarrow\mathcal{T}} \leq \mathsf{dis}_{\mathrm{RSD}}^{\mathcal{S}\leftrightarrow\mathcal{A}} + \mathsf{dis}_{\mathrm{RSD}}^{\mathcal{T}\leftrightarrow\mathcal{A}}$. (triangle inequality)

We first introduce the concept of weak majorization: Majorization is a preorder on vectors of real numbers. For a vector $\mathbf{a} \in \mathbb{R}^d$, we denote by $\mathbf{a}^\downarrow \in \mathbb{R}^d$ the vector with the same components, but sorted in descending order. Given $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, we say that $\mathbf{a}$ weakly majorizes (or dominates) $\mathbf{b}$ from below written as $\mathbf{a} \succ_w \mathbf{b}$ if and only if $\sum_{i=1}^{k} a_i^\downarrow \geq \sum_{i=1}^{k} b_i^\downarrow$ for $k = 1, \ldots, d$. Thus if we want to prove $\mathsf{dis}_{\mathrm{RSD}}^{\mathcal{S}\leftrightarrow\mathcal{T}} \leq \mathsf{dis}_{\mathrm{RSD}}^{\mathcal{T}\leftrightarrow\mathcal{A}} + \mathsf{dis}_{\mathrm{RSD}}^{\mathcal{S}\leftrightarrow\mathcal{A}}$, we can prove $|\sin\mathbf{\Theta}^{\mathcal{S}\leftrightarrow\mathcal{A}} - \sin\mathbf{\Theta}^{\mathcal{T}\leftrightarrow\mathcal{A}}| \prec_w \sin\mathbf{\Theta}^{\mathcal{S}\leftrightarrow\mathcal{T}}$ instead.

We first use several weak majorization results. Marshall et al. (1979) give a starting point, for Hermitian $\mathbf{A}$ and $\mathbf{B}$, we have

$$\Lambda(\mathbf{A}+\mathbf{B}) \prec_w \Lambda(\mathbf{A}) + \Lambda(\mathbf{B}), \tag{2}$$

where $\Lambda(\mathbf{A})$ denote the vector of all eigenvalues of $\mathbf{A}$ in nonincreasing order. Inspired by that, Horn & Johnson (2012) give anthor two results:

$$\Lambda(\mathbf{A}) - \Lambda(\mathbf{B}) \prec_w \Lambda(\mathbf{A}-\mathbf{B}), \tag{3}$$

$$\Sigma(\mathbf{A}\pm\mathbf{B}) \prec_w \Sigma(\mathbf{A}) + \Sigma(\mathbf{B}), \tag{4}$$

$$\Sigma(\mathbf{A}-\mathbf{B}) \prec_w \Sigma(\mathbf{A}) - \Sigma(\mathbf{B}), \tag{5}$$

where $\Sigma(\mathbf{A})$ denote the singular values of all eigenvalues of $\mathbf{A}$ in nonincreasing order.

Based on these results, Knyazev & Argentati (2007) prove $|\sin\mathbf{\Theta}^{\mathcal{S}\leftrightarrow\mathcal{A}} - \sin\mathbf{\Theta}^{\mathcal{T}\leftrightarrow\mathcal{A}}| \prec_w \sin\mathbf{\Theta}^{\mathcal{S}\leftrightarrow\mathcal{T}}$ as follows:

Denote by $P_{\mathcal{S}}$, $P_{\mathcal{T}}$ and $P_{\mathcal{A}}$ the corresponding orthogonal projectors onto the subspaces $\mathcal{S}$, $\mathcal{T}$ and $\mathcal{A}$, Knyazev & Argentati (2007) give anthor result:

$$
\begin{aligned}
&[\Sigma(P_{\mathcal{S}} - P_{\mathcal{T}}), 0, \ldots, 0] = \\
&[1, \ldots, 1, (\sin\Theta^{\mathcal{S}\leftrightarrow\mathcal{T}}, \sin\Theta^{\mathcal{S}\leftrightarrow\mathcal{T}})^\downarrow, 0, \ldots, 0],
\end{aligned}
\tag{6}
$$

where there are $|\dim(\mathcal{S}) - \dim(\mathcal{T})|$ extra 1s upfront. The set $\sin\Theta^{\mathcal{S}\leftrightarrow\mathcal{T}}$ is repeated twice and ordered, and extra 0s at the end may need to be added on either side to match the sizes.

