The Role of Embedding-complexity in Domain-invariant Representations

Anonymous Authors[1]

Abstract

Unsupervised domain adaptation aims to generalize the hypothesis trained in a source domain to an unlabeled target domain. One popular approach to this problem is to learn a domain-invariant representation for both domains. In this work, we study, theoretically and empirically, the explicit effect of the embedding on generalization to the target domain. In particular, the complexity of the class of embeddings affects an upper bound on the target domain's risk. This is reflected in our experiments, too.

1. Introduction

Domain adaptation is critical in many applications where collecting large-scale supervised data is prohibitively expensive or intractable, or conditions at prediction time can change. For instance, self-driving cars must be robust to various conditions such as different weather, change of landscape and traffic. In such cases, the model learned from limited source data should ideally generalize to different target domains. Specifically, unsupervised domain adaptation aims to transfer knowledge learned from a labeled source domain to similar but completely unlabeled target domains.

One popular approach to unsupervised domain adaptation is to learn domain-invariant representations (Ben-David et al., 2007; Long et al., 2015; Ganin et al., 2016), by minimizing a divergence between the representations of source and target domains. The prediction function is learned on the latent space, with the aim of making it domain-independent. A series of theoretical works justifies this idea (Ben-David et al., 2007; Mansour et al., 2009; Ben-David et al., 2010; Cortes & Mohri, 2011).

Despite the empirical success of domain-invariant representations, exactly matching the representations of source and target distribution can sometimes fail to achieve domain adaptation. For example, Wu et al. (2019) show that exact matching may increase target error if label distributions are different between source and target domain, and propose a new divergence metric to overcome this limitation. Zhao et al. (2019) establish lower and upper bounds on the risk when label distributions between source and target domains differ. Johansson et al. (2019) point out the information lost in non-invertible embeddings, and propose different generalization bounds based on the overlap of the supports of source and target distribution.

In contrast to previous analyses that focus on changes in the label distributions or on joint support, we here study the effect of the complexity of the joint representation. In particular, we show a general bound on the target risk that reflects a tradeoff between the embedding complexity and the divergence of source and target in the latent representation space. In particular, a too powerful class of embedding functions can result in overfitting the source data and the distribution matching, leading to arbitrarily high target risk. Hence, a restriction (taking into account assumptions about correspondences and invariances) is needed. Our experiments reflect these trends empirically, too.

2. Unsupervised Domain Adaptation

For simplicity, we consider binary classification tasks with input space $\mathcal{X} \subseteq \mathbb{R}^n$ and output space $\mathcal{Y} = \{0, 1\}$. Define $\mathcal{H}$ to be the hypothesis class from $\mathcal{X}$ to $\mathcal{Y}$. The learning algorithm obtains two datasets: labeled source data $\mathcal{X}_S$ with distribution $p_S(X)$), and unlabeled target data $\mathcal{X}_T$ with distribution $p_T(X)$. Unsupervised domain adaptation seeks a hypothesis $h \in \mathcal{H}$ that minimizes the risk in the target domain measured by a loss function ℓ (here, the zero-one loss):

$$R_T(h) = \mathbb{E}_{x,y\sim p_T}[\ell(h(x), y)]. \tag{1}$$

We will not assume common support between source and target domain, in line with standard benchmarks for domain adaptation such as adapting from MNIST to M-MNIST.

2.1. Domain-invariant Representations

A common approach to domain adaptation is to learn a joint embedding of source and target data (Ganin et al., 2016; Tzeng et al., 2017). The idea is that aligning source

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

and target distributions in this latent space $\mathcal{Z}$ results in a domain-invariant representation, and hence a subsequent classifier f from the embedding to $\mathcal{Y}$ will generalize from source to target. Formally, this results in the following objective function, where $\mathcal{G}$ is the class of embedding functions and we minimize a divergence d between the distributions $p_S(Z_g), p_T(Z_g)$ of source and target:

$$\min_{f\in\mathcal{F},g\in\mathcal{G}} R_S(f\circ g) + \alpha d(p_S(Z_g), p_T(Z_g)). \tag{2}$$

The divergence d could be the Jensen-Shannon (Ganin et al., 2016) or Wasserstein distance (Shen et al., 2017).

