CLASS-CONDITIONED DOMAIN GENERALIZATION VIA WASSERSTEIN DISTRIBUTIONAL ROBUST OPTIMIZATION

Jingge Wang [*], **Yang Li** [*], **Liyan Xie** [†], **Yao Xie** [†]

ABSTRACT

Given multiple source domains, domain generalization aims at learning a universal model that performs well on any unseen but related target domain. In this work, we focus on the domain generalization scenario where domain shifts occur among class-conditional distributions of different domains. Existing approaches are not sufficiently robust when the variation of conditional distributions given the same class is large. In this work, we extend the concept of distributional robust optimization to solve the class-conditional domain generalization problem. Our approach optimizes the worst-case performance of a classifier over class-conditional distributions within a Wasserstein ball centered around the barycenter of the source conditional distributions. We also propose an iterative algorithm for learning the optimal radius of the Wasserstein balls automatically. Experiments show that the proposed framework has better performance on unseen target domain than approaches without domain generalization.

1 INTRODUCTION

The distribution shift between training and testing data, a.k.a. domain shift, is a common problem in many realistic applications. One way to alleviate the adverse impact of domain shift is through domain generalization, which aims to learn a universal model based on available source datasets and in total absence of target data (Ghifary et al., 2016; Motiian et al., 2017). Most existing methods learn a domain-invariant representation on source domains (Muandet et al., 2013; Ghifary et al., 2015; Motiian et al., 2017; Li et al., 2018a;b). However, these methods may encounter problems in certain challenging domain generalization scenarios. Consider a data model $Y \to X$ defined on $\mathcal{X} \times \mathcal{Y}$. Given class label Y, feature X is generated by conditional distributions $D_y(X) = P_D(X|Y = y)$ where D denotes either a source domain ($S_m, m = 1, \ldots, M$) or the target domain T. For a fixed $y \in \mathcal{Y}$, most domain generalization methods assume the class-conditional of target domain $T(X|Y = y)$ is closer to at least one of the M source class-conditionals $S_m(X|Y = y)$ than to any distributions of another class (Figure 1 Left). In an ideal case, even simple k-NN based method can perform well. However, when the variation among class-conditionals of the same class is large, i.e., the closest conditional distribution to $T(X|Y = y)$ is some $S(X|Y = y')$ of class $y' \neq y$ (Figure 1 Right), aforementioned methods may not perform well (Krueger et al., 2020).

In this work, we propose a class-conditioned domain generalization method inspired by the concept of distributional robust optimization (Kuhn et al., 2019), which optimizes the worst-case performance of a hypothesis over a set of distributions, namely the uncertainty set centered around the observed reference distributions. In the domain generalization context, we assume that class conditionals of different domains form class-specific uncertainty sets, aiming to learn a universally robust classifier that is even discriminative over the worst-case distributions in these sets. One challenge is to define a robust and computable uncertainty set. When there is only one source domain, Gao et al. (2018) defined the uncertainty set using Wasserstein distance and formulated the robust optimization problem as a convex optimization problem which can be solved efficiently. However, in

[*]Jingge Wang, Yang Li (corresponding). Tsinghua-Berkeley Shenzhen Institute, Tsinghua University. Emails: wangjg19@mails.tsinghua.edu.cn, yangli@sz.tsinghua.edu.cn

[†]Liyan Xie, Yao Xie. School of Industrial and Systems Engineering, Georgia Institute of Technology. Emails: lxie49@gatech.edu, yao.xie@isye.gatech.edu

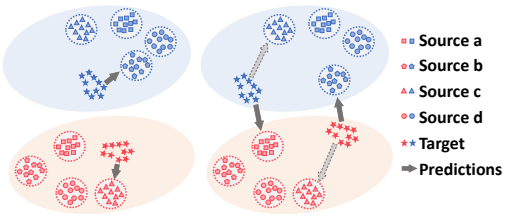

Figure 1: Classic and special setting with respect to class-conditional distributions (shaded circles indicate tasks, red and blue indicates different categories, best viewed in color.

