B. TECH. PROJECT REPORT

On

Multisource Wasserstein Distance based Domain Adaptation

BY

Saptarshi Ghosh

DISCIPLINE OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

DECEMBER 2019

Multisource Wasserstein Distance based Domain Adaptation

A PROJECT REPORT

Submitted in partial fulfillment of the
requirements for the award of the degree

of
BACHELOR OF TECHNOLOGY
in
COMPUTER SCIENCE AND ENGINEERING

Submitted by:
Saptarshi Ghosh

Guided by:
Dr. Surya Prakash, Associate Professor, Discipline of CSE, IIT Indore

INDIAN INSTITUTE OF TECHNOLOGY INDORE
December 2019

Declaration of Authorship

I hereby declare that the project entitled "Multi-source Wasserstein Distance based Domain Adaptation" submitted in partial fulfillment for the award of the degree of Bachelor of Technology in 'Computer Science and Engineering' completed under the supervision of Dr. Ke Yiping, Kelly, Assistant Professor, Computer Science and Engineering, NTU Singapore and Dr. Surya Prakash, Associate Professor, Computer Science and Engineering, IIT Indore is an authentic work.

Further, I declare that I have not submitted this work for the award of any other degree elsewhere.

Signed:

_____

Saptarshi Ghosh

Date:

Certificate

This is to certify that the B.Tech Project entitled "Multi-source Wasserstein Distance based Domain Adaptation" submitted by Saptarshi Ghosh, in partial fulfillment for the award of the degree of Bachelor of Technology in 'Computer Science and Engineering' embodies the work done by him at Computational Intelligence Lab, NTU, Singapore under the supervision of Dr. Ke Yiping, Kelly (NTU Singapore) and Dr. Surya Prakash (IIT Indore).

_____

Supervisor

Dr. Surya Prakash

Associate Professor,

Indian Institute of Technology, Indore

Date:

Acknowledgements

It is my privilege to express my gratitude to several persons who helped me directly or indirectly to conduct this research project work. I wish to express my heartful gratitude and deep indebtedness to my BTP guides **Dr. Surya Prakash** and **Dr. Ke Yiping, Kelly** for their sincere guidance and inspiration in completing this Project.

I sincerely appreciate **Mr. Wei Pengfei** for his cooperation and for his kind guidance and encouragement.

I would also like to thank all my friends and staff at IIT Indore, and NTU Computational Intelligence Lab for their technical support and enthusiastic help.

Finally, I would like to give special thanks to my parents for their consecutive support, understanding and love to me.

Abstract

While Domain Adaptation has been actively researched in the past years, most of them focus on Single Source Domain Adaptation. Most often, we have multiple datasets available which share a common feature space. As such, it is possible to learn feature representation across multiple sources which can lead to higher accuracies on Target Domain. We choose to work on unsupervised Domain Adaptation, wherein, it attempts to generalize a model learnt using multiple source domains to the Target Domain. The sources could have varying data distributions.

Inspired by Wasserstein GANs and Wasserstein Distance Guided Representation Learning for Domain Adaptation, we extend it for Domain Adaptation in Multiple Source Setting. To this end, we propose the Multi-Wasserstein Distance Based Neural Network (MWDNN). We also show an effective way to include weights for the different sources, which more often leads to a higher testing accuracies over the Target Dataset. We conduct extensive experiments over real world Datasets to demonstrate that the proposed MWDNN outperforms the state-of-the-art baselines.

Contents

Declaration of Authorship iii

Certificate v

Acknowledgements vii

Abstract ix

1 Introduction 1

2 Related Works 3

3 Theoretical Analysis 5
3.1 Generalization Bound for Multi-Source Wasserstein Distance . . . 5
3.2 Average Case Generalization Bound 9

4 Multi-source Wasserstein Distance based Neural Network 11

5 Comparison to other Multi-Source Approaches 15

6 Experimental Results 17
6.1 Amazon Review 17
6.2 Office-Caltech 19
6.3 Digits Dataset 20

7 Conclusion 21

Bibliography 25

Appendix 27

List of Figures

4.1 MWDNN Architecture 12

6.1 Wasserstein Distance : Kitchen 18
6.2 Wasserstein Distance : Books 18
6.3 Wasserstein Distance : Electronics 18

List of Tables

6.1 Accuracy comparison on Amazon Review Dataset 18
6.2 Accuracy comparison on Office-Caltech Dataset 19
6.3 Accuracy comparison on Digits Dataset 20

Chapter 1

Introduction

The success of Deep Learning based Models is partially attributed to rich datasets having large enough samples with abundant annotations.(Krizhevsky et al., 2012; Hinton et al., 2012; Russakovsky et al., 2015) Domain Adaptation caters to the scenario where sufficient labeled data is unavailable in the Target Domain as generating new datasets is prohibitively expensive. Domain Adaptation attempts to transfer knowledge from the source domains to the Target domain under presence of covariate shifts. Recently, Deep Learning Based Models have shown to perform significantly well for Domain Adaptation. (Glorot et al., 2011; Donahue et al., 2014; Yosinski et al., 2014; Bousmalis et al., 2016; Long et al., 2015; Ganin et al., 2016). However, few deal with multi-source scenarios. Furthermore, naive application of multi-source approaches is not guaranteed to improve upon learning good feature representation for good performance on Target Domain. Inclusion of multiple source may lead to worse performance on the Target Domain sometimes. (Han Zhao et al., 2017) To this end, we propose the Multi-Source Wasserstein Distance based Domain Adaptation. We also show an effective way to give weights to sources to overcome the above problem.

Adversarial based approaches for Unsupervised Domain Adaptation is gaining popularity in recent years. It builds upon the Generative Adversarial Network (GANs) (Goodfellow et al. 2014), which play a minimax game between two adversarial networks: the discriminator is trained to distinguish real data from the generated data, while the generator learns to generate high-quality data to fool the discriminator. It is intuitive to employ this minimax game for domain adaptation to make the source and target feature representations indistinguishable. Adversarial Approaches (Ganin et al. 2016; Tzeng et al. 2017; Han Zhao et al. 2017), use a domain classifier to reduce domain discrepancy through an

adversarial objective w.r.t domain classifier. But domain classifier suffers from the Vanishing Gradient Problem (Jian Shen et al., 2018). Once the domain classifier is sufficiently trained, it provides no further information for the feature extractor. This may lead to sub-optimal training. As shown by (Jian Shen et al. 2018; Arjovsky, Chintala, and Bottou 2017), A more reasonable solution would be to replace the domain discrepancy measure with Wasserstein distance, which provides more stable gradients even if two distributions are distant.

In this paper, we analyze the Generalization Bound for Multi-Wasserstein Distance based Neural Network (MWDNN). Our theoretical results build upon (Han Zhao et al., 2017) and (Jian Shen et al., 2018). Inspired by (Han Zhao et al., 2017), we generalize the bound for WDGRL for multiple source domains. Experiments on common Domain Adaptation Benchmarks demonstrate the superiority of MWDNN over other baselines, thus validating the effectiveness of the same.

Chapter 2

Related Works

The problem of effectively adapting a model from one domain to other can have different approaches for solution based on the problem setting. Depending on the availability of labels, domain adaptation may be classified into supervised, unsupervised and semi-supervised (Mei Wang, Weihong Deng 2018).

On basis of the feature space, DA may be categorized into homogeneous if the feature space of the source and target distributions are same and heterogeneous if they are different. Based on the solving approaches of DA it can be broadly classified into three categories namely Discrepancy Based (E. Tzeng, J. Hoffman 2015 etc.), Adversarial Based (E. Tzeng, J. Hoffman 2017 etc.) and Reconstruction Based (Z. Yi, H. Zhang, 2017 etc.).