2.2. Upper bounds on the target risk

Ben-David et al. (2007) introduced the $\mathcal{H}\Delta\mathcal{H}$-divergence to bound the worst-case loss from extrapolating between domains. Let $R_D(h,h') = \mathbb{E}_{x\sim D}[\ell(h(x),h'(x))]$ be the expected disagreement between two hypotheses, then the $\mathcal{H}\Delta\mathcal{H}$-divergence is defined as follows.

Definition 1. ($\mathcal{H}\Delta\mathcal{H}$-divergence) *Given two domain distributions p_S and p_T over $\mathcal{X}$, and a hypothesis class $\mathcal{H}$, the $\mathcal{H}\Delta\mathcal{H}$-divergence between p_S and p_T is*

$$d_{\mathcal{H}\Delta\mathcal{H}}(p_S,p_T) = \sup_{h,h'\in\mathcal{H}} |R_S(h,h') - R_T(h,h')|.$$

This divergence allows to bound the risk on the target domain:

Theorem 1. (Ben-David et al., 2010) *For all hypotheses $h\in\mathcal{H}$, the target risk is bounded as*

$$R_T(h) \le R_S(h) + d_{\mathcal{H}\Delta\mathcal{H}}(p_S,p_T) + \lambda_{\mathcal{H}},$$

where $\lambda_{\mathcal{H}}$ is the best joint in-class joint risk

$$\lambda_{\mathcal{H}} := \inf_{h\in\mathcal{H}} [R_S(h) + R_T(h)]$$

Similar results have been obtained for continuous labels (Cortes & Mohri, 2011; Mansour et al., 2009).

Theorem 1 is an influential theoretical result in unsupervised domain adaptation, and motivated work on domain invariant representations. For example, recent work (Ganin et al. (2016); Johansson et al. (2019)) applied Theorem 1 to the hypothesis space $\mathcal{F}$ that maps the representation space induced by an encoder g to the output space:

$$R_T(f\circ g) \le R_S(f\circ g) + d_{\mathcal{F}\Delta\mathcal{F}}(p_S(Z_g),p_T(Z_g)) + \lambda_{\mathcal{F}_g} \tag{3}$$

where $\lambda_{\mathcal{F}_g}$ is the best hypothesis risk defined as $\lambda_{\mathcal{F}_g} := \inf_{f\in\mathcal{F}}[R_S(f\circ g) + R_T(f\circ g)]$. The $\mathcal{F}\Delta\mathcal{F}$ divergence implicitly depends on the fixed g and can be small if g provides a suitable representation. However, if g induces a wrong alignment, then this divergence is large with any function class $\mathcal{F}$. The following example will illustrate such a situation, motivating to explicitly take the class of embeddings into account when bounding the target risk.

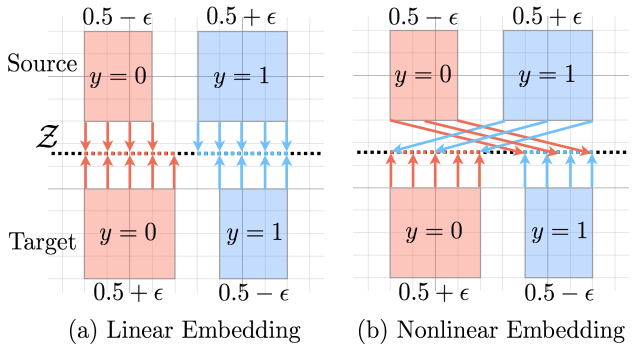


Figure 1. Illustrative example in 2D. The 1D representation space is illustrated as a dotted line and the arrows represent the embedding from 2D to 1D. (a) Optimal representations when $\mathcal{G}$ is the class of linear functions from 2D to 1D. (b) Optimal representation with a complex nonlinear function class with zero source and divergence loss; this representation destroys label consistency and leads to maximal target risk.

3. Influence of the representation

3.1. A Motivating example

We begin with an illustrative toy example. Figure 1 shows a binary classification problem in 2D with disjoint support and a slight shift in the label distributions from source to target: $p_S(y=1) = p_T(y=1) + 2\epsilon$. Assume the representation space is one dimensional, so the embedding g is a function from 2D to 1D. If we allow arbitrary, nonlinear embeddings, then, for instance, the embedding shown in Figure 1(b), together with an optimal predictor, achieves zero source loss and a zero divergence, and is hence optimal according to the loss (2). However, the target risk of this combination of embedding and predictor is maximal: $R_T(f\circ g) = 1$.