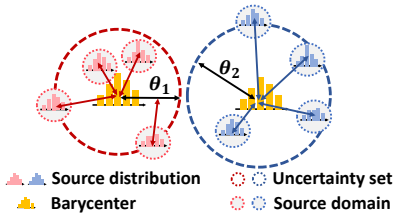

Figure 2: Initialization for θ_1, θ_2 using Wasserstein barycenter of source domain distributions.

domain generalization, there is no clear candidate for the reference distribution of each uncertainty set. Moreover, the radius of the Wasserstein uncertainty set, a fixed hyper-parameter in Gao et al. (2018), can largely impact the generalization performance. Our method uses Wasserstein barycenter as the reference distribution and an iterative algorithm for learning the optimal radius automatically.

2 Background

Gao et al. (2018) formulates the robust hypothesis testing as a minimax problem considering distributionally uncertainty, i.e. given two sets of distributions over X, $\mathcal{P}_1$ and $\mathcal{P}_2$, find distributions $P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2$ and detector ϕ that minimizes the maximum of type I and type II error by solving

$$\min_{\phi:\mathcal{X}\to\mathbb{R}} \max \left\{ \sup_{P_1\in\mathcal{P}_1} P_1\{x : \phi(x) < 0\}, \sup_{P_2\in\mathcal{P}_2} P_2\{x : \phi(x) \geq 0\} \right\}. \tag{1}$$

While there are various choices for the uncertainty sets, this work uses the Wasserstein distance to construct uncertainty ball of radius θ_k centered around empirical distribution $\mathcal{Q}_k$, i.e., $\mathcal{P}_k = \{P : \mathcal{W}(P, Q_k) \leq \theta_k\}, k = 1, 2$. Problem (1) can be transformed into an equivalent convex optimization problem of the following form

$$\min_{\phi} \max_{P_1\in\mathcal{P}_1, P_2\in\mathcal{P}_2} \Phi(\phi; P_1, P_2), \tag{2}$$

where Φ represents the risk under certain distribution P_1 and P_2. After interchanging the $\min$ and $\max$ operators, we can first get optimal detector ϕ^* for any given (P_1, P_2). Then what is left is optimizing the least-favorable distributions (LFDs) P_1^* and P_2^* with uncertainty set constraints. Reformulating Wasserstein metric constraints in (2) into equivalent linear constraints, the problem is finally transformed into a convex optimization problem.

3 Class-conditioned Domain Generalization

Suppose we have access to M diverse source domain with labeled data $\{(X_s^m, Y_s^m)\}, m = 1, \cdots, M$ which are representative of an underlying universal domain. Let $S_m(X|Y = y), m = 1, \cdots, M$ and $T(X|Y = y)$ denote class-conditional distributions for each class y in source and target domain, respectively. To construct class-specific uncertainty sets, the reference distribution and the radius need to be decided. Our method uses Wasserstein barycenter as the reference distribution, and introduces an iterative algorithm for learning the optimal radius. Without loss of generality, we consider the binary classification setting.

Estimation of Distribution Uncertainty Sets. Without access to any target domain data, we can no longer use the empirical distribution as the center of uncertainty sets. A natural choice for the center of an uncertainty set defined by the source domains is the 2-Wasserstein barycenter since it better capture the geometry among distributions (Rabin et al., 2011). Max value of all M distances between source distributions and the barycenter is taken as the initial uncertainty set radius. Therefore, we can construct the uncertainty Wasserstein ball $\mathcal{P}_1, \mathcal{P}_2$ of radius θ_1, θ_2, which can be seen as general class-conditioned domain, as shown in Figure 2.

