Graph-Relational Distributionally Robust Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Out-of-distribution (OOD) generalization is a challenging machine learning problem yet highly desirable in many high-stake applications. Distributionally robust optimization (DRO) is a promising learning paradigm to tackle this challenge but suffers from several limitations. To address this challenge, we propose graph-relational distributionally robust optimization that trains OOD-resilient machine learning models by exploiting the graph structure of data distributions. Our approach can uniformly handle both fully-known and partially-known graph structures. Empirical results on both synthetic and real-world datasets demonstrate the effectiveness and flexibility of our method.

1 Introduction

Recent years have witnessed a surge of applying machine learning (ML) in high-stake and safety-critical applications, such as health diagnoses and self-driving cars. Such applications pose an unprecedented *out-of-distribution (OOD) generalization challenge* [16]: ML models are constantly exposed to unseen distributions that lie outside their training space. Despite well-documented success for *interpolation*, modern ML models (*e.g.*, deep neural networks) are notoriously weak for *extrapolation*; a highly accurate model on average can fail catastrophically when presented with rare or unseen distributions [1]. Without addressing this challenge, ML models cannot be safely deployed.

A promising solution for out-of-distribution generalization is to conduct distributionally robust optimization (DRO) [13, 21, 11]. Different from empirical risk minimization (ERM) [24] that minimizes the average loss, DRO aims to optimize the *worst-case* generalization risk over a set of training groups. For instance, data with a similar distribution can compose a group [18]. However, it suffers from critical limitations. (1) DRO recklessly prioritizes the worst-case groups without assessing the risk that they might be outliers [27]; optimizing over outliers would fundamentally damage OOD generalization. (2) The worst-case groups are not necessarily the *influential* ones that are truly connected to unseen distributions; optimizing over the worst-case rather than influential groups would yield mediocre generalization performance.

To address these challenges, we propose a novel method for graph-relational distributionally robust optimization. Instead of the worst-case distributions, our key idea is to minimize the generalization risks over influential groups. Such influential groups can be identified by analyzing the graph of data distributions. Graph structures widely exist in the real world and can usually be represented by a graph. For instance, to capture the similarity of weather events in the U.S. [26], one can construct a graph where each state (group) realizes a node, and the physical adjacency between two states results in an edge. A significant merit of our approach is that it can uniformly handle various scenarios when the graph structure is either fully or partially available. Empirical results on both synthetic and real-world datasets demonstrate the superior performance of our method over SOTA.

Submitted to the Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

2 Related Work

Distributionally Robust Optimization. In the context of distributionally robust optimization (DRO), [3] and [20] argued that minimizing the maximal loss over a set of possible distributions can provide better generalization performance than minimizing the average loss. The robustness guarantee of DRO heavily relies on the quality of the uncertainty set which is typically constructed by moment constraints [2], f-divergence [13] or Wasserstein distance [19]. To avoid yielding overly pessimistic models [5], group DRO [8, 18] is proposed to leverage pre-defined data groups to formulate the uncertainty set as the mixture of these groups. However, none of these methods incorporate the physical prior that widely exists in real-world applications.

Out-of-Distribution Generalization. The goal of OOD generalization is to generalize models from source distributions to unseen target distributions. There are mainly two branches of methods to tackle OOD generalization: domain-invariant learning [1, 9, 12] and distributionally robust optimization. The goal of domain-invariant learning is to exploit the causally invariant correlations across multiple distributions. Invariant Risk Minimization (IRM) is one of the most representative methods which learns the optimal classifier across source distributions. However, recent work [17] shows IRM can fail catastrophically unless the test data are sufficiently similar to the training distribution.

3 Problem Formulation and Preliminary Works

Problem Formulation. We focus on the problem of out-of-distribution (OOD) generalization. Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ be the target space. (X, Y) are random variables defined over samples $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and the joint distribution $\mathbb{P}(X, Y)$. Since we cannot sample directly from $\mathbb{P}(X, Y)$, we usually assume data are drawn from a set of groups $\mathcal{E}_{\text{all}}$, where each group $e \in \mathcal{E}_{\text{all}}$ is sampled from a distinct distribution $\mathbb{P}(X^e, Y^e)$, *e.g.*, the distribution of medical images varies at different hospitals due to equipment or demographic differences. Let $\mathcal{E}_{\text{train}} \subsetneq \mathcal{E}_{\text{all}}$ be a finite subset of training groups, and assume that for each $e \in \mathcal{E}_{\text{train}}$, we have access to a dataset $\mathcal{D}^e := \left\{\left(x_j^e, y_j^e\right)\right\}_{j=1}^{n_e}$ sampled i.i.d. from $\mathbb{P}\left(X^e, Y^e\right)$. Given a function class $\mathcal{F}$ and a loss function ℓ, our goal is to learn a predictor $f \in \mathcal{F}$ using the data from $\mathcal{D}^e$ that minimizes the worst-case risk over the entire family of $\mathcal{E}_{\text{all}}$:

$$\min_{f \in \mathcal{F}} \max_{e \in \mathcal{E}_{\text{all}}} \mathbb{E}_{\mathbb{P}(X^e, Y^e)} \ell\left(f\left(X^e\right), Y^e\right). \tag{1}$$