If we restrict the class $\mathcal{G}$ of embeddings to linear maps $g(x) = \mathbf{W}x$ where $\mathbf{W}\in\mathbb{R}^{1\times 2}$, then the embeddings that are optimal with respect to the loss 2 are of the form $\mathbf{W} = [a, 0]$. Together with an optimal source classifier f, they achieve loss 2ϵ due to the shift in class distributions. However, these embeddings retain label correspondences, and can lead to a zero target risk.

This example illustrates that a too rich class of embeddings can "overfit" the alignment, and hence lead to arbitrarily bad solutions. Hence, the complexity of the encoder class plays an important role in learning domain invariant representation too.

3.2. Bounds for Domain-invariant Representations

Motivated by the example, we next expose the dependency of the bound on the target risk on the complexity of the embedding class. To do so, we apply Theorem 1 to the hypothesis $h = fg := f \circ g$:

$$R_T(fg) \leq R_S(fg) + d_{\mathcal{FG}\Delta\mathcal{FG}}(p_S, p_T) + \lambda_{\mathcal{FG}}. \quad (4)$$

Comparing the bound (4) to the previous bound (3), we notice two differences: the best in-class joint risk now minimizes over both $\mathcal{F}$ and $\mathcal{G}$, i.e.,

$$\lambda_{\mathcal{FG}} := \inf_{f\in\mathcal{F}, g\in\mathcal{G}} [R_S(fg) + R_T(fg)], \quad (5)$$

which is smaller than $\lambda_{\mathcal{F}_g}$ and reflects the fact that we are learning both f and g. In return, the divergence term $d_{\mathcal{FG}\Delta\mathcal{FG}}(p_S, p_T)$ becomes larger than the one in bound (3). To better understand these tradeoffs, we derive a more interpetable form of the bound on the target risk.

Lemma 1. *For all $f \in \mathcal{F}$ and $g \in \mathcal{G}$:*

$$R_T(fg) \leq R_S(fg) + \lambda_{\mathcal{FG}} + \sup_{f'\in\mathcal{F}, g'\in\mathcal{G}} |R_S(fg, f'g') - R_T(fg, f'g')|. \quad (6)$$

Deatiled proofs of all lemmas and theorems may be found in the Appendix. Different from Theorem 1, we do not directly bound the third term with $\mathcal{H}\Delta\mathcal{H}$-divergence (taking supremum over f and g). To gain more insight, we further derive an interpretable upper-bound of the last term.

Lemma 2. *For all $f \in \mathcal{F}$ and $g \in \mathcal{G}$,*

$$\sup_{\substack{f'\in\mathcal{F}\\ g'\in\mathcal{G}}} |R_S(fg, f'g') - R_T(fg, f'g')| \leq \sup_{\substack{f'\in\mathcal{F}\\ g'\in\mathcal{G}}} R_S(f'g, f'g') + \sup_{\substack{f'\in\mathcal{F}\\ g'\in\mathcal{G}}} R_T(f'g, f'g') + d_{\mathcal{F}\Delta\mathcal{F}}(p_S(Z_g), p_T(Z_g)) \quad (7)$$

This lemma shows that the third term can be bounded by the $\mathcal{F}\Delta\mathcal{F}$-divergence between latent distributions and the maximum disagreement of $f'g$ and $f'g'$ in souce and target domains. Combining Lemma 1 and Lemma 2 yields the following.

Theorem 2. *For all $f \in \mathcal{F}$ and $g \in \mathcal{G}$,*

$$R_T(fg) \leq R_S(fg) + \underbrace{d_{\mathcal{F}\Delta\mathcal{F}}(p_S(Z_g), p_T(Z_g))}_{\text{(i) } \mathcal{F}\Delta\mathcal{F}\text{-divergence}} + \underbrace{\sup_{\substack{f'\in\mathcal{F}\\ g'\in\mathcal{G}}} R_S(f'g, f'g') + \sup_{\substack{f'\in\mathcal{F}\\ g'\in\mathcal{G}}} R_T(f'g, f'g') + \lambda_{\mathcal{FG}}}_{\text{(ii) Complexity Trade-off}}. \quad (8)$$