Inference on Target Domain. Setting the generating function as exponential, the corresponding optimal detector is $\phi^* = 1/2 \log(P_1/P_2)$ (Gao et al., 2018). Using the uncertainty sets of radius θ_1,

Algorithm 1 Learning algorithm for class-conditioned domain generalization

Input: M diverse source tasks with labeled data $\mathcal{X}_s^m = \{(X_s^m, Y_s^m)\}, m = 1, \cdots, M$;
Output: The LFDs P_1^*, P_2^* supported on source task samples;
1. Initialization of θ_1, θ_2:
for each class y **do**
Barycenter distribution $C^*(X|y) \leftarrow \arg\min_{C(X|y)} \sum_{m=1}^{M} \frac{1}{M} \mathcal{W}_2^2 \left(C(X|y), S_m(X|y)\right)$;
Initial uncertainty set radius $\theta_y \leftarrow \max_{m=1,\cdots,M} \mathcal{W}_2 \left(C(X|y), S_m(X|y)\right)$;
end for
2. Dynamically Learning of θ_1, θ_2:
repeat
Solve $P_1^*, P_2^* \leftarrow \min_{T} \max_{P_1 \in \mathcal{P}_1(\theta_1), P_2 \in \mathcal{P}_2(\theta_2)} \Phi\left(T; P_1, P_2\right)$;
$\theta_1 \leftarrow \theta_1 - \Delta, \theta_2 \leftarrow \theta_2 - \Delta$;
until $\rho_\epsilon(P_1^*, P_2^*) < \gamma$

θ_2, problem (2) can produce worst-case distributions P_1^*, P_2^*, which are non-parametric functions of barycenter samples. To make inference on any target sample x_t, we define a weighted k-NN classifier as follows

$$\phi^*(x_t) = \begin{cases} 1, & \text{if } \frac{1}{K} \sum_{i=1}^{K} w_i \log\left(P_1^*(x_i)/P_2^*(x_i)\right) \geq 0 \\ 2, & \text{if } \frac{1}{K} \sum_{i=1}^{K} w_i \log\left(P_1^*(x_i)/P_2^*(x_i)\right) < 0, \end{cases} \tag{3}$$

where $x_1, \ldots, x_K$ are the K nearest neighbors of x_t measured by the Euclidean distance and w_i is inversely proportional to $\|x_i - x_t\|_2$. We use $K = 3$ in the experiments.

Dynamically Learning of θ_1, θ_2. In practice, the initial radius of Wasserstein balls tends to be too large and the optimal class-conditionals P_1^* and P_2^* may become indistinguishable. Thus we add one more constraint to ensure the LFDs are significantly different, using the chi-squared test, whose p-value is denoted as $\rho_\epsilon\left(P_1(\theta_1), P_2(\theta_2)\right)$, and the problem becomes as follows

$$\begin{aligned} &\min_{\phi} \max_{P_1 \in \mathcal{P}_1(\theta_1), P_2 \in \mathcal{P}_2(\theta_2)} \Phi\left(\phi; P_1, P_2\right) \\ &\text{s.t. } \rho_\epsilon\left(P_1(\theta_1), P_2(\theta_2)\right) < \gamma, \end{aligned} \tag{4}$$

where γ is the significance threshold taken as 0.05 in Chi-square testing. To avoid the difficulty of solving the optimization problem with non-convex constraint directly, we use a heuristic method to search for the optimal radius satisfying the constraint. As larger uncertainty set leads to less distinguishable LFDs, we initialize θ_1, θ_2 as the maximum Wasserstein distance between source distributions and the barycenter distribution. In each iteration, θ_1, θ_2 are dynamically decremented, i.e., $\theta_1 = \theta_1 - \Delta, \theta_2 = \theta_2 - \Delta$ with a small positive Δ, until P_1 and P_2 are significantly different. More details can be found in Algorithm 1.

4 Experimental Evaluation

Synthesized Data. We first evaluate our algorithm on data generated from Gaussian-like class-conditional distributions for each source and target domain. Setting four source domains and one target domain configured in a way similar to Figure 1 (right), we vary the source sample size from 20 to 200 and fix the target testing sample size at 60. Results shown in Figure 3 indicate that regardless of different sample size settings, the k-NN classifier trained by mixing all the source domain data shows far worse performance on the target testing data compared with our method. This simple experiment illustrates the potential of handling special data scenario of our method.