It is challenging to learn a predictor $f \in \mathcal{F}$ that generalizes from the finite set of training domains $\mathcal{E}_{\text{train}}$ to perform well on the set of all domains $\mathcal{E}_{\text{all}}$ since we do not have access to data from any unseen group $e \in \mathcal{E}_{\text{test}}$, where $\mathcal{E}_{\text{test}} = \mathcal{E}_{\text{all}} \setminus \mathcal{E}_{\text{train}}$.

Empirical Risk Minimization (ERM) [24]. ERM minimizes the average loss over the distribution of all training groups $\mathcal{E}_{\text{train}}$:

$$\min_{f \in \mathcal{F}} \sum_{e=1}^{m} \mathbb{E}_{\mathbb{P}(X^e, Y^e)}[\ell\left(f\left(X^e\right), Y^e\right)],$$

where $m = |\mathcal{E}_{\text{train}}|$ is the number of training groups. Models trained via ERM heavily rely on spurious correlations that do not always hold under distributional drifts [1].

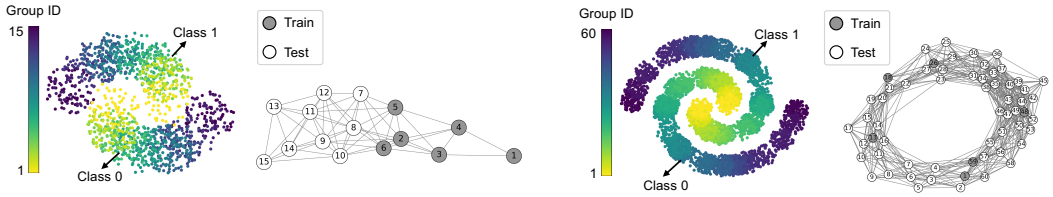
Distributionally Robust Optimization (DRO) [18]. Instead of minimizing the average loss, DRO minimizes the worst-combination loss of different training groups:

$$\min_{f \in \mathcal{F}} \max_{q \in \Delta_m} \sum_{e=1}^{m} q_e \mathbb{E}_{\mathbb{P}(X^e, Y^e)}[\ell\left(f\left(X^e\right), Y^e\right)], \tag{2}$$

where q is the mixture vector of $\mathcal{E}_{\text{train}}$ and $\Delta_m = \{q \in \mathbb{R}^m \mid \sum_{k=1}^{m} q_k = 1; \forall k, q_k \geq 0\}$. We empirically found that DRO blindly prioritizes the worst-case groups that incur higher losses than others. However, favoring the worst-case groups would inevitably ignore the *influential* ones that are truly connected to unseen distributions; optimizing over the worst-case rather than influential groups would yield compromised OOD resilience.

Table 1: Accuracy (%) on *DG-15* and *DG-60*. Our method sets the new SOTA on both datasets.

	ERM [24]	IRM [1]	REx [10]	SD [22]	DRO [18]	Ours
DG-15	58.00	57.87	57.22	57.56	43.22	**67.56**
DG-60	76.02	76.61	86.89	81.04	79.59	**89.19**



(a) DG-15 (b) DG-60

Figure 1: Data groups of (a) *DG-15* and (b) *DG-60* datasets.

4 Approach

4.1 Graph-Relational Distributionally Robust Optimization

Generalizing ML models to arbitrary unseen distributions without any prior knowledge or structural assumption is impossible [6]. Fortunately, the graph structures of $\mathcal{E}_{\text{all}}$ are often available as prior knowledge and can be captured by a graph $G = (V, E)$, where the nodes $V = \cup_{e\in\mathcal{E}_{\text{all}}} X^e$ symbolize the groups and the edges E represent interactions among groups. We assume the graph captures covariate shift ($P_e(X) \neq P_{e'}(X)$) rather that concept shift ($P_e(Y|X) \neq P_{e'}(Y|X)$) . Given the graph G, we can identify *influential groups* and incorporate them as a *physical prior* $\mathbf{p}$ (see Sec. 4.2) to constrain the optimization in Eq. 2: $\mathcal{D}(\mathbf{q}\|\mathbf{p}) \leq \tau$, where $\mathcal{D}(\cdot)$ is a distance metric over the space of distributions. τ is a fixed margin the controls the extent to which we enforce the prior constraint.