The first term of the bound is the source risk. The second term (i) is the $\mathcal{F}\Delta\mathcal{F}$-divergence between representation distribution $p_S(Z_g)$ and $p_T(Z_g)$ which also appears in the previous bound 3. The first two terms in (ii) measure the worst case disagreement of encoders which may decrease by reducing the encoder class $\mathcal{G}$. However, the best hypothesis risk $\lambda_{\mathcal{FG}}$ can increase with reduced capacity of $\mathcal{G}$. Therefore, (ii) makes a trade-off explicit between the expected disagreement and the model complexity.

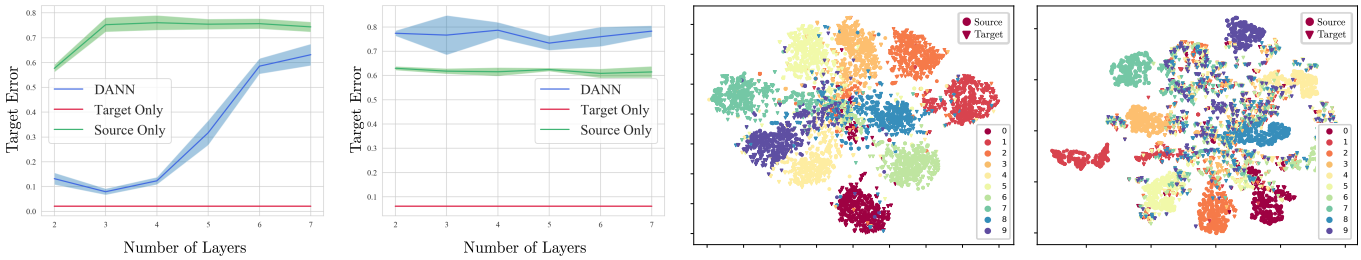
The last two terms in Theorem 1 also result in similar hypothesis complexity trade-off in (ii). However, with the fact that the hypothesis is composed with a encoder and a predictor, the conclusion drawn by simply applying Theorem 1 (Equation 4) is limited. It only implies that passively reducing the complexity of encoder and predictor may make the bound tighter. However, Theorem 2 shows that we can also make the bound tighter by minimizing the $\mathcal{F}\Delta\mathcal{F}$-divergence between latent distributions. Comparing to Equation 3, we do not assume a known encoder and instead quantify the effect of the encoder family, uncovering the missing complexity trade-off term. Without constraining the embedding-complexity, (ii) could be considerably large and result in a loose bound on target risk. We argue that restricting the model complexity is crucial for domain invariant representation learning.

3.3. Practical Implications

To reduce the worst case divergence (i), we should restrict the encoder family to those that can approximately minimize (i). Practically we can optimize the original objective of domain invariant representation showed in equation 2 to regularize the divergence of latent distribution induced by the encoder. Term (ii) implies that we should choose the minimal complexity encoder class $\mathcal{G}$ which is is still expressive enough to encode the data from both domains. Practically, this can be done by regularizing the encoder, e.g., restricting Lipschitz constants or norms of weight matrices. More explicitly, we can select the number of layers of a neural network or applying inductive bias via selecting network architectures. For instance, comparing to fully connected networks (FCs), convolutional neural networks (CNNs) restrict the output to be spatially consistent with respect to the input which has been applied in numerous computer vision tasks. Therefore, we can also interpret inductive biases as a form of complexity restriction.

4. Experiments

To empirically examine our theorems, we implement Domain-Adversarial Neural Networks (DANN) by Ganin et al. (2016) which minimize latent divergence based on domain discriminator (Jensen-Shannon divergence). We



(a) Target Error (CNN) (b) Target Error (MLP) (c) t-SNE (CNN) (d) t-SNE (MLP)

Figure 2. Experiment Results on MNIST → M-MNIST. (a) (b): Target error with respect to number of encoder's layers. Complexities of encoders have a direct impact on the target errors in CNN case. However, without inductive bias, DANNs with MLP encoder consistently perform worse than hypotheses only trained in source domain. (c) (d): t-SNE projections of representations with different inductive biases. CNN encoders result in target representations that are well align with those in source domain. However, MLP encoders lose label-consistency while minimizing the latent divergence between domains.

evaluate the performance on standard benchmark of domain adaptation MNIST → MNIST-M (Ganin & Lempitsky (2014)). The task is to classify unlabeled handwritten digits overlayed with random photographs (MNIST-M) based on labeled images of digits alone (MNIST). Here we consider two categories of complexity: number of layers and the inductive bias.