Given a source domain with ground truth and a target domain without labels, the main problem statement of Unsupervised Domain Adaptation is to learn a model that performs well on target distribution. As the source and target domains have different distributions, the main aim is to reduce the domain shift between the two domains. Various metrics are used to define and reduce the distance between the domains like $H\Delta H$ divergence (Ben-David et al., 2010), K-L divergence, CORAL (Sun, Feng, and Saenko 2016) and Wasserstein distance (Jian Shen et al., 2018; Arjovsky, Chintala, and Bottou 2017). In this paper, we prove the superiority of Wasserstein distance in multi-source domain adaptation over other divergences.

Generative Adversarial Network (GANs) can also be used for domain adaptation as they have a common problem of solving a mini-max game.

Various GAN based methods are used for domain adaptation tasks and the main idea for using Wasserstein Distance for domain adaptation in (Jian Shen et al., 2018) came from the Wasserstein GAN itself (Arjovsky, Chintala, and Bottou 2017). Wasserstein Distance has also been very popular for solving problems of Optimal Transport (Courty, Flamary, and Tuia 2014; Courty et al. 2017).

Chapter 3

Theoretical Analysis

3.1 Generalization Bound for Multi-Source Wasserstein Distance

We first introduce the notations used in the paper.

Notations. Let $\mathcal{X}$ represent the input space, and a labeling function $f : \mathcal{X} \to [0, 1]$, that holds for all the k domains, where domain represents a distributions $\mathcal{D}$ on $\mathcal{X}$. A Hypothesis class H is the set of predictor functions, $\forall h \in H, h : \mathcal{X} \to [0, 1]$. The error of a hypothesis h w.r.t labeling function f is defined as: $\varepsilon_S(h, f) = \mathbb{E}_{x\sim\mathcal{D}_S}[|h(x) - f(x)|]$. We use the shorthand $\varepsilon_S(h) = \varepsilon_S(h, f)$. It is similarly defined for the target domain as well.

Let $\{\mathcal{D}_{S_i}\}_{i=1}^k$ and $\mathcal{D}_T$ be k source domains and the target domain, respectively. We also use $\{\mu_{s_i}\}_{i=1}^k$ and μ_t to represent the corresponding distributions of the domains on $\mathcal{X}$.

The Wasserstein metric is a distance measure between probability distributions on a given metric space (M, ρ), where $\rho(x, y)$ is a distance function for two instances x and y in the set M. The p-th Wasserstein distance between two Borel probability measures $\mathbb{P}$ and $\mathbb{Q}$ is defined as:

$$W_p(\mathbb{P}, \mathbb{Q}) = \Big(\inf_{\mu\in\Gamma(\mathbb{P},\mathbb{Q})} \int \rho(x, y)^p d\mu(x, y) \Big)^{\frac{1}{p}}$$

where $\mathbb{P}, \mathbb{Q} \in \{\mathbb{P} : \int \rho(x, y)^p d\mathbb{P}(x) < \infty, \forall y \in M\}$ are two probability measures on M with finite p-th moment, and $\Gamma(\mathbb{P}, \mathbb{Q})$ is the set of all measures on $M \times M$ with marginals $\mathbb{P}$ and $\mathbb{Q}$.

Kantorovich-Rubinstein theorem shows when M is separable, first Wasserstein Distance can be defined as (Villani 2008):

$$W_1(\mathbb{P}, \mathbb{Q}) = \sup_{||f||_L \leq 1} \mathbb{E}_{x\sim\mathbb{P}}[f(x)] - \mathbb{E}_{x\sim\mathbb{Q}}[f(x)] \tag{3.1}$$

where $\|f\|_L = \sup |f(x) - f(y)|/\rho(x, y)$ represents the Lipschitz semi-norm.

Definition 3.1. We re-define the Wasserstein Distance Function to find the distance between $\mathcal{D}_T$ and a set of source domains $\{\mathcal{D}_{S_i}\}_{i=1}^k$ as follows:

$$W(\mathcal{D}_T; \{\mathcal{D}_{S_i}\}_{i=1}^k) := \max_{i\in[k]} W(\mathcal{D}_T; \mathcal{D}_{S_i}) = \max_{i\in[k]} \sup_{||f||_L \leq 1} \mathbb{E}_{x\sim\mu_t}[f(x)] - \mathbb{E}_{x\sim\mu_{s_i}}[f(x)]$$

Let h^* be the optimal hypothesis that achieves the minimum combined risk, λ:

$$\lambda := \varepsilon_T(h^*) + \max_{i\in[k]} \varepsilon_{S_i}(h^*)$$

We now proceed to show that the error in Target Domain is bounded by the error across source Domain, whilst using the Wasserstein Metric. The Proofs are provided in the Appendix.

Lemma 3.1. *(Jian Shen, Yanru Qu, Weinan Zhang and Yong Yu 2018) Assume $\forall h \in H, h$ is K-Lipschitz continous for some K. Then the following holds:*

$$\varepsilon_T(h, h') \leq \varepsilon_S(h, h') + 2KW_1(\mu_t, \mu_s)$$

Theorem 3.2.

$$\varepsilon_T(h) \leq \max_{i\in[k]} \varepsilon_{S_i}(h) + 2KW(\mathcal{D}_T; \{\mathcal{D}_{S_i}\}_{i=1}^k) + \lambda \tag{3.2}$$

Remark. The Above theorem shows a Generalization Bound over the True distribution. However, most of the times, we do not have access to the True distributions

and correspondingly the True Error over the Source Domains. Hence, we proceed to show a bound over the Empirical Distribution.

Theorem 3.3. (Han Zhao *et al.* 2018) Let $\{\mathcal{D}_{S_i}\}_{i=1}^k$ be k source distributions over $\mathcal{X}$. Let H be a hypothesis class where $VC\dim(H) = d$. If $\{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k$ are the empirical distributions of $\{\mathcal{D}_{S_i}\}_{i=1}^k$ generated with m $i.i.d$ samples from each domain, then, for $\epsilon > 0$, we have:

$$\Pr\Big(\sup_{h\in H}\Big|\max_{i\in[k]}\varepsilon_{S_i}(h) - \max_{i\in[k]}\widehat{\varepsilon}_{S_i}(h)\Big| \geq \epsilon\Big) \leq 2k\Big(\frac{em}{d}\Big)^d \exp\left(-2m\epsilon^2\right)$$

Lemma 3.4. *((Bolley, Guillin, and Villani 2007), Theorem 2.1; (Redko, Habrard, and Sebban 2016), Theorem 1) Let μ be a probability measure in $\mathbb{R}^d$ satisfying $T_1(\lambda)$ inequality. Let $\hat{\mu} = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ be its associated empirical measure defined on a sample of independent variables $\{x_i\}_{i=1}^N$ drawn from μ. Then for any $d' > d$ and $\lambda' < \lambda$ there exists some constant N_0 depending on d' and some square exponential moment of μ such that for any $\epsilon > 0$ and $N \geq N_0 \max(\epsilon^{-(d'+2)}, 1)$*

$$\mathbb{P}[W_1(\mu, \hat{\mu}) > \epsilon] \leq \exp(-\frac{\lambda'}{2}N\epsilon^2)$$

where d', λ' can be calculated explicitly.