4.2 Implementation of Physical Prior

Motivated by centrality analysis [14] in social networks, we propose to assess the *group centrality* to identify *influential groups* that are truly connected to unseen distributions, which can be calculated using graph measurements [23] such as degree, betweenness, and closeness. In this paper, we calculate the betweenness centrality of each node in G as a *physical prior* $\mathbf{p}$ to identify influential groups. Betweenness centrality measures how often a node is on the shortest path between two other nodes in the graph. [4] indicates that nodes with higher betweenness centrality would have more control over the graph as more information will pass through them. We consider two scenarios: graph structure is fully known and partially known.

Fully-known structure denotes the graph structure of all groups $\mathcal{E}_{\text{all}}$ is available. Let $s \in \mathcal{E}_{\text{train}}$ and $t \in \mathcal{E}_{\text{test}}$ be the start and end of a path in G. We define the centrality of group e as the fraction of shortest paths that pass through it: $c_e^{\text{full}} = \sum_{s\in\mathcal{E}_{\text{train}}, t\in\mathcal{E}_{\text{test}}} \frac{\sigma(s,t|e)}{\sigma(s,t)}$, where $\sigma(s,t)$ is the number of shortest paths between groups s and t in the graph ((s,t)-paths), and $\sigma(s,t \mid e)$ is the number of (s,t)-paths that go through group e.

Partially-known structure denotes only the graph structures of training groups $\mathcal{E}_{\text{train}}$ is available. The underlying assumption is that the unobserved part of the graph should not be very different from the observed part and training groups with high centrality also exert strong influence on unseen groups. Instead of sampling groups pairs from two separate sets, we sample (s,t) from $\mathcal{E}_{\text{train}}$. We define the centrality of group e as: $c_e^{\text{partial}} = \sum_{s,t\in\mathcal{E}_{\text{train}}} \frac{\sigma(s,t|e)}{\sigma(s,t)}$.

We use softmax function to normalize c_e and the prior probability for group e is: $p_e = \exp(c_e)/\sum_{e=1}^{m} \exp(c_e)$. In Sec. 5, we empirically found that the proposed method with c_e^{partial} still outperforms other baselines by a large margin and is only slightly worse than that with c_e^{full}.

Table 2: Mean Squared Error (MSE) of task N (24) $\rightarrow$ S (24) on *TPT-48* [25]. Our method achieves the lowest MSE of all test groups.

Group	ERM [24]	IRM [1]	REx [10]	SD [22]	DRO [18]	Ours
Hop-1	1.084	1.133	**0.487**	1.169	0.931	0.889
Hop-2	1.265	1.312	**0.944**	1.354	1.170	0.991
Hop-3	1.975	2.021	2.266	2.091	2.027	**1.678**
All	1.426	1.474	1.194	1.523	1.356	**1.177**

Table 3: Ablation study on partially-known graph structure. Ours (partial) outperforms other baselines by a large margin and is only slightly worse than Ours (full).

	DG-15(↑)	E(24)→W(24)(↓)	N(24)→S(24)(↓)
ERM [24]	58.00	1.716	1.426
DRO [18]	43.22	1.684	1.356
Ours (partial)	66.44	1.471	1.301
Ours (full)	**67.56**	**1.466**	**1.177**

5 Experiments

Datasets. (1) *DG-15* [26] is a synthetic binary classification dataset with 15 groups. Each group contains 100 data points. In this dataset, adjacent groups have similar decision boundaries. Following [26], we use six connected groups as the training groups, and use others as test groups. (2) *DG-60* [26] is another synthetic dataset generated using the same procedure as *DG-15*, except that it contains 60 groups, with 6,000 data points in total. We randomly select six groups as the training groups, and use others as test groups. Visualization of *DG-15* and *DG-60* are shown in Fig. 1. (3) *TPT-48* [25] contains the monthly average temperature for the 48 contiguous states in the US from 2008 to 2019. We focus on the regression task to predict the next 6 months' temperature based on the previous first 6 months' temperature. We consider two generalization tasks: E(24) $\rightarrow$ W(24): we use the 24 eastern states as training groups and the 24 western states as test groups; N(24) $\rightarrow$ S(24): we use the 24 northern states as training groups, the 24 southern states as test groups.