4.1. Number of Layers of Encoder

To analyze the effect of the encoder's complexity, we augment the original 2 layers CNN encoders with 1 to 5 additional CNN layers and leave the other setting unchanged. We retrain each model for 5 times and plot the mean and standard deviation of target error with respect to the number of layers in Figure 2 (a). We can see that the target error decreases at the beginning and then increases when more layers are added. This corroborates our theory: the CNN encoder without additional layers does not have enough expressive power which consequently makes the best hypothesis risk term $\lambda_{\mathcal{FG}}$ larger. However, when more layers are added, the complexity increases and subsequently makes the disagreements larger.

4.2. Inductive Bias of Encoder

To investigate the importance of inductive bias in domain invariant representation, we replace the CNN encoder with MLP one. The experiment result are shown in 2 (b). Comparing the target error between (a) and (b) in Figure 2, we can see that the target error with MLP encoder is significantly higher than CNN encoder's. Comparing to CNNs which encode invariancy via pooling and learned filters, MLPs do not assume any inductive bias and therefore have more complexity, which lead to worse performance. We essentially show that without inductive bias, learning domain invariant representation can even worsen the performance. To gain deeper insight, we use t-SNE (Maaten & Hinton (2008)) projection to visualize source and target domain representation distributions in Figure 2 (c)(d). We can see that with the inductive bias of CNNs, the representations of target domain well align with those in source domain. However, despite the overlap between source and target domains, MLP encoder will result in serious label-mismatch.

4.3. Discussion

We show that the complexity of encoder has direct effect on the target error. A more complex encoder class will lead to looser bound on target error, resulting in a significant performance drop in target domain. On the other hand, the inductive biases are also crucial for learning domain-invariant representations. With proper inductive bias such as CNNs, DANN can still achieve higher performance than the one with MLP encoder even the number of layers of the encoder is double. CNNs have been applied to various vision tasks and are themselves suitable for digit recognition. However, explicit supervision may be required to identify the encoder class when we have no prior knowledge about the task (Motiian et al., 2017; Chen & Chien, 2015).

5. Conclusion

We attempt to explain the success of domain invariant representation for unsupervised domain adaptation. We theoretically and empirically show that restricting the representation encoder is necessary for successful adaptation which is often overlooked by previous works. Without carefully selecting the encoder class, learning domain invariant representation might even harm the performance. For future works, analyses of different complexity regularization and inductive biases could provide more insights into the role of embedding-complexity. Understanding the strength of domain invariant representations in other modalities such as natural language and speech is also an interesting direction.

References

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pp. 137–144, 2007.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Chen, H.-Y. and Chien, J.-T. Deep semi-supervised learning for domain adaptation. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE, 2015.

Cortes, C. and Mohri, M. Domain adaptation in regression. In *International Conference on Algorithmic Learning Theory*, pp. 308–323. Springer, 2011.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014.

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*, 17(1):2096–2030, 2016.

Johansson, F. D., Ranganath, R., and Sontag, D. Support and invertibility in domain-invariant representations. *arXiv preprint arXiv:1903.03448*, 2019.

Long, M., Cao, Y., Wang, J., and Jordan, M. I. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov): 2579–2605, 2008.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Motiian, S., Jones, Q., Iranmanesh, S., and Doretto, G. Few-shot adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 6670–6680, 2017.

Shen, J., Qu, Y., Zhang, W., and Yu, Y. Wasserstein distance guided representation learning for domain adaptation. *arXiv preprint arXiv:1707.01217*, 2017.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, 2017.

Wu, Y., Winston, E., Kaushik, D., and Lipton, Z. Domain adaptation with asymmetrically-relaxed distribution alignment. *arXiv preprint arXiv:1903.01689*, 2019.

Zhao, H., Combes, R. T. d., Zhang, K., and Gordon, G. J. On learning invariant representation for domain adaptation. *arXiv preprint arXiv:1901.09453*, 2019.