Combining Above Theorems and Lemma, we have the following Theorem:

Theorem 3.5. Under the Assumption of Lemma 3.4, Let $\mathcal{D}_T$ and $\{\mathcal{D}_{S_i}\}_{i=1}^k$ be the target distributions and k source distributions over $\mathcal{X}$. Let H be the hypothesis class where $VC\dim(H) = d$. If $\widehat{\mathcal{D}}_T$ and $\{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k$ are the empirical distributions of $\mathcal{D}_T$ and $\{\mathcal{D}_{S_i}\}_{i=1}^k$ generated with m $i.i.d.$ samples from each domain. then, for $0 < \delta < 1$, with probability of atleast $1 - \delta$, we have:

$$\varepsilon_T(h) \leq \max_{i\in[k]} \widehat{\varepsilon}_{S_i}(h) + \sqrt{\frac{1}{2m}\left(\log\frac{2k}{\delta} + d\log\frac{em}{d}\right)} + 2KW(\widehat{\mathcal{D}}_T; \{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k) + 4K\sqrt{\frac{2}{\lambda'}\log\left(\frac{1}{\delta}\right)} \cdot \left(\sqrt{\frac{1}{m}}\right) + \lambda \tag{3.3}$$

Remark. The Above Theorem shows the Worst Case Generalization Bound of Domain Adaptation using the Wasserstein Metric. The First Term measures the worst Case Accuracy over the Source Domains, whilst the third term measures the distance between the source and the target. For a successful Domain Adaptation, we hope to minimize both the Source Training Error, whilst decreasing the Wasserstein Distance among the source and Target as well. As described in (Han Zhao et al., 2017), the bound depends on the worst case source Domain. As such natively incorporating a source is not a good idea. We hope to tackle the problem in the next part, by giving weights to the sources.

3.2 Average Case Generalization Bound

We extend the definitions from Def 3.1 to include a convex combination α of the k sources.

Definition 3.2. Let $\alpha \in \mathbb{R}^k$ such that $\alpha \geq 0$ and $\sum_{i\in[k]} \alpha_i = 1$. Define $W_\alpha(\mathcal{D}_T; \{\mathcal{D}_{S_i}\}_{i=1}^k)$ as follows:

$$W_\alpha(\mathcal{D}_T; \{\mathcal{D}_{S_i}\}_{i=1}^k) := \sum_{i\in[k]} \alpha_i \cdot W(\mathcal{D}_T; \mathcal{D}_{S_i}) = \sum_{i\in[k]} \alpha_i \cdot \sup_{||f||_L \leq 1} \mathbb{E}_{x\sim\mu_t}[f(x)] - \mathbb{E}_{x\sim\mu_{s_i}}[f(x)]$$

Let h_α^* be the optimal hypothesis that achieves the minimum combined risk, λ_α:

$$\lambda_\alpha := \varepsilon_T(h^*) + \sum_{i\in[k]} \alpha_i \cdot \varepsilon_{S_i}(h^*)$$

Theorem 3.6.

$$\varepsilon_T(h) \leq \sum_{i\in[k]} \alpha_i \cdot \varepsilon_{S_i}(h) + 2KW_\alpha(\mathcal{D}_T; \{\mathcal{D}_{S_i}\}_{i=1}^k) + \lambda_\alpha \tag{3.4}$$

Theorem 3.7. (Han Zhao *et al.* 2018) Let $\{\mathcal{D}_{S_i}\}_{i=1}^k$ be k source distributions over $\mathcal{X}$. Let H be a hypothesis class where $VC\dim(H) = d$. If $\{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k$ are the empirical distributions of $\{\mathcal{D}_{S_i}\}_{i=1}^k$ generated with m $i.i.d$ samples from each domain, then, for $\epsilon > 0$, we have:

$$\Pr\Big(\sup_{h\in H}\Big|\sum_{i\in[k]} \alpha_i \cdot \varepsilon_{S_i}(h) - \sum_{i\in[k]} \alpha_i \cdot \widehat{\varepsilon}_{S_i}(h)\Big| \geq \epsilon\Big) \leq 2k\Big(\frac{em}{d}\Big)^d \exp\left(-2m\epsilon^2\right)$$

Theorem 3.8. Under the Assumption of Lemma 3.4, Let $\mathcal{D}_T$ and $\{\mathcal{D}_{S_i}\}_{i=1}^k$ be the target distributions and k source distributions over $\mathcal{X}$. Let H be the hypothesis class where $VC\dim(H) = d$. If $\widehat{\mathcal{D}}_T$ and $\{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k$ are the empirical distributions of $\mathcal{D}_T$ and $\{\mathcal{D}_{S_i}\}_{i=1}^k$ generated with m $i.i.d.$ samples from each domain. then, for $\alpha \in \mathbb{R}^k, \alpha \geq 0, \sum_{i\in[k]} \alpha_i = 1$, for $0 < \delta < 1$, with probability of atleast $1-\delta$, we have:

$$\varepsilon_T(h) \leq \sum_{i\in[k]} \alpha_i \cdot \widehat{\varepsilon}_{S_i}(h) + \sqrt{\frac{1}{2m}\left(\log\frac{2k}{\delta} + d\log\frac{em}{d}\right)} + 2KW_\alpha(\widehat{\mathcal{D}}_T; \{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k) + 4K\left(\sqrt{\frac{2}{\lambda'}\log\left(\frac{1}{\delta}\right)} \cdot \left(\sqrt{\frac{1}{m}}\right)\right) + \lambda_\alpha \tag{3.5}$$

Since the Wasserstein Distance between the source i and target represents a distance between the two distributions, we choose to set the weights in proportion to their Wasserstein Distances. Setting

$$\alpha_i \propto 1 - \frac{W(\widehat{\mathcal{D}}_T; \widehat{\mathcal{D}}_{S_i})}{\sum_{j\in[k]} W(\widehat{\mathcal{D}}_T; \widehat{\mathcal{D}}_{S_j})} \tag{3.6}$$

It gives higher weights to the source with the least Wasserstein Distance. We discuss the implications of our choice of weights in Chapter 5.

Chapter 4

Multi-source Wasserstein Distance based Neural Network

Next, we describe the basic architecture used to optimize the generalization bound.

Consider, we are given labeled samples from k source domains, each containing m instance-label pairs, along with m unlabeled instances from target domain. The Hypothesis class H is dependent on the choice of architecture, i.e. no. of layers, nodes in each layer etc. Deciding the architecture of the Neural Network, fixes the Hypothesis Class H, which in turn fixes the square root terms and combined risk error λ_a. We can only hope to minimize the weighted source training error and the Wasserstein distance between the source domains and target domain. We choose to work on the average generalization bound since the worst case generalization bound minimizes error for only one source at a time. (Han Zhao et al., 2017) showed that the soft-MDAN performed better than the hard version.

The Goal is to learn a transferable classifier over H that minimizes $\varepsilon_T(h) = \mathbb{E}_{x\sim\mathcal{D}_T}[|h(x) - f(x)|] = \Pr_{x\sim\mathcal{D}_T}[f(x) \neq h(x)], h \in H$.

The Architecture can be broken into roughly three parts : 1) Feature Extractor 2) Label Classifier and 3) Domain Discriminator. Key Point to be noted in a domain adaptation setting is that the extracted features are indistinguishable across the domains yet informative enough to perform classification accurately.

Every Domain Discriminator independently calculates the Wasserstein Metric between $\mathcal{D}_{S_i}$ and $\mathcal{D}_T$. Let's denote feature extractor, label classifier, and domain discriminator by the corresponding functions f_g, f_c, f_w with corresponding network parameters $\theta_g, \theta_c, \theta_w$. For every domain discriminator, the loss can be represented as :