Baselines. We compare our method with the following methods: (1) Empirical Risk Minimization (ERM) [24]; (2) Group distributionally robust optimization (DRO) [18]; (3) Invariant Risk Minimization (IRM) [1]; (4) Risk Extrapolation (REx) [10]; (5) Spectral Decoupling (SD) [15]. Following [7], we perform model selection based on a validation set constructed from training groups only.

Results. Results of *DG-15* and *DG-60* are summarized in Tab. 1. As seen, in both datasets, our method achieves the best performance. In *DG-15*, all other baselines are inferior or ERM while ours outperforms ERM by 9.56%. We show the results for task N (24) $\rightarrow$ S (24) on *TPT-48* in Tab. 2. As observed, our method achieves the lowest average MSE. We also report the average MSE of Hop-1, Hop-2, and Hop-3 test groups. Although REx achieves the lowest error on Hop-1 and Hop-2 groups, it yields the highest prediction error on Hop-3 groups. Our method achieves the best performance on Hop-3 groups, indicating its generalization capability under large distributional drifts.

Ablation Study. We evaluate our method with partially-known graph structure. In this scenario, we assume only the graph structure of training groups are available. We report the results in Tab. 3. As seen, in all datasets, ours (partial) is only slightly worse than ours (full), indicating the strong effectiveness and flexibility of our method.

6 Conclusion

In this paper, we proposed Graph-Relational Distributionally Robust Optimization. We integrate graph information into distributionally robust optimization to develop OOD resilience. Our method strikes a good balance between the worst-case and influential groups, preventing the model from overfitting to worst-case groups and rationally improving generalization performance. Empirical results on both synthetic and real-world datasets demonstrate the effectiveness of our method.

References

[1] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

[3] J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[4] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[5] C. Frogner, S. Claici, E. Chien, and J. Solomon. Incorporating unlabeled data into distributionally robust learning. *Journal of Machine Learning Research*, 22(56):1–46, 2021.

[6] V. Garg, A. T. Kalai, K. Ligett, and S. Wu. Learn to expect the unexpected: Probably approximately correct domain generalization. In *International Conference on Artificial Intelligence and Statistics*, pages 3574–3582. PMLR, 2021.

[7] I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

[8] W. Hu, G. Niu, I. Sato, and M. Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.

[9] M. Koyama and S. Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. 2020.

[10] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

[11] D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

[12] J. Liu, Z. Hu, P. Cui, B. Li, and Z. Shen. Heterogeneous risk minimization. In *International Conference on Machine Learning*, pages 6804–6814. PMLR, 2021.

[13] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

[14] M. E. Newman. A measure of betweenness centrality based on random walks. *Social networks*, 27(1):39–54, 2005.

[15] M. Pezeshki, O. Kaba, Y. Bengio, A. C. Courville, D. Precup, and G. Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[16] A. Robey, G. J. Pappas, and H. Hassani. Model-based domain generalization. *Advances in Neural Information Processing Systems*, 34:20210–20229, 2021.

[17] E. Rosenfeld, P. Ravikumar, and A. Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, volume 9, 2021.

[18] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2019.

[19] S. Shafieezadeh Abadeh, V. A. Nguyen, D. Kuhn, and P. M. Mohajerin Esfahani. Wasserstein distributionally robust kalman filtering. *Advances in Neural Information Processing Systems*, 31, 2018.

[20] S. Shalev-Shwartz and Y. Wexler. Minimizing the maximal loss: How and why. In *International Conference on Machine Learning*, pages 793–801. PMLR, 2016.

[21] M. Staib and S. Jegelka. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[22] P. Su et al. Gradient regularized contrastive learning for continual domain adaptation. In *AAAI*, 2021.

[23] Y. Tian, L. Zhao, X. Peng, and D. Metaxas. Rethinking kernel methods for node representation learning on graphs. *Advances in neural information processing systems*, 32, 2019.

[24] V. Vapnik. Statistical learning theory, 1998.

[25] R. Vose, S. Applequist, M. Squires, I. Durre, M. Menne, C. Williams Jr, C. Fenimore, K. Gleason, and D. Arndt. Gridded 5km ghcn-daily temperature and precipitation dataset (nclimgrid) version 1. *Information, NNCfE, editor. Maximum Temperature, Minimum Temperature, Average Temperature, and Precipitation*, 2014.

[26] Z. Xu, G.-H. Lee, Y. Wang, H. Wang, et al. Graph-relational domain adaptation. In *International Conference on Learning Representations*, 2022.

[27] R. Zhai, C. Dan, Z. Kolter, and P. Ravikumar. Doro: Distributional and outlier robust optimization. In *International Conference on Machine Learning*, pages 12345–12355. PMLR, 2021.