Appendix for The Role of Embedding-complexity in Domain-invariant Representations

Anonymous Authors[1]

1. Proofs

Theorem 1. (Ben-David et al., 2010) *For all hypotheses $h \in \mathcal{H}$, the target risk is bounded as*

$$R_T(h) \leq R_S(h) + d_{\mathcal{H}\Delta\mathcal{H}}(p_S, p_T) + \lambda_{\mathcal{H}},$$

where $\lambda_{\mathcal{H}}$ is the best joint in-class joint risk

$$\lambda_{\mathcal{H}} \coloneqq \inf_{h \in \mathcal{H}} [R_S(h) + R_T(h)]$$

Proof. The proof relies on the triangle inequality for classification error and the definition of $\mathcal{H}\Delta\mathcal{H}$-divergence.

$$\begin{aligned}
R_T(h) &\leq R_T(h^\star) + R_T(h, h^\star) && (1)\\
&\leq R_T(h^\star) + R_S(h, h^\star) + |R_S(h, h^\star) - R_T(h, h^\star)| && (2)\\
&\leq R_T(h^\star) + R_S(h, h^\star) + \sup_{h, h' \in \mathcal{H}} |R_S(h, h') - R_T(h, h')| && (3)\\
&\leq R_T(h^\star) + R_S(h, h^\star) + d_{\mathcal{H}\Delta\mathcal{H}(p_S, p_T)} && (4)\\
&\leq R_T(h^\star) + R_S(h) + R_S(h^\star) + d_{\mathcal{H}\Delta\mathcal{H}(p_S, p_T)} && (5)\\
&= R_S(h) + \lambda_{\mathcal{H}} + d_{\mathcal{H}\Delta\mathcal{H}(p_S, p_T)} && (6)
\end{aligned}$$

□

For simplicity, we replace the notation of function composition $f \circ g$ with fg in the following sections.

Lemma 1. *For all $f \in \mathcal{F}$ and $g \in \mathcal{G}$:*

$$R_T(fg) \leq R_S(fg) + \lambda_{\mathcal{F}\mathcal{G}} + \sup_{f' \in \mathcal{F}, g' \in \mathcal{G}} |R_S(fg, f'g') - R_T(fg, f'g')|. \tag{7}$$

Proof. The proof relies on the triangle inequality for classification error (Ben-David et al. (2007)).

$$\begin{aligned}
R_T(fg) &\leq R_T(f^\star g^\star) + R_T(fg, f^\star g^\star) && (8)\\
&\leq R_T(f^\star g^\star) + R_S(fg, f^\star g^\star) + |R_S(fg, f^\star g^\star) - R_T(fg, f^\star g^\star)| && (9)\\
&\leq R_T(f^\star g^\star) + R_S(fg, f^\star g^\star) + \sup_{f' \in \mathcal{F}, g' \in \mathcal{G}} |R_S(fg, f'g') - R_T(fg, f'g')| && (10)\\
&\leq R_T(f^\star g^\star) + R_S(fg) + R_S(f^\star g^\star) + \sup_{f' \in \mathcal{F}, g' \in \mathcal{G}} |R_S(fg, f'g') - R_T(fg, f'g')| && (11)\\
&= R_S(fg) + \lambda_{\mathcal{F}\mathcal{G}} + \sup_{f' \in \mathcal{F}, g' \in \mathcal{G}} |R_S(fg, f'g') - R_T(fg, f'g')| && (12)
\end{aligned}$$

□

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Lemma 2. *For all $f \in \mathcal{F}$ and $g \in \mathcal{G}$,*

$$\sup_{\substack{f' \in \mathcal{F} \\ g' \in \mathcal{G}}} |R_S(fg, f'g') - R_T(fg, f'g')| \leq \sup_{\substack{f' \in \mathcal{F} \\ g' \in \mathcal{G}}} R_S(f'g, f'g') + \sup_{\substack{f' \in \mathcal{F} \\ g' \in \mathcal{G}}} R_T(f'g, f'g') + d_{\mathcal{F}\Delta\mathcal{F}}(p_S(Z_g), p_T(Z_g)) \tag{13}$$

Proof. We first decompose the term with using the property of absolute value and supremum.