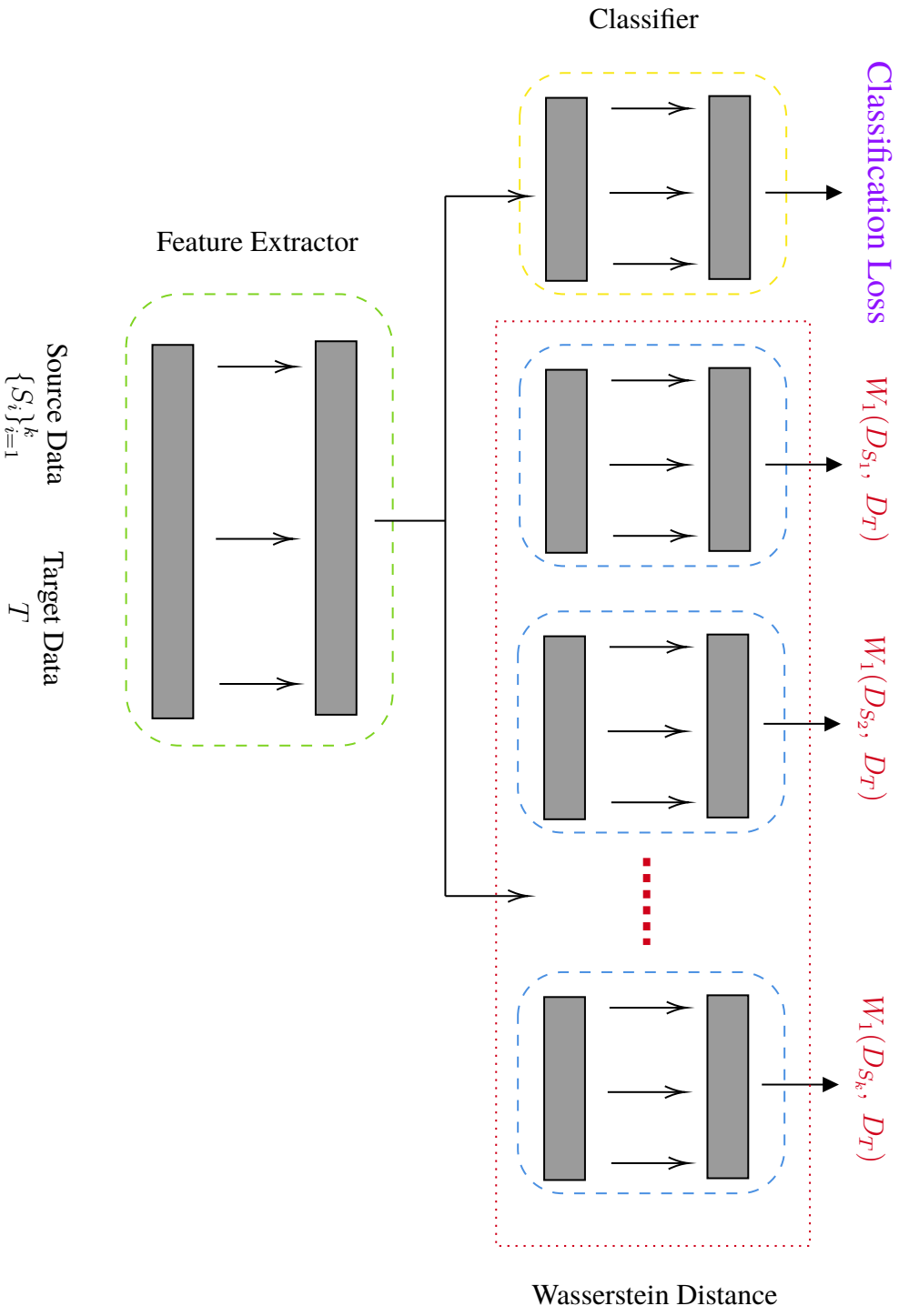

Figure 4.1: MWDNN Architecture

$$L_{wd_i} = \frac{1}{m} \sum_{x \sim \mathcal{D}_{S_i}} f_{w_i}(f_g(x)) - \frac{1}{m} \sum_{x \sim \mathcal{D}_T} f_{w_i}(f_g(x))$$

Common Activation Functions used in Neural Network like Sigmoid, ReLU, tanh etc. are Lipschitz continuous. As shown by (Jian Shen et al., 2018) and (Gulrajani et al. 2017), a reasonable way to enforce the Lipschitz constraint is via gradient penalty. The gradient penalty is defined as

$$L_{grad_i}(\hat{h}) = (\|\nabla_{\hat{h}} f_{w_i}(\hat{h})\|_2 - 1)^2$$

where the feature representation $\hat{h}$ are defined not only at the source and target but along random points along the straight line between source and target representations. The Wasserstein distance can be estimated by solving the following objective:

$$\max_{\theta_{w_i}} \{L_{wd_i} - \gamma L_{grad_i}\}$$

Representing the classification Loss by L_{c_i} for every source domain, the final loss function which minimizes the average generalization bound can be written as :

$$\min_{\theta_g, \theta_c} \Big\{ \sum_{i \in [k]} \alpha_i \Big(L_{c_i} + \lambda \cdot \max_{\theta_w} (L_{wd_i} - \gamma L_{grad_i}) \Big) \Big\}$$

where, λ controls the balance between discriminative and transferable feature learning, whereas balancing coefficient γ should be set to 0, during minimizing phase.

Algorithm 1 Multi-Source Wasserstein Distance based Domain Adaptation

Require: source instance-label pair $\{X^{s_i}\}_{i=1}^{k}$; target instance X^t; coefficient $\lambda, \gamma, \eta = 0.7$; domain critic learning rate β_1; classifier & feature extractor learning rate β_2; batch size m

Initialize random weights $\theta_g, \theta_c, \{\theta_{w_i}\}_{i=1}^{k}$
for $t = 1$ to ∞ **do**
  Sample batch $\{S_i^{(t)}\}_{i=1}^{k}$ and $T^{(t)}$ from $\{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^{k}$ and $\widehat{\mathcal{D}}_T$
  for $i = 1$ to k **do**
    for $j = 1$ to n **do**
      $h^{s_i} \leftarrow f_g(x^{s_i})$, $h^t \leftarrow f_g(x^t)$
      Sample h as random points along straight line h^{s_i} and h^t
      $\hat{h} \leftarrow \{h^{s_i}, h^t, h\}$
      $\theta_{w_i} \leftarrow \theta_{w_i} + \beta_1 \nabla_{\theta_{w_i}} [L_{wd_i}(x^{s_i}, x^t) - \gamma L_{grad_i}(\hat{h})]$
    end for
  end for
  for $i = 1$ to k **do**
    $\alpha_i = \eta w\Big(1 - \frac{W(\widehat{\mathcal{D}}_T; \widehat{\mathcal{D}}_{S_i})}{\sum_{j\in[k]} W(\widehat{\mathcal{D}}_T; \widehat{\mathcal{D}}_{S_j})}\Big)^{(t)} + (1-\eta) w\Big(1 - \frac{W(\widehat{\mathcal{D}}_T; \widehat{\mathcal{D}}_{S_i})}{\sum_{j\in[k]} W(\widehat{\mathcal{D}}_T; \widehat{\mathcal{D}}_{S_j})}\Big)^{(t-1)}$
  end for
  $\theta_c = \theta_c - \beta_2 \nabla_{\theta_c} \Big[\sum_{i\in[k]} \alpha_i \cdot L_c(x^{s_i}, y^{s_i})\Big]$
  $\theta_g = \theta_g - \beta_2 \nabla_{\theta_g} \Big[\sum_{i\in[k]} \alpha_i \cdot [L_c(x^{s_i}, y^{s_i}) + L_{wd_i}(x^{s_i}, x^t)]\Big]$
end for

The above algorithm can be implemented with a standard back-propagation based algorithm. w is a normalizing function that makes the sum of weights to 1. We used a sigmoid like function here which gives higher weights to sources with lower Wasserstein Distances, while penalizing the rest. The steep slope was used to harshly penalize some sources while minimizing the penalty in sources with lower Wasserstein Distance. Exponential Average was used to minimize the fluctuations in weights due to fluctuating Wasserstein Distances as approximated by the Domain Critic. The Wasserstein Distance Approximated by Neural Network fluctuates in practice, which leads to a fluctuations in weights as well.

Chapter 5

Comparison to other Multi-Source Approaches

Most of the Domain Adaptation in the literature focus mostly on single source setting. As such, few multi-sources approaches exist. The closest is the MDAN (Han Zhao et al., 2017), which extends the DANN (Ganin et al., 2016) to multiple sources. However as stated in (Jian Shen et al., 2018), the binary classifier used in DANN, suffers from vanishing Gradient Problem. Once it correctly begins to correctly classify the source and target, it provides no further useful information during training. However, Wasserstein Distance continues to provide stable gradients, thus avoiding the gradient vanishing problem.