In our definiation of RSD, all representation subspaces are the same dimension. Thus this result in our case is:

$$
\begin{aligned}
&[\Sigma(P_{\mathcal{S}} - P_{\mathcal{T}}), 0, \ldots, 0] = \\
&[(\sin\Theta^{\mathcal{S}\leftrightarrow\mathcal{T}}, \sin\Theta^{\mathcal{S}\leftrightarrow\mathcal{T}})^\downarrow, 0, \ldots, 0],
\end{aligned}
\tag{7}
$$

Based on that, Knyazev & Argentati (2007) give a new result: $(P_{\mathcal{S}} - P_{\mathcal{A}}) - (P_{\mathcal{T}} - P_{\mathcal{A}}) = (P_{\mathcal{S}} - P_{\mathcal{T}})$. Using (5), we have $\Sigma(P_{\mathcal{S}} - P_{\mathcal{A}}) - \Sigma(P_{\mathcal{T}} - P_{\mathcal{A}}) = \Sigma(P_{\mathcal{S}} - P_{\mathcal{T}})$. Then we use (7), the set of nonzero entries of $|\Sigma(P_{\mathcal{S}} - P_{\mathcal{A}}) -$

[1]School of Software, BNRist, Tsinghua University.
Xinyang Chen <chenxinyang95@gmail.com>. Correspondence to: Mingsheng Long <mingsheng@tsinghua.edu.cn>.

Proceedings of the 38^{th} *International Conference on Machine Learning*, PMLR 139, 2021.

Table 1. Sum of MAE across three regression tasks on dSprites: unsupervised domain adaptation (ResNet-18).

Method	C → N	C → S	N → C	N → S	S → C	S → N	Avg
ResNet-18 (He et al., 2016)	0.94 ± 0.06	0.90 ± 0.08	0.16 ± 0.02	0.65 ± 0.02	0.08 ± 0.01	0.26 ± 0.03	0.498
ResNet-18 w/ BN	0.97 ± 0.03	0.92 ± 0.06	0.18 ± 0.02	0.67 ± 0.02	0.08 ± 0.00	0.30 ± 0.02	0.520
DANN (Ganin et al., 2016)	0.47 ± 0.07	0.46 ± 0.07	0.16 ± 0.02	0.65 ± 0.05	$\mathbf{0.05} \pm 0.00$	0.10 ± 0.01	0.315
DANN w/ BN	0.68 ± 0.05	0.71 ± 0.04	0.18 ± 0.02	0.63 ± 0.05	0.06 ± 0.01	0.11 ± 0.01	0.393

Table 2. Sum of MAE across two regression tasks and cosine values of principal angles on three transfer tasks on MPI3D.

	RL → T			**T → RL**			**T → RC**		
Method	MAE	$\cos\theta_1^{\mathcal{S}\leftrightarrow\mathcal{T}}$	$\cos\theta_{36}^{\mathcal{S}\leftrightarrow\mathcal{T}}$	MAE	$\cos\theta_1^{\mathcal{S}\leftrightarrow\mathcal{T}}$	$\cos\theta_{36}^{\mathcal{S}\leftrightarrow\mathcal{T}}$	MAE	$\cos\theta_1^{\mathcal{S}\leftrightarrow\mathcal{T}}$	$\cos\theta_{36}^{\mathcal{S}\leftrightarrow\mathcal{T}}$
ResNet-18 (He et al., 2016)	0.44	0.951	0.029	0.51	0.938	0.017	0.50	0.935	0.016
Geodesic Distance	0.32	0.973	0.097	0.52	0.962	0.083	0.46	0.966	0.076
RSD (ours)	0.23	0.991	0.049	0.41	0.987	0.032	0.42	0.985	0.035