Real-world Data. We adopt five datasets collected from electric charge-discharge tests of power batteries under different experimental conditions. Using one dataset as the target domain and the other four as source domains, our goal is to decide whether a battery should be retired based on its charge-discharge features. For each sample, 17-dimension features represents the lab testing result, and binary label represents the battery state determined by its capacity level. By varying the capacity range of each battery state, we create six experimental scenarios with different classification difficulty levels denoted by numbers 1-6 in Table 1. The higher the level, the harder the task. To simulate

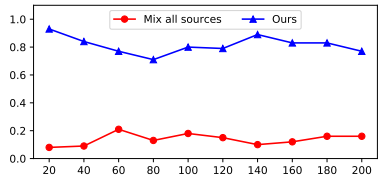

Figure 3: Average accuracy of unseen Gaussian data from target domain using source Gaussian data with different sample size.

Table 1: The range of for each class, data size and initial uncertainty set radius in varying difficulty settings for battery datasets.

	Difficultiy level					
	1	2	3	4	5	6
Capacity range	[0.95,1] [0,0.85]	[0.945,1] [0,0.855]	[0.94,1] [0,0.86]	[0.93,1] [0,0.87]	[0.92,1] [0,0.88]	[0.905,1] [0,0.895]
Training	16	18	20	24	28	38
Testing	84	92	102	122	146	196
θ_1	3.44	3.33	3.36	3.20	3.04	2.65
θ_2	4.02	3.96	3.63	3.04	2.80	2.37

Table 2: Comparison of binary classification accuracy under 6 difficulty grade settings.

Method		Difficulty level					
		1	2	3	4	5	6
Single domain	Target only (supervised)	**0.935**	0.927	0.925	**0.899**	0.850	**0.782**
	Source a only (unsupervised)	0.929	**0.934**	**0.927**	0.884	**0.872**	0.761
	Source b only (unsupervised)	0.224	0.218	0.184	0.184	0.228	0.270
	Source c only (unsupervised)	0.918	0.927	0.909	0.866	0.822	0.751
	Source d only (unsupervised)	0.083	0.085	0.078	0.108	0.160	0.247
Domain generalization	Source a+b+c+d (unsupervised)	0.525	0.538	0.451	0.476	0.469	0.478
	Source a+b+c+d+target (semi-supervised)	0.680	0.691	0.626	0.609	0.601	**0.568**
	Ours w/o radius learning(unsupervised)	0.740	0.670	0.608	0.402	0.599	0.517
	Ours w/ radius learning(unsupervised)	**0.806**	**0.795**	**0.765**	**0.681**	**0.628**	0.530

the realistic setting of scarce labeled data, in each trial we randomly sample 1/10 of the training set as source domain data and calculate the average accuracy over 20 trials for all experiments. Sampled training and testing sizes for each source domain are also shown in Table 1.

We compare our method and its truncated version without dynamically learning of θ on target testing data with the following baselines based on k-NN: (1) Target only: use k-NN on target training data; (2) Source a/b/c/d only: use data from one of the four source domains; (3) Source a+b+c+d: mix all source training data; (4) Source a+b+c+d+target: mix source and target training data together. The truncated version uses the initial θ shown in the last two rows in Table 1, which is used as initialization in the standard version.

Results in Table 2 shows that our proposed method outperforms the approach of mixing available data in both unsupervised and semi-supervised way. It is inferior to two source only methods but outperforms the other two. This shows the usability of our method especially when we have no idea which source is more similar with the target in an unsupervised setting. The truncated version only yields competitive results 0.740 compared with semi-supervised approach in the easiest grade, implying the necessity of learning for θ.

5 Conclusions

In this paper, we present a robust domain generalization method that can effectively learn a universal classifier invariant to domain shifts in the class conditional distributions. Testing on both synthesized and real-world data, we show that this method has promising performance using only limited source domains. In the future, we will extend our method to more learning scenarios, such as multi-class classification, and evaluate our framework on more real-world applications.

Acknowledgments

This research is funded by Natural Science Foundation of China 62001266.

REFERENCES

Rui Gao, Liyan Xie, Yao Xie, and Huan Xu. Robust hypothesis testing using wasserstein uncertainty sets. In *NeurIPS*, pp. 7913–7923, 2018.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pp. 2551–2559, 2015.

Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430, 2016.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pp. 130–166. INFORMS, 2019.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5400–5409, 2018a.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b.

Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5715–5725, 2017.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 435–446. Springer, 2011.