We choose to work with the average case generalization bound as it presents a tighter bound than the worst case. The theoretical bound grows as $\mathcal{O}(\sqrt{\log k})$.

One thing that differentiates this method from previous approaches is the choice of weights. Earlier works (Han Zhao et al., 2017 etc.) try to find domain invariant feature representations, i.e. they try to find a common subspace of features amongst all domains. When multiple sources are to be considered, the choice of sources are crucial. Inclusion of a poor choice of source can negatively affect the performance on target domain. In our domain Adaptation setting, we are motivated to find the best performing model over the Target Domain. Hence, it is intuitive to filter out sources that can potentially affect the training negatively. In other words, instead of finding the common subspace, we try to find a subspace that is potentially better for the Target Domain.

It is reflected on our choice of weights

$$\alpha_i \propto 1 - \frac{W(\widehat{\mathcal{D}}_T; \widehat{\mathcal{D}}_{S_i})}{\sum_{j \in [k]} W(\widehat{\mathcal{D}}_T; \widehat{\mathcal{D}}_{S_j})}$$

Compare this to the weights assigned in the MDAN Model (Han Zhao et al., 2017):

$$\alpha_i = \frac{\varepsilon_i}{\sum_{j \in [k]} \varepsilon_j}$$

The choice of weights is latter assigns higher weights to sources with higher error. This approach is good when finding the common subspace, but does not guarantee the best performance over the Target Domain. In particular, it will continuously assign a higher weight to a poor choice of source. In contrast, our approach assigns the least weight to sources with highest error. It tries to align with sources which are common to the Target Domain. The continuous and exponential averaging of weights prevents abrupt changes in weights as well as leaves room for other sources to be assigned a higher weight should the distributions become similar later on during training.

We note that in some cases, our choice of weights may be so low that it effectively leads to a Single Source Domain Adaptation. However, in all cases, our choices perform better than previous approaches.

Chapter 6

Experimental Results

In this section, we prove our mathematical findings by testing on three standard domain adaptation datasets. We propose two variants of Multi-Source Wasserstein Distance algorithm, one unweighted and other with weights namely MSWD and w-MSWD respectively. The first is Amazon Review dataset which is a text based dataset and widely used for sentiment analysis. The second is Office-Caltech dataset (Gong et al., 2012) which is an image dataset and has 10 overlapping categories from Office and Caltech dataset. We worked on processed features on both the datasets. Last is the Digits Dataset namely, MNIST (LeCun et al., 1998), MNIST-M (Ganin et al., 2016), SVHN (Netzer et al., 2011), Synth-Digits (Ganin et al., 2016). The codes are implemented in Tensorflow 2.0 and Python 3.7.4.

6.1 Amazon Review

The dataset contains classes namely books, electronics, dvd and kitchen appliances. For each of the domains we have 2000 labelled and 4000 unlabelled reviews. We have compared our proposed approach of MSWD to various multi-source approaches as well as single source approaches. In multi-source we compared to MDAN (Han Zhao et al., 2017) and mSDA (Chen et al., 2012). For single source approaches we like B-DANN (Han Zhao et al., 2017; Ganin et al., 2016) and SWD (Jian Shen et al., 2018) we have taken the best case domain transfer accuracy for all domains.

We see that our method outperforms the compared approaches by a significant margin and w-MSWD gives best results.

Train/Test	mSDA	B-DANN	MDAN	SWD-Best	MSWD	w-MSWD
D+E+K/B	76.98	76.50	78.63	80.81	81.32	**81.926**
B+E+K/D	78.61	77.32	80.65	83.15	84.30	**84.412**
B+D+K/E	81.98	83.81	85.34	86.83	86.48	**87.267**
B+D+E/K	84.26	84.33	86.26	88.16	88.81	**88.966**
Average	80.46	80.49	82.72	84.74	85.23	**85.643**

Table 6.1: Accuracy comparison on Amazon Review Dataset

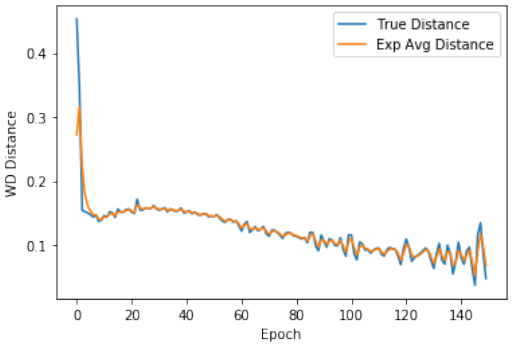

Figure 6.1: Wasserstein Distance : Kitchen

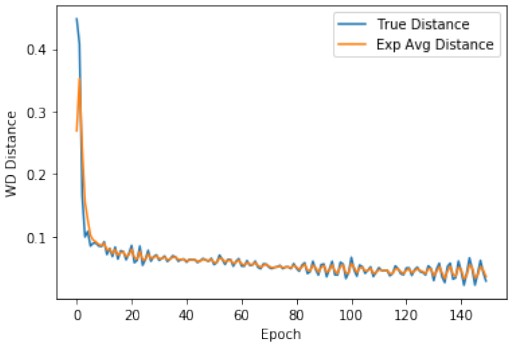

Figure 6.2: Wasserstein Distance : Books

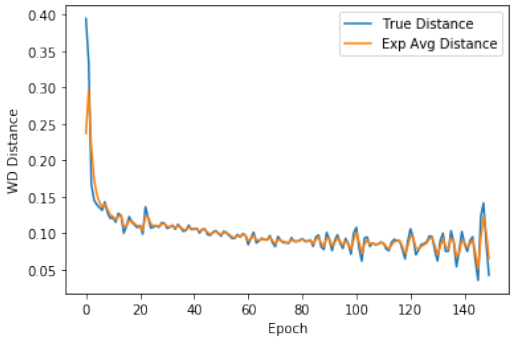

Figure 6.3: Wasserstein Distance : Electronics

Figure : Wasserstein Distance Loss : Target (DVD)

We note that Books and DVD are closely related, so are Kitchen and Electronics. While the Target is DVD, we note that Books has the least Wasserstein Distance as well, followed by Electronics then Kitchen.

The weights that were automatically assigned to the sources were Books ~ 0.7, Electronics ~ 0.2, Kitchen ~ 0.1. The steep slope of sigmoid like function in w, was used to increase the gap in the weights that would have been assigned otherwise.

6.2 Office-Caltech

The Office-Caltech dataset is a relatively small size dataset as compared to the amazon review dataset and contains 10 overlapping categories of the Office and Caltech datasets. The results on this data are in accordance with the findings of (Jian Shen et al., 2018) that using wasserstein distance as a metric on even less data points is much more stable than the compared approaches.

Train/Test	Source-only	M-DAN	SWD-Best	MSWD	w-MSWD
C+D+W/A	92.35	93.27	93.67	93.06	**93.93**
A+D+W/C	84.55	87.80	90.24	90.95	**91.06**
A+C+W/D	98.25	**100**	**100**	**100**	**100**
A+C+D/W	88	**98.95**	97.89	98.68	**98.95**
Average	90.79	95.00	95.45	95.67	**95.98**

Table 6.2: Accuracy comparison on Office-Caltech Dataset

We note that Webcam and DSLR are highly correlated. This is also verified by their individual source-only training. As such our algorithm gives higher weights to Webcam and DSLR pair. But, as training progresses, the wasserstein distance decreases across all sources. At later stages, the weights assigned become almost equal for all the sources. This also shows that our algorithm doesn't completely neglect the other sources.

6.3 Digits Dataset

Next, we test on the Digits Dataset namely, MNIST (Mt), MNIST-M (Mm), SVHN (Sv), Synth-Digits (Sy).