$\Sigma(P_\mathcal{T} - P_\mathcal{A})|$ consists of nonzeros entries of $|\sin\Theta^{\mathcal{S}\leftrightarrow\mathcal{A}} - \sin\Theta^{\mathcal{T}\leftrightarrow\mathcal{A}}|$ repeated twice. And$\Sigma(P_\mathcal{S} - P_\mathcal{T})$ consists of nonzeros entries of $\sin\Theta^{\mathcal{S}\leftrightarrow\mathcal{T}}$ repeated twice. Thus we can get $|\sin\mathbf{\Theta}^{\mathcal{S}\leftrightarrow\mathcal{A}} - \sin\mathbf{\Theta}^{\mathcal{T}\leftrightarrow\mathcal{A}}| \prec_w \sin\mathbf{\Theta}^{\mathcal{S}\leftrightarrow\mathcal{T}}$.

□

A.2. More Experimental Details

For Figure 1(a)(b) in the paper, we use L2 regularization to change the Frobenius norm of feature matrix:

$$\mathsf{reg}_{\mathrm{norm}} = (\||\mathbf{F}^s\|_F^2 - R)^2, \tag{8}$$

where R is the expected Frobenius norm of source feature matrix. We give 0.05 as a trade-off hyper-parameter for $\mathsf{reg}_{\mathrm{norm}}$ and we run 10000 iterations.

For Figure 1(c) in the paper, we tune the best hyperparamters in DANN and AFN. Average of the Frobenius norm of source feature matrix is reported. The reason why we only report results in source domain is that when distribution discrepancy is minimized, the Frobenius norm of target feature matrix is the same as the Frobenius norm of source feature matrix.

A.3. Normalization

It is well known that normalization techniques are useful in deep learning, and they have the potential to solve the problem of feature scaling. Ioffe & Szegedy (2015), Ba et al. (2016) and Wu & He (2018) are the most widely used techniques. We choose batch normalization (BN) to solve the problem of feature scaling (using BN before regressor). Results are shown in Table 1. We observe that add a BN layer before regressor is harmful to domain adaptation performances. This is consistent with the observation of (Lathuilière et al., 2019): add a BN layer before the regressor in deep regression may have negative effects on ResNet.

A.4. Geodesic Distance

Geodesic Distance (GD) is a commonly used geometrical distance to measure the similarity of subspaces: $\mathsf{dis}_{\mathrm{GD}}^{\mathcal{S}\leftrightarrow\mathcal{T}}(\mathbf{U}^s, \mathbf{U}^t) = \left\|\mathbf{\Theta}^{\mathcal{S}\leftrightarrow\mathcal{T}}\right\|_2$. And it can be used as a regularizer in the representation learning process. The main difference is that RSD gives larger gradients to the minimum angle (due to sin and L1-norm), while GD gives larger gradients to the maximum angle (the L2-norm amplifies it). Since a large angle means low similarity, it is unreasonable to match two dissimilar bases with GD minimization. We conduct experiments using GD regularization in Table 2. It is observed that using GD will make small principal angles smaller, while RSD prefers to make the large principal angles smaller.

References

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research (JMLR)*, 17(1): 2096–2030, 2016.

Golub, G. H. and Van Loan, C. F. *Matrix computations (3rd ed.)*. DBLP, 1996.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE conference on* *computer vision and pattern recognition (CVPR)*, pp. 770–778, 2016.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge university press, 2012.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning (ICML)*, pp. 448–456. PMLR, 2015.

Knyazev, A. V. and Argentati, M. E. Majorization for changes in angles between subspaces, ritz values, and graph laplacian spectra. *SIAM journal on matrix analysis and applications*, 29(1):15–32, 2007.

Lathuilière, S., Mesejo, P., Alameda-Pineda, X., and Horaud, R. A comprehensive analysis of deep regression. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2019.

Marshall, A. W., Olkin, I., and Arnold, B. C. *Inequalities: theory of majorization and its applications*, volume 143. Springer, 1979.

Wu, Y. and He, K. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.