$$\sup_{f' \in \mathcal{F}, g' \in \mathcal{G}} |R_S(fg, f'g') - R_T(fg, f'g')| \tag{14}$$

$$= \sup_{f' \in \mathcal{F}, g' \in \mathcal{G}} |R_S(fg, f'g) + R_S(fg, f'g') - R_S(fg, f'g) - R_T(fg, f'g) - R_T(fg, f'g') + R_T(fg, f'g)| \tag{15}$$

$$\leq \sup_{f' \in \mathcal{F}, g' \in \mathcal{G}} (|R_S(fg, f'g) - R_T(fg, f'g)| + |R_S(fg, f'g') - R_S(fg, f'g)| + |R_T(fg, f'g') - R_T(fg, f'g)|) \tag{16}$$

$$\leq \sup_{f' \in \mathcal{F}} |R_S(fg, f'g) - R_T(fg, f'g)| + \sup_{f' \in \mathcal{F}, g' \in \mathcal{G}} |R_S(fg, f'g') - R_S(fg, f'g)| + \sup_{f' \in \mathcal{F}, g' \in \mathcal{G}} |R_T(fg, f'g') - R_T(fg, f'g)| \tag{17}$$

Before moving on, we first show the first term can be bounded by $\mathcal{F}\Delta\mathcal{F}$-divergence on representation space

$$\sup_{f' \in \mathcal{F}} |R_S(fg, f'g) - R_T(fg, f'g)| = \sup_{f' \in \mathcal{F}} |R_{g \circ S}(f, f') - R_{g \circ T}(f, f')| \tag{18}$$

$$\leq \sup_{f, f' \in \mathcal{F}} |R_{g \circ S}(f, f') - R_{g \circ T}(f, f')| \tag{19}$$

$$= d_{\mathcal{F}\Delta\mathcal{F}}(p_S(Z_g), p_T(Z_g)) \tag{20}$$

Then, we simplify the last two terms. Note that for $h_1, h_2, h_3 \in \mathcal{H}$,

$$|R_D(h_1, h_2) - R_D(h_1, h_3)| = |\, \mathbb{E}_{x \sim D}[\, \|h_1(x) - h_2(x)\| - \|h_1(x) - h_3(x)\| \,]\,| \tag{21}$$

$$\leq \mathbb{E}_{x \sim D}[\, |\, \|h_1(x) - h_2(x)\| - \|h_1(x) - h_3(x)\| \,|\,] \tag{22}$$

$$\leq \mathbb{E}_{x \sim D}[\, \|(h_1(x) - h_2(x)) - (h_1(x) - h_3(x))\| \,] \qquad \text{(By reverse triangle inequality)}$$

$$= \mathbb{E}_{x \sim D}[\, \|h_3(x) - h_2(x)\| \,] \tag{23}$$

$$= R_D(h_3, h_2) \tag{24}$$

Apply the bound above to the last two terms, we have

$$\sup_{f' \in \mathcal{F}, g' \in \mathcal{G}} |R_S(fg, f'g') - R_T(fg, f'g')| \leq d_{\mathcal{F}\Delta\mathcal{F}}(p_S(Z_g), p_T(Z_g)) + \sup_{f' \in \mathcal{F}, g' \in \mathcal{G}} R_S(f'g, f'g') + \sup_{f' \in \mathcal{F}, g' \in \mathcal{G}} R_T(f'g, f'g') \tag{25}$$

□

Theorem 2. *For all $f \in \mathcal{F}$ and $g \in \mathcal{G}$,*

$$R_T(fg) \leq R_S(fg) + \underbrace{d_{\mathcal{F}\Delta\mathcal{F}}(p_S(Z_g), p_T(Z_g))}_{\text{(i) } \mathcal{F}\Delta\mathcal{F}\text{-divergence}} + \underbrace{\sup_{\substack{f' \in \mathcal{F} \\ g' \in \mathcal{G}}} R_S(f'g, f'g') + \sup_{\substack{f' \in \mathcal{F} \\ g' \in \mathcal{G}}} R_T(f'g, f'g') + \lambda_{\mathcal{F}\mathcal{G}}}_{\text{(ii) Complexity Trade-off}} . \tag{26}$$

Proof. It is a consequence of Lemma 1 and Lemma 2. □

References

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pp. 137–144, 2007.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.