Train/Test	Best-Single Source	B-DANN	MDAN	w-MSWD
Mt+Sv+Sy/Mm	51.90	59.11	68.72	**94.12**
Mm+Sv+Sy/Mt	96.43	96.70	97.99	**98.35**
Mt+Mm+Sy/Sv	81.41	81.82	81.60	**82.76**
Average	76.58	79.21	82.77	**91.74**

Table 6.3: Accuracy comparison on Digits Dataset

Proceeding on a similar line, MNIST and MNIST-M are more related than others. In fact, MNIST-M is generated from MNIST itself. And Indeed, our algorithm does give higher weightage to Mt-Mm Source/Target pair.

We note that during Mt-Mm Source/Target Pair, the weights assigned to SVHN and Synth-Digits become as low as ~ 0.1 or less, thus significantly focusing on a single domain rather than finding domain invariant features across all sources. But even so, the testing accuracies over the Target Domain remain higher.

Chapter 7

Conclusion

The theoretical analysis as well as the experimental results on the three datasets prove the superiority of our proposed method in multi-source setting over both:-

- state of the art multi source methods like M-DAN
- best case single source domain adaptation methods

Also, our method is stable and invariant of the number of data-points as it performs equally well on small size as well as large size datasets, apart from being computationally inexpensive unlike CORAL and MMD. The choice of weights is unlike previous approaches, which focused more on finding invariant representations rather than focusing over the Target Domain.

Bibliography

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein gan". In: *arXiv preprint arXiv:1701.07875* (2017).

[2] Shai Ben-David et al. "A theory of learning from different domains". In: *Machine learning* 79.1-2 (2010), pp. 151–175.

[3] François Bolley, Arnaud Guillin, and Cédric Villani. "Quantitative concentration inequalities for empirical measures on non-compact spaces". In: *Probability Theory and Related Fields* 137.3-4 (2007), pp. 541–593.

[4] Konstantinos Bousmalis et al. "Domain separation networks". In: *Advances in neural information processing systems*. 2016, pp. 343–351.

[5] Minmin Chen et al. "Marginalized denoising autoencoders for domain adaptation". In: *arXiv preprint arXiv:1206.4683* (2012).

[6] Nicolas Courty, Rémi Flamary, and Devis Tuia. "Domain adaptation with regularized optimal transport". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2014, pp. 274–289.

[7] Nicolas Courty et al. "Optimal transport for domain adaptation". In: *IEEE transactions on pattern analysis and machine intelligence* 39.9 (2016), pp. 1853–1865.

[8] Jeff Donahue et al. "Decaf: A deep convolutional activation feature for generic visual recognition". In: *International conference on machine learning*. 2014, pp. 647–655.

[9] Yaroslav Ganin et al. "Domain-adversarial training of neural networks". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 2096–2030.

[10] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. "Domain adaptation for large-scale sentiment classification: A deep learning approach". In: *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011, pp. 513–520.

[11] Boqing Gong et al. “Geodesic flow kernel for unsupervised domain adaptation”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, pp. 2066–2073.

[12] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[13] Ishaan Gulrajani et al. “Improved training of wasserstein gans”. In: *Advances in neural information processing systems*. 2017, pp. 5767–5777.

[14] Geoffrey Hinton et al. “Deep neural networks for acoustic modeling in speech recognition”. In: *IEEE Signal processing magazine* 29 (2012).

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[16] Yann LeCun, Corinna Cortes, and Christopher JC Burges. “The MNIST database of handwritten digits, 1998”. In: *URL http://yann.lecun.com/exdb/mnist* 10 (1998), p. 34.

[17] Mingsheng Long et al. “Learning transferable features with deep adaptation networks”. In: *arXiv preprint arXiv:1502.02791* (2015).

[18] Yuval Netzer et al. “Reading digits in natural images with unsupervised feature learning”. In: (2011).

[19] Ievgen Redko, Amaury Habrard, and Marc Sebban. “Theoretical analysis of domain adaptation with optimal transport”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2017, pp. 737–753.

[20] Olga Russakovsky et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[21] Jian Shen et al. “Wasserstein distance guided representation learning for domain adaptation”. In: *arXiv preprint arXiv:1707.01217* (2017).

[22] Baochen Sun, Jiashi Feng, and Kate Saenko. “Return of frustratingly easy domain adaptation”. In: *Thirtieth AAAI Conference on Artificial Intelligence*. 2016.

[23] Eric Tzeng et al. “Adversarial discriminative domain adaptation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 7167–7176.

[24] Eric Tzeng et al. “Simultaneous deep transfer across domains and tasks”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 4068–4076.

[25] Mei Wang and Weihong Deng. “Deep visual domain adaptation: A survey”. In: *Neurocomputing* 312 (2018), pp. 135–153.

[26] Jason Yosinski et al. “How transferable are features in deep neural networks?” In: *Advances in neural information processing systems*. 2014, pp. 3320–3328.

[27] Han Zhao et al. “Multiple source domain adaptation with adversarial training of neural networks”. In: *arXiv preprint arXiv:1705.09684* (2017).

Appendix

Generalization Bound for Multi-Source Wasserstein Distance

This section presents the proofs of the various Theorem.

Definition 3.1. We re-define the Wasserstein Distance Function to find the distance between $\mathcal{D}_T$ and a set of source domains $\{\mathcal{D}_{S_i}\}_{i=1}^k$ as follows:

$$W(\mathcal{D}_T; \{\mathcal{D}_{S_i}\}_{i=1}^k) := \max_{i\in[k]} W(\mathcal{D}_T; \mathcal{D}_{S_i}) = \max_{i\in[k]} \sup_{||f||_L\leq 1} \mathbb{E}_{x\sim\mu_t}[f(x)] - \mathbb{E}_{x\sim\mu_{s_i}}[f(x)]$$

Let h^* be the optimal hypothesis that achieves the minimum combined risk, λ:

$$\lambda := \varepsilon_T(h^*) + \max_{i\in[k]} \varepsilon_{S_i}(h^*)$$

Lemma 3.1. *(Jian Shen, Yanru Qu, Weinan Zhang and Yong Yu 2018) Assume $\forall h \in H, h$ is K-Lipschitz continous for some K. Then the following holds:*

$$\varepsilon_T(h, h') \leq \varepsilon_S(h, h') + 2KW_1(\mu_t, \mu_s)$$

Theorem 3.2.

$$\varepsilon_T(h) \leq \max_{i\in[k]} \varepsilon_{S_i}(h) + 2KW(\mathcal{D}_T; \{\mathcal{D}_{S_i}\}_{i=1}^k) + \lambda$$

Proof.

$$\begin{aligned}
\varepsilon_T(h) &\leq \varepsilon_T(h^*) + \varepsilon_T(h, h^*) \\
&= \varepsilon_T(h^*) + \varepsilon_T(h, h^*) - \max_{i\in[k]} \varepsilon_{S_i}(h, h^*) + \max_{i\in[k]} \varepsilon_{S_i}(h, h^*) \\
&\leq \varepsilon_T(h^*) + |\varepsilon_T(h, h^*) - \max_{i\in[k]} \varepsilon_{S_i}(h, h^*)| + \max_{i\in[k]} \varepsilon_{S_i}(h, h^*) \\
&\leq \varepsilon_T(h^*) + 2KW(\mathcal{D}_T; \{\mathcal{D}_{S_i}\}_{i=1}^k) + \max_{i\in[k]} \varepsilon_{S_i}(h, h^*) \\
&\leq \varepsilon_T(h^*) + 2KW(\mathcal{D}_T; \{\mathcal{D}_{S_i}\}_{i=1}^k) + \max_{i\in[k]} \varepsilon_{S_i}(h) + \max_{i\in[k]} \varepsilon_{S_i}(h^*) \\
&= \max_{i\in[k]} \varepsilon_{S_i}(h) + 2KW(\mathcal{D}_T; \{\mathcal{D}_{S_i}\}_{i=1}^k) + \lambda
\end{aligned}$$

□

Theorem 3.3. (Han Zhao *et al.* 2018) Let $\{\mathcal{D}_{S_i}\}_{i=1}^k$ be k source distributions over $\mathcal{X}$. Let H be a hypothesis class where $VC\dim(H) = d$. If $\{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k$ are the empirical distributions of $\{\mathcal{D}_{S_i}\}_{i=1}^k$ generated with m $i.i.d$ samples from each domain, then, for $\epsilon > 0$, we have:

$$\Pr\Big(\sup_{h\in H}\Big|\max_{i\in[k]}\varepsilon_{S_i}(h) - \max_{i\in[k]}\widehat{\varepsilon}_{S_i}(h)\Big| \geq \epsilon\Big) \leq 2k\Big(\frac{em}{d}\Big)^d \exp\left(-2m\epsilon^2\right)$$

Proof.

$$\begin{aligned}
&\Pr\Big(\sup_{h\in H}\Big|\max_{i\in[k]}\varepsilon_{S_i}(h) - \max_{i\in[k]}\widehat{\varepsilon}_{S_i}(h)\Big| \geq \epsilon\Big)\\
\leq &\Pr\Big(\sup_{h\in H}\max_{i\in[k]}\Big|\varepsilon_{S_i}(h) - \widehat{\varepsilon}_{S_i}(h)\Big| \geq \epsilon\Big)\\
= &\Pr\Big(\max_{i\in[k]}\sup_{h\in H}\Big|\varepsilon_{S_i}(h) - \widehat{\varepsilon}_{S_i}(h)\Big| \geq \epsilon\Big)\\
\leq &\sum_{i=1}^k \Pr\Big(\sup_{h\in H}\Big|\varepsilon_{S_i}(h) - \widehat{\varepsilon}_{S_i}(h)\Big| \geq \epsilon\Big)\\
\leq &k\cdot\Pi_H(m)\Pr\left(|\varepsilon_{S_i}(h) - \widehat{\varepsilon}_{S_i}(h)| \geq \epsilon\right)\\
\leq &k\cdot\Pi_H(m)\cdot 2\exp\left(-2m\epsilon^2\right)\\
\leq &2k\Big(\frac{em}{d}\Big)^d \exp\left(-2m\epsilon^2\right)
\end{aligned}$$

□

Lemma 3.4. *((Bolley, Guillin, and Villani 2007), Theorem 2.1; (Redko, Habrard, and Sebban 2016), Theorem 1) Let μ be a probability measure in $\mathbb{R}^d$ satisfying $T_1(\lambda)$ inequality. Let $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}$ be its associated empirical measure defined on a sample of independent variables $\{x_i\}_{i=1}^{N}$ drawn from μ. Then for any $d' > d$ and $\lambda' < \lambda$ there exists some constant N_0 depending on d' and some square exponential moment of μ such that for any $\epsilon > 0$ and $N \geq N_0 \max(\epsilon^{-(d'+2)}, 1)$*

$$\mathbb{P}[W_1(\mu, \hat{\mu}) > \epsilon] \leq \exp(-\frac{\lambda'}{2}N\epsilon^2)$$

where d', λ' can be calculated explicitly.

Theorem 3.5. Under the Assumption of Lemma 3.4, Let $\mathcal{D}_T$ and $\{\mathcal{D}_{S_i}\}_{i=1}^k$ be the target distributions and k source distributions over $\mathcal{X}$. Let H be the hypothesis class where $VC\dim(H) = d$. If $\widehat{\mathcal{D}}_T$ and $\{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k$ are the empirical distributions of $\mathcal{D}_T$ and $\{\mathcal{D}_{S_i}\}_{i=1}^k$ generated with m $i.i.d.$ samples from each domain. then, for $0 < \delta < 1$, with probability of atleast $1 - \delta$, we have:

$$\varepsilon_T(h) \leq \max_{i\in[k]} \widehat{\varepsilon}_{S_i}(h) + \sqrt{\frac{1}{2m}\left(\log\frac{2k}{\delta} + d\log\frac{em}{d}\right)} + 2KW(\widehat{\mathcal{D}}_T; \{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k) + 4K\sqrt{\frac{2}{\lambda'}\log\left(\frac{1}{\delta}\right)} \cdot \left(\sqrt{\frac{1}{m}}\right) + \lambda$$

Proof. $\forall h \in H$, let $i_j := \arg\max_{i\in[k]} W(\mathcal{D}_T; \mathcal{D}_{S_i})$

$$\begin{aligned}
\varepsilon_T(h) &\leq \max_{i\in[k]} \varepsilon_{S_i}(h) + 2KW(\mathcal{D}_T; \{\mathcal{D}_{S_i}\}_{i=1}^k) + \lambda \\
&= \max_{i\in[k]} \varepsilon_{S_i}(h) + 2KW(\mathcal{D}_T; \mathcal{D}_{S_{i_j}}) + \lambda \\
&\leq \max_{i\in[k]} \varepsilon_{S_i}(h) + 2KW(\mathcal{D}_T; \widehat{\mathcal{D}}_T) + 2KW(\widehat{\mathcal{D}}_T; \mathcal{D}_{S_{i_j}}) + \lambda \\
&\leq \max_{i\in[k]} \varepsilon_{S_i}(h) + 2KW(\mathcal{D}_T; \widehat{\mathcal{D}}_T) + 2KW(\widehat{\mathcal{D}}_T; \widehat{\mathcal{D}}_{S_{i_j}}) + 2KW(\widehat{\mathcal{D}}_{S_{i_j}}; \mathcal{D}_{S_{i_j}}) + \lambda \\
&\leq \max_{i\in[k]} \varepsilon_{S_i}(h) + 2KW(\mathcal{D}_T; \widehat{\mathcal{D}}_T) + 2KW(\widehat{\mathcal{D}}_T; \{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k) + 2KW(\widehat{\mathcal{D}}_{S_{i_j}}; \mathcal{D}_{S_{i_j}}) + \lambda \\
&\leq \max_{i\in[k]} \varepsilon_{S_i}(h) + 2KW(\widehat{\mathcal{D}}_T; \{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k) + 2K\sqrt{\frac{2}{\lambda'}\log\left(\frac{1}{\delta}\right)} \cdot \left(\sqrt{\frac{1}{m}} + \sqrt{\frac{1}{m}}\right) + \lambda \\
&\leq \max_{i\in[k]} \varepsilon_{S_i}(h) + 2KW(\widehat{\mathcal{D}}_T; \{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k) + 4K\sqrt{\frac{2}{\lambda'}\log\left(\frac{1}{\delta}\right)} \cdot \left(\sqrt{\frac{1}{m}}\right) + \lambda \\
&\leq \max_{i\in[k]} \widehat{\varepsilon}_{S_i}(h) + \sqrt{\frac{1}{2m}\left(\log\frac{2k}{\delta} + d\log\frac{em}{d}\right)} + 2KW(\widehat{\mathcal{D}}_T; \{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k) \\
&\qquad + 4K\sqrt{\frac{2}{\lambda'}\log\left(\frac{1}{\delta}\right)} \cdot \left(\sqrt{\frac{1}{m}}\right) + \lambda
\end{aligned}$$

Average Case Generalization Bound

We extend the definitions from Def 3.1 to include a convex combination α of the k sources.

Definition 3.2. Let $\alpha \in \mathbb{R}^k$ such that $\alpha \geq 0$ and $\sum_{i\in[k]} \alpha_i = 1$. Define $W_\alpha(\mathcal{D}_T; \{\mathcal{D}_{S_i}\}_{i=1}^k)$ as follows:

$$W_\alpha(\mathcal{D}_T; \{\mathcal{D}_{S_i}\}_{i=1}^k) := \sum_{i\in[k]} \alpha_i \cdot W(\mathcal{D}_T; \mathcal{D}_{S_i}) = \sum_{i\in[k]} \alpha_i \cdot \sup_{||f||_L \leq 1} \mathbb{E}_{x\sim\mu_t}[f(x)] - \mathbb{E}_{x\sim\mu_{s_i}}[f(x)]$$

Let h_α^* be the optimal hypothesis that achieves the minimum combined risk, λ_α:

$$\lambda_\alpha := \varepsilon_T(h^*) + \sum_{i\in[k]} \alpha_i \cdot \varepsilon_{S_i}(h^*)$$

Theorem 3.6.

$$\varepsilon_T(h) \leq \sum_{i\in[k]} \alpha_i \cdot \varepsilon_{S_i}(h) + 2KW_\alpha(\mathcal{D}_T; \{\mathcal{D}_{S_i}\}_{i=1}^k) + \lambda_\alpha$$

Proof.

$$\begin{aligned}
\varepsilon_T(h) &\leq \varepsilon_T(h^*) + \varepsilon_T(h, h^*) \\
&= \varepsilon_T(h^*) + \Big(\sum_{i\in[k]} \alpha_i\Big) \cdot \varepsilon_T(h, h^*) - \sum_{i\in[k]} \alpha_i \cdot \varepsilon_{S_i}(h, h^*) + \sum_{i\in[k]} \alpha_i \cdot \varepsilon_{S_i}(h, h^*) \\
&= \varepsilon_T(h^*) + \sum_{i\in[k]} \alpha_i \cdot (\varepsilon_T(h, h^*) - \varepsilon_{S_i}(h, h^*)) + \sum_{i\in[k]} \alpha_i \cdot \varepsilon_{S_i}(h, h^*) \\
&\leq \varepsilon_T(h^*) + \sum_{i\in[k]} \alpha_i \cdot 2KW(\mathcal{D}_T; \mathcal{D}_{S_i}) + \sum_{i\in[k]} \alpha_i \cdot \varepsilon_{S_i}(h) + \sum_{i\in[k]} \alpha_i \cdot \varepsilon_{S_i}(h^*) \\
&= \sum_{i\in[k]} \alpha_i \cdot \varepsilon_{S_i}(h) + 2KW_\alpha(\mathcal{D}_T; \{\mathcal{D}_{S_i}\}_{i=1}^k) + \lambda_\alpha
\end{aligned}$$

□

Theorem 3.7. (Han Zhao *et al.* 2018) Let $\{\mathcal{D}_{S_i}\}_{i=1}^{k}$ be k source distributions over $\mathcal{X}$. Let H be a hypothesis class where $VC\dim(H) = d$. If $\{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^{k}$ are the empirical distributions of $\{\mathcal{D}_{S_i}\}_{i=1}^{k}$ generated with m $i.i.d$ samples from each domain, then, for $\epsilon > 0$, we have:

$$\Pr\Big(\sup_{h\in H}\Big|\sum_{i\in[k]}\alpha_i\cdot\varepsilon_{S_i}(h) - \sum_{i\in[k]}\alpha_i\cdot\widehat{\varepsilon}_{S_i}(h)\Big| \geq \epsilon\Big) \leq 2k\Big(\frac{em}{d}\Big)^{d}\exp\left(-2m\epsilon^2\right)$$

Theorem 3.8. Under the Assumption of Lemma 3.4, Let $\mathcal{D}_T$ and $\{\mathcal{D}_{S_i}\}_{i=1}^k$ be the target distributions and k source distributions over $\mathcal{X}$. Let H be the hypothesis class where $VC\dim(H) = d$. If $\widehat{\mathcal{D}}_T$ and $\{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k$ are the empirical distributions of $\mathcal{D}_T$ and $\{\mathcal{D}_{S_i}\}_{i=1}^k$ generated with m $i.i.d.$ samples from each domain. then, for $\alpha \in \mathbb{R}^k, \alpha \geq 0, \sum_{i\in[k]} \alpha_i = 1$, for $0 < \delta < 1$, with probability of atleast $1-\delta$, we have:

$$\varepsilon_T(h) \leq \sum_{i\in[k]} \alpha_i \cdot \widehat{\varepsilon}_{S_i}(h) + \sqrt{\frac{1}{2m}\left(\log\frac{2k}{\delta} + d\log\frac{em}{d}\right)} + 4K\left(\sqrt{\frac{2}{\lambda'}\log\left(\frac{1}{\delta}\right)}\cdot\left(\sqrt{\frac{1}{m}}\right)\right) + 2KW_\alpha(\widehat{\mathcal{D}}_T; \{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k) + \lambda_\alpha$$

Proof.

$$\begin{aligned}
\varepsilon_T(h) &\leq \sum_{i\in[k]} \alpha_i \cdot \varepsilon_{S_i}(h) + 2KW_\alpha(\mathcal{D}_T; \{\mathcal{D}_{S_i}\}_{i=1}^k) + \lambda_\alpha \\
&= \sum_{i\in[k]} \alpha_i \cdot \varepsilon_{S_i}(h) + 2K\sum_{i\in[k]} \alpha_i \cdot W(\mathcal{D}_T; \mathcal{D}_{S_i}) + \lambda_\alpha \\
&\leq \sum_{i\in[k]} \alpha_i \cdot \varepsilon_{S_i}(h) + 2K\sum_{i\in[k]} \alpha_i \cdot \Big(W(\mathcal{D}_T; \widehat{\mathcal{D}}_T) + W(\widehat{\mathcal{D}}_T; \widehat{\mathcal{D}}_{S_{i_j}}) + W(\widehat{\mathcal{D}}_{S_{i_j}}; \mathcal{D}_{S_{i_j}})\Big) + \lambda_\alpha \\
&\leq \sum_{i\in[k]} \alpha_i \cdot \varepsilon_{S_i}(h) + 2K\sum_{i\in[k]} \alpha_i \cdot \Big(W(\mathcal{D}_T; \widehat{\mathcal{D}}_T) + W(\widehat{\mathcal{D}}_{S_{i_j}}; \mathcal{D}_{S_{i_j}})\Big) + 2K\sum_{i\in[k]} \alpha_i \cdot W(\widehat{\mathcal{D}}_T; \widehat{\mathcal{D}}_{S_{i_j}}) + \lambda_\alpha \\
&\leq \sum_{i\in[k]} \alpha_i \cdot \varepsilon_{S_i}(h) + 4K\sum_{i\in[k]} \alpha_i \cdot \left(\sqrt{\frac{2}{\lambda'}\log\left(\frac{1}{\delta}\right)}\cdot\left(\sqrt{\frac{1}{m}}\right)\right) + 2K\sum_{i\in[k]} \alpha_i \cdot W(\widehat{\mathcal{D}}_T; \widehat{\mathcal{D}}_{S_{i_j}}) + \lambda_\alpha \\
&\leq \sum_{i\in[k]} \alpha_i \cdot \varepsilon_{S_i}(h) + 4K\left(\sqrt{\frac{2}{\lambda'}\log\left(\frac{1}{\delta}\right)}\cdot\left(\sqrt{\frac{1}{m}}\right)\right) + 2KW_\alpha(\widehat{\mathcal{D}}_T; \{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k) + \lambda_\alpha \\
&\leq \sum_{i\in[k]} \alpha_i \cdot \widehat{\varepsilon}_{S_i}(h) + \sqrt{\frac{1}{2m}\left(\log\frac{2k}{\delta} + d\log\frac{em}{d}\right)} + 4K\left(\sqrt{\frac{2}{\lambda'}\log\left(\frac{1}{\delta}\right)}\cdot\left(\sqrt{\frac{1}{m}}\right)\right) \\
&\qquad + 2KW_\alpha(\widehat{\mathcal{D}}_T; \{\widehat{\mathcal{D}}_{S_i}\}_{i=1}^k) + \lambda_\alpha
\end{aligned}$$